LA INVESTIGACION EVALUATIVA: UNA PERSPECTIVA EXPERIMENTALISTA

Francisco Alvira Martín

En los últimos quince años, los movimientos paralelos de *Indicadores Sociales* e *Investigación Evaluativa* han hecho cada vez más cercana la vieja tesis de Comte del sociólogo-gobernante o las tesis de la ingeniería social predominantes en los inicios de la Sociología como ciencia.

Cuando la NASA y el Gobierno estadounidense a mitad de los años sesenta empezaron a intentar evaluar el impacto del programa espacial sobre su propia sociedad, surgió inmediatamente una doble necesidad:

— *indicadores sociales* que midieran a un nivel adecuado los posibles efectos o consecuencias de dicho programa;
— *procedimientos* para evaluar correctamente dichos impactos, es decir, diseños evaluativos.

La intervención de la Administración en el campo de la educación, servicios sociales, vivienda, rehabilitación de delincuentes, etc., hizo más perentoria la necesidad de *evaluar* dichas intervenciones.

Investigación evaluativa, evaluación de programas, es simplemente la acumulación de información sobre una intervención —programa— sobre su funcionamiento y sobre sus efectos y consecuencias. Más formalmente

> «La investigación evaluativa es, ante todo, el proceso de aplicar procedimientos científicos para acumular evidencia válida y fiable sobre la manera y grado en que un conjunto de actividades específicas produce resultados o efeectos concretos» (L. Ruthman, 1977: 16).

Conjunto de actividades específicas es la *intervención o programa* y los procedimientos científicos se centran en la *medición* y en el *diseño* de la evaluación.

La idea explícita en esta definición de Ruthman de que un *programa produce efectos o consecuencias* (o no los produce, que no es lo mismo pero es igual), conduce sin solución de continuidad a la «Sociedad experimentalista» de Donald Campbell. Pero ésta no es la única posibilidad, aunque quizá sí la más prometedora.

Lyons Morris y Taylor Fitz-Gibbon (1977) distinguen, por ejemplo, seis tipos de modelos evaluativos:

— evaluación orientada a metas/objetivos;
— evaluación orientada a la toma de decisiones;
— evaluación transaccional;
— investigación evaluativa;
— evaluación libre de metas/objetivos;
— evaluación de alternativas.

Insistiendo en su propio modelo de evaluación, el modelo desarrollado en el «Centro para el estudio de la evaluación» (CEV) de la Universidad de California en Los Angeles, modelo que recoge prácticamente los seis anteriores, puesto que hace hincapié en la evaluación como proceso.

M. Scriven (1976) habla de evaluación no condicionada por el coste *(cost-free),* I. Deutscher (1976) insiste en la evaluación sin límites de metas/objetivos, K. Arrow (1976), por último, sin nombrarla, habla de la evaluación sistémica.

Al igual que sucede con la dicotomía evaluación cualitativa/cuantitativa, los diferentes tipos o modelos de evaluación propuestos (sistémica, sin costes, sin límites de metas) plantean problemas metodológicos o epistemológicos más que técnicos, pero al igual que sucede con dicha dicotomía, puede llegarse a una «especie de tregua»: cada modelo o tipo es adecuado y debe utilizarse en determinados tipos de evaluación y en otros no.

Tomemos el caso de la evaluación sistémica, por ejemplo. K. Arrow *(op. cit.)* señala que la utilización de experimentos sociales en la evaluación de programas plantea básicamente problemas de escala:

— en experimentación se analiza una pequeña parte de la realidad social, parte que está conectada con otras y forma un sistema interrelacionado;

— por otra parte, los efectos o consecuencias observados en experimentación suelen ser efectos a corto o medio plazo. No es posible estudiar efectos a largo plazo. Y, sin embargo, estos tipos de efectos son dominantes en Ciencias Sociales.

En el fondo, el problema, según Arrow, está en cómo estudiar y analizar sistemas complejos e interrelacionados; no es un problema de imposibilidad de predicción por el «libre albedrío» humano, ni, por tanto, un problema propio sólo de las Ciencias Sociales. Se plantea también en la Meteorología, Astronomía, etc., donde resulta ridículo hablar de experimentación.

Por ello, en la evaluación sistémica, el primer paso es el análisis y descripción del sistema o subsistema que se pretenda evaluar como, por ejemplo, hace R. Gordon Cassidy en *A System approach to plannig and evaluation in criminal justice systems*. Parece evidente que no todo programa de intervención tiene por qué ser visto como un subsistema —o parte de uno— que afecte a todo el sistema social. Algunos programas deberán ser estudiados sistémicamente, pero esto es una cuestión empírica. Por otra parte, los dos problemas de escala señalados por Arrow son problemas relacionables con la *validez externa* de los diseños de investigación y pueden resolverse en parte. De hecho, en los estudios longitudinales a largo plazo se pueden estimar efectos no inmediatos.

Se debe reconocer que el alcance de un programa puede ser menor del alcance que debe tener la correspondiente evaluación en línea con la perspectiva sistémica. Así, el famoso programa sobre control social de la delincuencia en Hyde Park en Chicago estaba restringido a dicha área urbana. La evaluación se hizo sólo en el área, resultando positiva: se logró que los atracos en la calle disminuyeran, pero probablemente a costa de «expulsar» hacia otros barrios contiguos dicha delincuencia. Hyde Park forma parte de un sistema urbano y como tal debería de ser tratado. Ahora bien, no sólo cabe acotar el espacio, más amplio, afectado por el tema del control de la delincuencia en Hyde Park, sino que, además, puede que no se «expulsara» delincuencia, sino que simplemente el control llevado a cabo fuera totalmente disuasor de dichos actos.

La perspectiva sistémica complementaría en este caso una perspectiva estricta de «investigación evaluativa» o «evaluación de fines específicos».

El modelo desarrollado por el CEV, como ya dije, es a la vez más eléctrico y más comprensivo. En este modelo la evaluación se desarrolla en cuatro fases:

— evaluación de *necesidades* que sirve para determinar los fines u objetivos del programa;
— planificación del programa;
— evaluación «formativa».
— evaluación «sumativa».

En el caso de la evaluación «formativa», se intenta no sólo ayudar a que el programa empiece a funcionar, sino ante todo a conceptualizar en qué consiste el programa y cómo funciona en el papel y en la realidad. En la evaluación «sumativa» el objetivo es el análisis del impacto o consecuencias del programa. Resulta evidente que una investigación evaluativa específica puede desarrollar las cuatro fases o solamente una o varias de ellas, es decir, puede simplemente evaluar las necesidades, o analizar cómo está funcionando el programa o evaluar si los fines perseguidos se están cumpliendo, etc., o todo ello conjuntamente.

La posibilidad de evaluación de programas o intervenciones sociales

La evaluación «formativa» puede servir —y sirve—, entre otras cosas, para determinar *a priori* si es posible la evaluación de un programa y en qué sentido es posible dicha evaluación.

J. S. Wholey (1977) plantea como objetivos de esta estimación de la posibilidad de evaluación los siguientes:

1. Descripción detallada del *programa* y de su funcionamiento.
2. Delimitación precisa de los *fines* perseguidos.
3. Determinación del *mecanismo* que une el conjunto de actividades del programa y los fines o metas que se quieren conseguir.

Esta delimitación de los tres aspectos que constituyen una precondición a la subsiguiente evaluación, implica moverse a dos niveles distintos:

— De una parte, al nivel del modelo teórico/legal del programa que estará en el conjunto de documentos que han puesto en marcha el programa, en la mente de sus impulsores o legisladores, etc., y que Wholey llama el modelo *retórico*.
— De otra, el modelo que describe simplificadamente el funcionamiento real del programa y que está en la gente que realmente hace funcionar día a día el programa y en su propio funcionamiento.

La evaluación de la posibilidad de evaluación implica contrastar el modelo teórico/retórico con el modelo en funcionamiento y derivar un posible modelo a evaluar. Ahora bien, este *modelo evaluable* no sólo debe deducirse por el investigador, sino que además una vez «producido» por éste, debe discutirse, tanto con los responsables del modelo teórico/retórico, como con los responsables operativos del funcionamiento del programa.

Precisamente uno de los puntos clave de esta «negociación» del modelo evaluable es el llegar a un acuerdo de que constituye *evidencia válida y fiable* en la evaluación de dicho programa. La evidencia o información sobre la rela-

ción entre actividades y objetivos debe ser válida y fiable desde el punto de vista científico, pero también lo debe ser desde el punto de vista de los *usuarios* de la evaluación. Desde el punto de vista de éstos, debe gozar de *credibilidad* y los criterios de esta credibilidad deben ser fijados antes de la evaluación.

Resulta evidente que esta fase de estimación de la posibilidad de evaluar un programa es en sí misma una evaluación de un tipo determinado: una evaluación *a priori* de la evaluación.

Por ello, aunque la decisión última sea que es imposible llegar a un modelo evaluable por la ambigüedad de los objetivos o su inexistencia, porque no existen criterios de credibilidad tajantes, etc., lo cierto es que se habrá efectuado una evaluación del programa, aunque sea indirecta.

Esta fase de evaluación «formativa» y de evaluación de la «evaluabilidad», conecta directamente con un cierto modelo fenomenológico de diseños evaluativos expuesto, por ejemplo, por I. Mitroff (1983).

El tratamiento de problemas poco —o mal— estructurados requiere, según Mitroff un enfoque que permita su estructuración y su posterior relación con la acción a través de la participación de usuarios de la evaluación, de usuarios del programa, de los expertos evaluadores, en una palabra, de todos los implicados, tanto en la evaluación como en el programa. La premisa básica es que «la naturaleza del problema está en los ojos de los implicados» (I. Mitroff, *op. cit.*: 167), y por tanto sólo a través de su participación se puede llegar a un conocimiento —evaluación— válido.

La participación sirve para llegar a determinar los supuestos básicos del problema y preparar el camino para el análisis del mismo a través de la argumentación. A la síntesis del conocimiento se llega a través de una síntesis de grupos, fase última muy parecida a la señalada antes de contrastar y discutir los modelos teórico/retórico, operacional y evaluable con los participantes para llegar a un modelo evaluable consensuado.

La lógica experimental en la evaluación de programas

Un programa o una intervención en la realidad social es un *conjunto de actividades más o menos complejas* que se pretende produzcan un determinado impacto. Precisamente un *experimento* es simplemente la manipulación de un fenómeno o conjunto de fenómenos para ver sus efectos o consecuencias. Cuando durante el proceso de manipulación o intervención y de medición y análisis de sus efectos se controlan todas aquellas posibles *explicaciones alternativas* que no sean las de que esa *intervención* ha producido *esos efectos determinados,* entonces estamos ante un experimento que tiene validez interna. (Véase Campbell y Stanley, 1967, y F. Alvira y otros, 1979.)

En experimentación, la base del control de estas hipótesis explicativas alternativas radica en el propio hecho de la manipulación. El experimentador

no tiene que esperar a que se produzca un fenómeno y observar/analizar sus efectos, sino que crea el fenómeno, y lo aplica a quien quiere y en el lugar y tiempos adecuados. Esto permite el control de dichas hipótesis alternativas básicamente a través de dos técnicas:

— el uso del pretest/postest;
— y el uso de grupos de control.

En el caso de la evaluación de programas, el concepto de validez interna y las técnicas para lograrla son aplicables en diferente grado, dependiendo del tipo de programa y sus circunstancias. Un programa en marcha que deba ser evaluado hoy, no permite el uso del pretest, es decir, de la medición «antes» de la intervención; más importante aún, normalmente el «quién» del programa, es decir, a quién debe afectar, no es negociable e incluso puede depender de la voluntad de los propios interesados.

Realmente, lo único necesario para lograr un control de hipótesis alternativas es conseguir un grupo de control y un grupo «experimental» que puedan ser comparados. Para ello, es preciso que el grupo de control sea *auténticamente* igual al grupo experimental excepto en lo que respecta a la aplicación del programa y la única manera para lograrlo es la *formación aleatoria* de ambos grupos. Es esta condición de aleatorización la que resulta difícil de cumplir en la evaluación de programas.

Resulta difícil, pero no imposible, y sobre todo la lógica experimental debe actuar como una meta a la que aproximarse más que como un o «todo» o «nada». Desde esta idea de aproximarse al experimento aleatorizado se han desarrollado tres enfoques diferentes (véase L. Kish, 1979, y R. F. Boruch, 1976).

1. Aquellos que siguen la idea de utilizar el control estadístico *a posteriori* para «igualar» los grupos que se estén comparando y que Boruch denomina «igualadores». Utilizan todo tipo de técnicas de análisis estadístico como el emparejamiento, el análisis de regresión o covarianza de modo que las hipótesis explicativas alternativas queden controladas y sólo quede el efecto del programa que nos interesa.

2. Los «ajustadores» en la terminología de Boruch, que siguen el método hipotético-deductivo casi al pie de la letra y se encuadran casi siempre dentro de una perspectiva sistémica. Se elabora un modelo teórico más o menos complejo y se deducen de él las relaciones empíricas que se deben observar si el modelo es verdadero. Se «ajustan» los datos obtenidos en la investigación a los datos derivados del modelo y se analizan los desajustes o residuos. Los seguidores de «modelos de ecuaciones estructurales» o «modelos econométricos» u otro tipo de modelos, sigue este camino.

3. Los dos enfoques anteriores pueden clasificarse como enfoques de control *a posteriori,* sobre todo el primero. El último enfoque se basa en el

control *a priori* (véase L. Kish, *op. cit.*). Se trata de buscar *diseños* de evaluación que se aproximen al máximo a los experimentos aleatorizados. Este es el camino seguido por D. Campbell o T. Cook en el desarrollo de diseños cuasi-experimentales, por ejemplo.

Las tres perspectivas no son independientes entre sí y pueden/deben utilizarse conjuntamente. Precisamente en esa dirección se encaminan los desarrollos en evaluación de programas que intentan lograr una alta validez interna. Veamos algunos ejemplos.

a) *Utilización conjunta de experimentos y cuasiexperimentos*

Algunos programas sociales permiten utilizar con ciertos grupos auténticos experimentos aleatorizados, mientras que no lo permiten con otros grupos para los que debe diseñarse cuasiexperimentos con grupos de control «aparentemente» iguales. Tener este doble diseño permite estimar el efecto del programa (con el experimento) y el efecto de variables perturbadoras (en el cuasiexperimento). Además, ambas estimaciones pueden combinarse, de modo que se logre una estimación del efecto total del programa sobre la población utilizando técnicas de regresión y covarianza (véase Boruch, *op. cit.*: 42).

b) *Experimentos dentro de cuasiexperimentos y cuasiexperimentos dentro de experimentos*

En el *primer caso,* el programa objeto de evaluación no permite la utilización de experimentación pero ésta puede realizarse en algunas partes del programa o en algunos momentos o aspectos del mismo. Un caso interesante se produce en los diseños de *discontinuidad en la regresión* (véase Campbell y Stanley, *op. cit.)* en que los sujetos que van a beneficiarse del programa deben reunir unos ciertos requisitos (rentas por debajo de un cierto nivel, cociente intelectual alto, etc.). En estos casos *todas las personas* que cumplen los requisitos para ser beneficiarios reciben el programa y no puede aleatorizarse la asignación. Sin embargo, puede que los criterios establecidos tengan un nivel de ambigüedad en el que no esté claro si una persona es beneficiario o no. En estos niveles de ambigüedad puede aplicarse la aleatorización, de modo qe al final pueda disponerse de cuatro grupos:

1. Grupo que no se beneficia del programa por no cumplir los requisitos.
2. Grupo que no se beneficia porque está dentro del intervalo de ambigüedad y aleatoriamente se le deja fuera del programa.
3. Grupo que se beneficiará al ser incluido aleatoriamente por estar en el intervalo de ambigüedad.

4. Grupo que se beneficia del programa por cumplir inequívocamente los requisitos establecidos.

De este modo, contaremos con un experimento en el marco de un cuasi-experimento. El control *a posteriori* de tipo estadístico permitirá la estimación de efectos globales al igual que en el caso *a)*.

En el *segndo caso,* la experimentación puede realizarse a un cierto nivel de agregación y análisis —por ejemplo, colegios o institutos—, pero no puede ralizarse al nivel más desagregado —alumnos o profesores—. De este modo, se diseña un experimento a un nivel y al nivel de desagregación se planea un cuasiexperimento.

c) *Desfase en la aplicación del programa*

Aun dentro de un programa que afecte a toda una población cabe muchas veces dosificar la aplicación del mismo incluso por razones técnicas y de gestión. Es decir, aleatoriamente se extrae un subgrupo que aunque se beneficie del programa lo haga en una fase posterior, de tal manera que pueda servir de grupo de control mientras no recibe el tratamiento.

Siempre que sea posible puede subdividirse este grupo de control de modo que se formen varios grupos que van sucesivamente recibiendo los tratamientos del programa y van siendo utilizados uno tras otro como grupos de control. Naturalmente, estos grupos deben formarse aleatoriamente y el análisis y estudio posterior de efectos implica utilizar técnicas de análisis estadístico de control *a posteriori.*

En algunos casos, incluso, esto no es posible, pero existe una variante derivada de los diseños factoriales. Si no es posible lograr un grupo de control estricto puede lograrse que haya diferentes niveles de tratamiento o tratamientos alternativos de tal modo que se formen varios grupos aleatoriamente que sean sometidos a distintos niveles de tratamiento o a tratamientos alternativos. Unos grupos actúan aquí de control de los otros como sucede en los diseños experimentales factoriales.

La *validez interna* no es el único tipo de validez que una investigación evaluativa debe lograr; la capacidad de poder generalizar los resultados obtenidos o *validez externa* es igualmente importante. De hecho, algunas de las técnicas señaladas incrementan tanto la validez interna como la externa, puesto que ambas están estrechamente interrelacionadas.

La lógica experimental se ha centrado ante todo en la validez interna bien porque se pensará que la validez externa lógicamente debe seguir a la interna (¿qué generalizar si no sabemos interpretar los resultados?), bien porque se creyera, como lo hacía el primer Campbell, que la paradoja de la inducción impedía lograr validez externa y que la única escapatoria era la repetición de investigaciones.

Sin embargo, últimamente, se han producido desarrollos no sólo para lograr métodos para aumentar dicho tipo de validez, sino también centrados en la clasificación y codificación de los diferentes aspectos relacionados con la posibilidad de generalizar los resultados obtenidos en una investigación evaluativa.

Desde un punto de vista metateórico parece claro que las amenazas a la generalización de resultados provienen siempre de efectos interactivos. Claro está que este conocimiento no resuelve el problema. De un modo general existen cuatro grandes fuentes de factores que restan validz externa a un diseño:

— problemas de selección;
— errores de medición;
— factores asociados con el tratamiento, y
— factores de la situación o contexto en que tiene lugar el programa.

Los aspectos relacionados con la medición forman parte del problema de la validez «de constructo» y el lector interesado en este tema puede recurrir al artículo de Campbell y Cook (véase Campbell y Cook, 1976).

Prescindiendo, por tanto, de los aspectos relativos a la validez de constructo quedan dos grandes apartados relacionados con las unidades de análisis (selección) y el contexto en el que tiene lugar la investigación evaluativa o el programa. El problema de la confusión de «tratamientos» se refiere en cambio a la identificación precisa y clara de que constituye el conjunto de actividades del programa y a su separación en la práctica de otros fenómenos que pueden darse a la vez.

En lo que respecta a la *selección,* los problemas más graves surgen cuando se trata de una selección de *voluntarios,* es decir, el programa está abierto a toda persona que lo solicita o se produce una selección por *conveniencia,* es decir, el programa se aplica a las personas que están más disponibles. Obviamente, los problemas surgen no con programas que realizan este tipo de selección, sino con las investigaciones evaluativas que no pueden evitar el mismo tipo de selección.

En ambos casos, este tipo de selección no aleatoria plantea problemas de validez interna y validez externa: resulta difícil separar el efecto debido al programa del efecto debido al haberse ofrecido como voluntario que produce normalmente un efecto placebo. Al no poder separar ambos efectos, la generalización resulta también difícil de llevar a cabo.

Muy parecido a este efecto placebo provocado por una selección voluntaria, es el efecto Hawthorne que en cambio debe ser conceptualizado dentro de los factores de tipo situacional o ecológico que pueden interferir con la validez externa. El reconocimiento por parte de los participantes en una evaluación de un programa de que están participando en dicha evaluación puede producir cambios debidos exclusivamente a este «saberse observados».

Más aún, si se sabe que del resultado de la evaluación puede depender la supervivencia de un programa determinado.

La especificidad del personal que administra el problema, el contexto geográfico y el momento concreto temporal en que se efectúa la investigación evaluativa y/o la aplicación del programa son otros tantos aspectos de los problemas situacionales/ecológicos relacionados con la validez externa.

Por último, quiero resaltar una serie de aspectos relacionados con los tratamientos, o sea aquello que constituye el programa. En muchos casos resulta especialmente difícil poder saber en qué consiste realmente aquello a que ha sido sometido un grupo de personas y sobre todo asegurarse de que todas las personas del grupo que recibe el tratamiento (el programa) reciban el mismo tipo de tratamiento. La uniformidad de tratamientos es vital para poder interpretar los resultados y luego poder generalizarlos.

Este hecho es tanto más importante cuanto mayor alcance y extensión geográfica tenga el programa. Por ejemplo, la rehabilitación y reinserción de drogadictos ha atraído a múltiples asociaciones, cada una de las cuales dispone de su propio método de rehabilitación. No sólo es muy difícil determinar en qué consiste este método y su unicidad sino que una asociación que disponga de varios centros a lo largo y ancho de España no lo aplicará de una manera similar y uniforme en todos ellos. La evaluación de este método de rehabilitación de drogadictos resultará muy difícil por esta razón, que afecta tanto a la validez interna como a la externa.

La lógica experimental no debe verse como algo rígido que implique el planteamiento de diseños experimentales clásicos como el de Solomon, o el de grupo de control con pretest y postest, sino que debe verse como un conjunto de medidas u operaciones a tomar en el planteamiento de diseños de evaluación de programas para lograr el control de hipótesis alternativas, es decir, para lograr una correcta interpretación de los resultados obtenidos en la evaluación.

Esto se consigue siguiendo diferentes pasos:

1. Siempre que sea posible, formar uno o varios grupos de control aleatoriamente.
2. En cualquier caso, buscar siempre un grupo de control equiparable utilizando las variables relevantes y si es necesario el pretest.
3. Cuando no se pueda utilizar la aleatorización en la formación de los grupos, aparte del pretest, los diseños de series temporales resultan tan eficaces en muchas situaciones como los propios diseños experimentales.
4. Por último, a falta de aleatorización, se pueden combinar la igualación *a priori* del grupo de control con el control estadístico *a posteriori*. Ni que decir tiene que todo este proceso debe estar guiado por la teoría y por los conocimientos sobre el programa y cómo actúa —o cómo se espera que actúe— adquiridos durante la fase de evaluación formativa.

Un ejemplo de investigación evaluativa

Un programa social o un programa de intervención social no tiene por qué ser algo excesivamente complicado. J. Bentham y los utilitaristas ingleses proporcionaron una justificación filosófico/psicológica a la pena como disuasora de la comisión de delitos dentro del sistema judicial. Lo que puso en marcha Bentham fue un auténtico programa de intervención social cuyo objetivo era la prevención de la delincuencia a través de un mecanismo determinado: la pena o sanción y su gradación de modo que contrarrestara las posibles ganancias derivadas de la comisión de delitos. La pena capital correspondería así a delitos de sangre y otros similares en su magnitd y beneficios derivables por su comisión.

Dentro del movimiento de abolición de la pena de muerte en los Códigos penales, uno de los argumentos utilizados ha sido su falta de poder disuasorio. T. Sellin, en trabajos pioneros, investigó la relación entre pena capital y tasas de homicidios utilizando dos diseños diferentes pero similares:

1. Comparaba Estados de Estados Unidos con y sin pena de muerte en ss Códigos en lo que respecta al número de homicidios cometidos. Se trata de un diseño típico no experimental con medición postest y un grupo de control no equivalente.

2. Analizaba las variaciones de la tasa de homicidios en diferentes Estados antes y después de la abolición —o reestablecimiento— de la pena de muerte. Este es un diseño típico de series temporales o discontinuidad en la regresión.

Sus conclusiones fueron que no existía un efecto disuasor sobre las tasas de homicidio. No había diferencias significativas en las tasas de homicidio de Estados con y sin pena de muerte.

La *pena capital* se relaciona en estos estudios de Sellin y en otros similares con la *tasa de homicidios* pero, obviamente, el primer problema estriba en que las consecuencias o efectos de la pena capital pueden ser otros: puede tener un efecto disuasor sobre otros delitos o sobre los homicidios cometidos contra los encargados del control social. Pero esto no es el principal problema y para todo lo referente al problema general de la disuasión remito al lector a F. Alvira (1984).

El problema básico es si realmente la interpretación de los resultados ofrecida por Sellin es la correcta a la vista de que *no se han controlado las posibles hipótesis alternativas*. Los Estados que se comparan no son equiparables ni han sido igualados aleatoriamente, no existe control sobre los cambios ocurridos en el sistema judicial ni en el sistema policial, etc.

Precisamente estos problemas han hecho resurgir la investigación sobre el efecto disuasor de la pena capital de la mano de economistas y econometras que han cambiado el énfasis desde diseños cuasiexperimentales a modelos

econométricos utilizando análisis de regresión múltiple o modelos derivados de Box & Jenkings. I. Eearlich (1979), por ejemplo, utiliza una serie de modelos de regresión en sucesivos análisis y llega a la conclusión de que sí se produce un importante efecto disuasor sobre las tasas de homicidio.

La aceptación de los resultados de Earlich y otros autores que escriben dentro de la misma línea debe estar condicionada, entre otras cosas, por:

— cerciorarse de que el modelo incluye todas y cada una de las variables relevantes; es decir, asegurarse de que está correctamente especificado;
— confirmación de que los indicadores utilizados realmente son fiables y válidos.

En cualquier caso, la conclusión será siempre y en el mejor de los casos que los datos no contradicen el modelo, pero nunca que los datos confirman el modelo. L. S. Friedman lo entiende perfectamente al afirmar que

> «el análisis de regresión multivariado nunca puede dar una prueba absoluta de que un factor causa otro, aunque pueda ofrecer evidencia circunstancial plausible de casalidad» (L. S. Friedman, 79: 63).

Lo que habría que hacer es intentar un diseño evaluativo que juntara las tres perspectivas apuntadas por Boruch: la de los «ajustadores», «diseñadores» e «igualadores». Es decir, intentar buscar diseños cuasiexperimentales menos ingenuos que los utilizados en la primera época de los estudios sobre el carácter disuasor de la pena de muerte introduciendo técnicas de control e igualación de grupos *a priori* y *a posteriori*, efectuando también análisis mediante modelos econométricos o matemáticos en general.

Claro está que si se parte de la premisa de que la abolición de la pena de muerte es una cuestión ético/moral, entonces nos encontraríamos con que no haría falta realizar ninguna evaluación, dado que sería un programa no evaluable.

BIBLIOGRAFIA

ALVIRA, F.: "El efecto disuador de la pena", *Estudios penales y criminológicos*, Santiago de Compostela, 1984.
— y otros: *Los dos métodos de las ciencias sociales*, CIS, 1979.
BERNSTEIN, I. N.: *Validity Issues in Evaluation Research*, Sage, 1976.
BORUCH, R. F.: "Combining Randomized Experiments and Approximations to Experiments in Social Program Evaluation", en I. N. BERNSTEIN: *Validity Issues in Evaluative Research*, Sage, 1976.
CAMBELL, D., y COOK, T.: "Diseños experimentales y cuasi-experimentales", en H. M. DUNNETTE: *Handbook of Organization and Industrial Psychology*, Rand McNally, 1976.
CAMPBELL, D., y STANLEY, J.: *Diseños experimentales y cuasi-experimentales en educación*, Amorrortu, 1967.

EARLICH, I.: "The Economic Approach to Crime" y "Capital Punishment and Deterrence: some Futher Throughts and Additional Evidence", en S. L. MESSINGER y E. BITTNER (eds.): *Criminology Review Book*, vol. 1, Sage, 1979.

FRIEDMAN, L. S.: En S. L. MESSINGER y E. BITTNER (eds.): *Criminology Review Book*, vol. 1, Sage, 1979.

GORDON, R.: "A System Approach to Planning and Evaluation in Criminal Justice Systems", *Socio-Economic Planning Sciences*, 1975.

KENDEL, E., y MICHAEL, R.: "Evaluating and Planning of a Componnet in the Criminal Justice System", *Socio-Economic Planning Sciences*, vol. 6, 1972.

KISH, L.: "Representación, aleatorización y control", en F. ALVIRA y otros: *Los dos métodos de las ciencias sociales*, CIS, 1979.

LYONS, L., y TAYLOR, C.: *Evaluator's Handbook*, Sage, 1978; *How to Deal with Goals and Objectives*, Sage, 1978; *How to Design a Program Evaluation*, Sage, 1978; *How to Measure Program Implementation*, Sage, 1978; *How to Measure Program Implementation*, Sage, 1978; *How to Measure Achievement*, Sage, 1978; *How to Measure Attitudes*, Sage, 1978; *How to Calculate Statistics*, Sage, 1978; *How to Present an Evaluation Report*, Sage, 1978.

MCDOWALL, D. y col.: *Interrumpted Time Series Analysis*, Sage, 1980.

MITROFF, I.: "Beyond Experimentation: New Methods for a New Age", en E. SEIDMAN: *Handbook of Social Intervention*, Sage, 1983.

RUTHMAN, L.: *Evaluation Research Methods: a Basic Guide*, Sage, 1977.

SAXE, L., y FINE, M.: *Social Experiments. Methods for Design and Evaluation*, Sage, 1981.

SPECTOR, P.: *Research Design*, Sage, 1981.

STERN, P.: *Evaluating Social Science Research*, Oxford Univ. Press, 1979.

WALL, W., y WILLIAMS, H.: *Longitudinal Studies and Social Sciences*, Heineman, 1970.

WHOLEY, J.: "Evaluability Assessment", en L. RUTHMAN: *Evaluation Research Methods: a Basic Guide*, Sage, 1977.