

NACIONES
UNIDAS



CEPAL

ILPES

INSTITUTO LATINOAMERICANO
DE PLANIFICACION
ECONOMICA Y SOCIAL

PROGRAMA DE CAPACITACION

Documento ASD-117

EVALUACION Y EXPERIMENTACION *

- * El presente documento, que se reproduce para uso exclusivo de los participantes de cursos de la Dirección de Programas de Capacitación, se ha tomado del libro "Análisis de formación de decisiones políticas" de Edward S. Quade, Instituto de Estudios Fiscales, Madrid, 1989.



EVALUACION Y EXPERIMENTACION

La presentación eficiente de servicios públicos exige algo más que la manifestación y el montaje de lo que los políticos han decidido que es el mejor programa que ellos pueden diseñar. En fechas recientes, los políticos han comenzado a darse cuenta de que también necesitan determinar cómo se comportan realmente estos programas --medir los resultados bajo las condiciones de funcionamiento y descubrir si el programa está llevando a cabo lo que pretendía y, de no ser así, estar en situación de interrumpirlo o de mejorarlo. La investigación necesarias ha pasado a denominarse evaluación: en este sentido, la evaluación ha logrado prominencia general durante los últimos veinte años, aunque ha tenido una historia medianamente larga en los campos de la educación y de la medicina. Con la difusión de nuevos programas sociales, ha tenido un amplio reconocimiento por parte de funcionarios públicos la necesidad de la evaluación. La legislación federal la exige actualmente; se ha proporcionado dinero, se ha creado y fortalecido el personal evaluador, se han emprendido estudios de evaluación importantes, hecho uso de algunos métodos experimentales rigurosos y ha prestado atención a los resultados.

La evaluación, en su utilización más general, es un proceso que se lleva a cabo para determinar el mérito. Es, de este modo, una parte de cualquier análisis de decisión pública. Pese a su utilización más amplia, en este capítulo el término se aplica como antes en un sentido más restringido y técnico como una investigación para medir hasta qué punto la realización efectiva de un programa en curso (o, menos frecuentemente, un programa completo) concuerda con la ejecución prevista. No es, sin embargo, totalmente retrospectiva, ya que el objetivo puede ser señalar un cambio en la asignación de los recursos, mejorar el funcionamiento o planificar actividades futuras. El elemento clave para distinguir la evaluación de, pongamos por caso, el análisis coste-beneficio de un programa propuesto es que en este último no tiene que manejárselas con una actividad en curso de realización y, de este modo, con la gente que participa en la actividad o que es afectada por ella. Esta gente obra entre sí e influye sobre una valoración por unas vías que deben ser tomadas en cuenta por los analistas.

Las evaluaciones cubren un amplio espectro, tal como muestran los siguientes títulos:

HEAD START: Final Report of the Evaluation of the Second Year Program, L.M. Vogt et. al., Urban Institute, Washington, 1973.

Evaluation Results for the Los Angeles Woman's Job Corps Center, S. Carrol et. atl. R-745-OEO, The Rand Corporation, Santa Mónica, 1971.

The Indianapolis Police Fleet Plan: An Example of Program Evaluation for Local Government, Donald M. Fisk, Urban Institute, Washington, octubre, 1970.

Evaluation of the JOBS Program in Nine Cities (Informe final) TM-WD-(L)-313/001/000. Systems Development Corporation, Falls Church, Virginia, septiembre 1969.

Final Report: Evaluation of Office of Economic Opportunity Legal Services Program. Volúmenes 1 y 2. John D. Katelle Corporation, Paoli, Pennsylvania, 1971.

La esencia de la evaluación es la valoración del resultado de un programa o de una actividad --lo que sucede y no habría sucedido en su ausencia-- y de la eficiencia relativa dentro de conjuntos de programas o actividades-- ¿qué proyectos concretos o clases de proyectos funcionan mejor? La razón para llevar a cabo una evaluación es ofrecer información bien a los políticos, sobre el coste y los efectos de sus programas y proyectos y ayudar en la asignación eficaz de los recursos, o a los gestores de los programas, para proporcionarles ayuda en la gestión eficaz de sus programas, o bien ambas cosas. En otras palabras, la evaluación de los proyectos o programas públicos es un proceso de valoración diseñado para ofrecer información sobre las operaciones reales y su eficiencia con el fin de prestar ayuda en la toma de decisiones sobre el futuro.¹ Desgraciadamente, mucha actividad clasificada

¹ Puede resultar útil trazar una distinción entre evaluación y estudios tradicionales de intervención a posteriori, incluso aunque los auditores cambien su punto de vista (Staats, 19973). Fueron realizados para objetivos diferentes --en ciertas cosas, casi opuestos. La intervención tradicional a posteriori trata de comprobar la legalidad y corrección de las transacciones financieras y asegurar que los techos presupuestarios no se sobrepasarán. No se trataba realmente de evaluación, porque no informaba a los políticos o gestores de programas de lo que han realizado las actividades objeto del programa. Trata con inputs, no con outputs. Además, la intervención a posteriori convencional tendía a mirar hacia atrás intentaba buscar las culpas. Sólo contribuía a mejorar en el sentido de que servía a una función disuasoria, previniendo probablemente que se produjeran ciertos abusos. Una evaluación, por otro lado, debería mirar

como evaluación ha sido irrelevante porque no produjo la clase de información necesaria a quienes toman decisiones.

El siguiente ejemplo, sugerido por Kathleen Archibald, puede ayudar a ilustrar las dificultades que surgen.

Supongamos que usted es un funcionario público, que toma decisiones y que está preocupado por la elevada tasa de desempleo de latinoamericanos en San José (o de los negros en Albuquerque o de los trabajadores de la aeronáutica en Seattle). Usted gestiona la prestación de cierto dinero federal para poner en marcha un proyecto de formación de mano de obra destinado a ellos --el único existente en la comunidad. Estando verdaderamente interesado por la eficiencia del programa, contrata una evaluación. Usted decide, y el evaluador concuerda, que necesita conocer si el programa produce sus efectos. Se realiza entonces un estudio competente que, casualmente, cuesta bastante dinero. Supongamos que descubre lo siguiente:

1. La tasa de desempleo entre los participantes en el programa es la misma que en el grupo de control, es decir, no existe evidencia de que su programa esté produciendo un efecto.
2. La tasa de desempleo de los participantes es un x por 100 menor que en el grupo de control.

Con esta información, ¿qué se supone que hace usted? Si el gasto no tiene impacto alguno, no le dice cómo cambiar el programa. Si la tasa de participantes es más baja, ¿significa esto que está haciendo una buena labor? Incluso si la tasa de desempleo de los participantes es cero, sería bueno conocer si podría reducir los servicios, añadir más participantes y alcanzar todavía un buen resultado. El tema está en que la información sobre posibles cambios para mejorar la ejecución o incluso la forma en que

principalmente hacia adelante y ayudar a la dirección a decidir lo que ha de hacer después.

La clave de la distinción es decidir los efectos deseados de la investigación. Si la investigación está proyectada para producir un impacto sobre una decisión tocante al tamaño del programa, estrategia o funcionamiento, se trata de una investigación adecuada, si controla el cumplimiento de regulaciones, directrices, cargas laborales o prácticas administrativas, entonces se trata de un control de conformidad. El estudio de estas últimas cosas por lo general no aporta ninguna luz sobre el valor del proyecto, ya que la conformidad con los fines legislativos y con las regulaciones administrativas puede tener poco que ver con el rendimiento.

funciona el programa, en algún sentido relativo es probable que sea más útil que la información en un sentido absoluto.

Debido a que la evaluación es una forma de análisis político, son aplicables los procedimientos del análisis político. Los objetivos del programa a evaluar deben identificarse, descubrirse medidas de eficiencia y determinarse los criterios para comparación. Las alternativas a comparar son entonces investigadas, los datos recopilados y los costes y otras consecuencias valoradas. Existen siempre al menos dos alternativas --continuar el programa como está o suspenderlo-- y, por lo general, más, por lo que el interés debería estar en mejorar el funcionamiento.

En la evaluación, nos gustaría comparar lo que realmente está sucediendo o lo que sucederá como consecuencia del programa con lo que sucedería si el programa se suspendiera o modificara. Para llevar a cabo las comparaciones pueden hacerse varios bocetos. Por ejemplo, Hatry² et al. (1973) observan:

Idealmente, nos gustaría comparar lo que "realmente sucedió" con lo que "habría sucedido si el mundo hubiera sido exactamente el mismo, salvo que el programa no se hubiera puesto en funcionamiento". Puesto que es imposible determinar exactamente lo que "habría ocurrido si ...", el problema es utilizar procedimientos para aproximarse a esto.

Con este fin presente, Hatry y sus colegas indican cinco aproximaciones para identificar y cuantificar los efectos debidos al programa.

1. Antes vs. después de comparar el programa. Comparar los resultados del programa desde la misma jurisdicción medida en dos momentos distintos --inmediatamente antes de poner en marcha el programa y en su momento adecuado después.
2. Proyección de la tendencia temporal de los datos anteriores al programa vs. datos reales después del programa. Comparar los datos reales después del programa con los datos estimados a partir de un número de períodos temporales anteriores al mismo.

² Su definición de programa de evaluación excluye el proceso de desarrollar alternativas a los programas existentes y analizar los probables costes y efectos futuros de estas alternativas, centrándonos en su lugar en la ejecución anterior real de programas existentes o consumados.

3. Comparaciones con jurisdicciones o segmentos de población no cubiertos por el programa. Comparar los datos de la jurisdicción donde funciona el programa con los de otras jurisdicciones donde no funciona.
4. Experimentación controlada. comparar grupos similares, preseleccionados, algunos de ellos atendidos y otros no (o atendidos de forma diferente). El aspecto fundamental es que los grupos comparados son fijados previamente antes de poner en funcionamiento el programa de tal suerte que los grupos sean tan similares como sea posible, excepto para el tratamiento del programa.
5. Comparaciones de rendimientos planificados vs. reales: Comparar los datos posteriores al programa real con los objetivos fijados en fechas anteriores --bien antes de poner en práctica el programa o bien en cualquier período a partir de su puesta en funcionamiento.

La experimentación controlada es la aproximación más provechosa, pero también la más cara y la más difícil de llevar a cabo. Un problema importante en todos estos esquemas consiste en identificar y tener en cuenta la influencia sobre la comparación de efectos externos no relacionados con el programa --por ejemplo, un cambio en la economía u otros programas que podrían realizarse simultáneamente.

Existen diferencias en la forma de procurar recopilar los datos para calcular el rendimiento de los programas alternativos mientras se encuentran todavía en la fase de propuestas y evaluarlos después de su entrada en funcionamiento. A la mayoría de la gente no le gusta la evaluación de su trabajo, incluso cuando está de acuerdo en que la evaluación es una buena idea. Y cuanto más se figuran que pueden estar realizando un mal trabajo, menos les agrada ser evaluados. A veces, el temor a ser evaluados llega casi a la "paranoia". La gente vinculada a un programa, tanto gestores como participantes, está preocupada porque el analista pueda no estar de su parte (especialmente cuando procede de fuera de la organización y, en particular, si trabaja para una agencia presupuestaria). Incluso cuando su intención declarada es mejorar las actuaciones, no están seguros de tomar su objetivo declarado por su valor nominal. Realmente, la palabra "evaluación" por sí sola es suficiente para desencadenar una reacción negativa y esto puede obstruir el camino y dificultar la obtención de datos.

Antes de que una propuesta o programa se financie y ponga en funcionamiento, los partidarios creen en el mismo --no han visto sus defectos en la medida en que alguien que haya participado o puesto en práctica realmente el programa los ve. Saben que su propuesta es buena y trabajarán para persuadir al analista.

En términos egoístas, un ejecutivo tiene interés en que se lleve a cabo un buen análisis previo al programa. Desea considerar cuidadosamente las alternativas antes de proponer un programa nuevo/ tiene una mayor posibilidad de realizar el programa aprobado de esta manera. Cuando evalúa cuidadosamente los programas existentes, entonces el balance entre el personal ejecutivo puede ser contrario a la evaluación, especialmente si los resultados pueden convertirse en un tema de dominio público. Esta puede ser una razón de por qué la evaluación es aún una de las actividades menos convincentes en el sector público y en el sistema federal. Igualmente, no es sorprendente que el congreso pueda haber jugado o juegue un papel más importante que la rama ejecutiva en la promoción de las actividades evaluatorias.

Consideramos primero la evaluación como ayuda a un decisor externo en la tarea de asignación de los recursos. Seguimos con la consideración de la evaluación para la realización de mejores, para el consumo interno. Después de esto, consideramos algunos problemas de experimentación y concluimos después con unas cuantas observaciones sobre la evaluación en general.

Evaluación para influir en la asignación de recursos

La evaluación destinada a influir en la asignación de recursos se proyecta con el fin de calcular el valor o la eficiencia de un programa o proyecto en curso para ayudar a determinar su financiación (o, posiblemente, otros recursos) que deberían atribuirsele. A veces, supone una elección entre el uso de fondos para proseguir o finalizar un programa, aunque lo más frecuente es la decisión para asignar recursos en el margen, aumentándolos en los programas que parecen estar funcionando bien y recortando, o no aumentando, los restantes.

El problema de los criterios supone aquí un tropiezo en la misma medida en que lo es en todos los análisis. Calcular el valor relativo del efecto o eficiencia de un programa puede compararse con las necesidades no localizadas, con objetivos de rendimiento especificados, con un rendimiento anterior del programa, con las ejecuciones de los programas compitiendo por los mismos activos y con sus propios costes.

Para mostrar las mejoras, a veces los gobiernos hacen uso de la comparación con el rendimiento anterior. Esto puede resultar engañoso. Un programa sanitario que cura el 70 por 100 de los casos este año pero sólo el 20 por 100 el año pasado puede considerarse bueno sobre esa base, pero no si las nuevas técnicas que posibilitan el aumento curan el 98 por 100.

La evaluación para comparar programas alternativos como todo análisis político necesita tener muy claros los objetivos de los programas considerados y de las vías para medir su consecución. Si

pueden elegirse medidas de eficiencia que sean clara y estrechamente relacionadas con los fines de los programas, la evaluación puede ser completa. Por ejemplo, el objetivo del proyecto "Poner en libertad bajo fianza al preso" de la Fundación Vera es demostrar que los presos libertados sin fianza comparecerían en el tribunal. Un cálculo para determinar el porcentaje de personas libertadas con su propia obligación y presentadas después al tribunal cuando se comparó con la fracción que compareció tras pagar la fianza, demostró la viabilidad del procedimiento. Como consecuencia de ello, se adoptó ampliamente por su evidente eficiencia en el coste.

Las metas de los programas en curso pueden ser tan ambiguas múltiples y conflictivas como los fines a que se enfrenta un analista político cuando aborda prácticamente cualquier estudio. Es, por lo tanto, necesario para el evaluador pedir una clara definición de los fines y prioridades entre ellos, antes de que pueda actuar convenientemente.

Por ejemplo, un programa para trabajar con grupos de adolescentes puede asegurar que su función es mejorar el "comportamiento social" de sus participantes. El éxito se mediría en términos de lo que los participantes hacen de un modo diferente después de haber participado. Pero ¿qué comportamiento social debería medirse: el número de detenciones?, ¿el número de sentencias?, ¿el número de querellas?, ¿el tiempo en casa después de las diez horas?, ¿la asistencia escolar?, ¿la asistencia a las sesiones del programa? Parte de la respuesta descansa en las difíciles realidades de tiempo, dinero y acceso. Decidir qué aspectos utilizar no es probablemente una tarea fácil. Elegir un aspecto o una combinación de ellos sin un conocimiento de cómo están relacionados, probablemente sea engañoso. Otra consideración es la utilización de lo que la evaluación se propone: ¿qué grado de éxito es necesario? La medición es algo deseable, al igual que la elegancia y la precisión, pero pueden tener que utilizarse indicadores de éxito no estrictamente cuantitativos, incluyendo las opiniones de profesores, funcionarios perezosos, policía, almacenistas del vecindario, y así sucesivamente. Un perfil o tarjeta formada por un gran número de indicadores puede que tenga que ofrecerse a quienes toman decisiones y formulan juicios.

Desgraciadamente, puesto que es tan difícil el tema de los criterios, los proyectos de evaluación tienden a medir la carga del trabajo antes que el verdadero rendimiento. Las medidas del esfuerzo del trabajo pueden no ponernos de manifiesto si el programa está logrando sus objetivos o con qué perfección lo está haciendo. Nos informan de que alguna labor se está realizando, pero sin clarificar si es lo que el programa desea hacer. Archibald (1970) asegura:

Por ejemplo, un programa de formación de mano de obra está ofreciendo formación a 200 de los más pobres al año con un coste de 1.000 dólares por individuo a forma --dicho en números redondos. Otro proyecto muy similar consiste en formar anualmente a 400 sujetos, siendo el coste total igual, razón por la cual este proyecto funciona sólo con 500 dólares por individuo. Admirable --funcionan con el doble al mismo coste. Pero eso no es una evaluación porque el objetivo no es disponer de la gente a través de un programa de formación --cabe presumir que es aumentar sus posibilidades de lograr un empleo y mantenerlo.

Para decidir qué proyecto da más por su coste, uno tiene que saber cuántos obtienen empleos. Incluso cuando se dispone de esa información, puede todavía o ser fácil decidir entre ambos proyectos. Digamos que un 60 por 100 obtiene empleo en el que trabajan 200 alumnos a un coste de 1.000 dólares por alumno, y un 40 por 100 en el que trabajan el coste de 400 a 500 dólares por alumno. Esto significaría que 120 graduados del primer proyecto obtienen empleo y 160 del segundo proyecto. El coste total de ambos proyectos es el mismo, por lo cual si el coste por empleo obtenido es el criterio, habría tenido que optarse por el de 400 alumnos a un coste de 500 dólares cada uno. Este proyecto ofrece más empleos, pero proporciona también mas "fracasos", 240 no obtienen empleo, en comparación con sólo 80 del otro proyecto. Podemos estar preocupados por los efectos del fracaso, las consecuencias de las expectativas suscitadas y el no satisfacerlas. Si estamos dispuestos a invertir algo en evitar esto, podría preferirse el proyecto de 1.000 dólares por alumno.

Este es un ejemplo muy sencillo; existen otras cosas que deberían tenerse en cuenta. Por ejemplo, quizás el proyecto de 1.000 dólares por alumno esté tratando con una clientela diferente --que, digámoslo, es más desventajoso. O bien, si los proyectos están funcionando en localidades diferentes, puede haber importantes diferencias en el mercado de trabajo. Asimismo, probablemente, se quiera considerar durante cuánto tiempo se mantienen los empleos.

Si la decisión a tomar es del tipo si o no, es decir, proseguir o suspender el programa, en tal caso, la evaluación debería proyectarse con el fin de comparar costes y beneficios del programa con costes y beneficios de no contar con dicho programa. No es seguro suponer que tanto costes como beneficios, de carecer del programa, sean cero.

Como cabría esperar, los análisis coste-beneficio y coste-eficiencia pueden aplicarse en este caso. Si los resultados del programa pueden expresarse en dólares (para un programa de rehabilitación profesional aquellos pueden ser los ingresos vitalicios o los rendimientos imponibles), entonces la diferencia

resultante entre beneficio y costes del programa puede emplearse para ayudar a tomar decisiones de asignación entre el programa sometido a evaluación y los restantes usos que pueden darse en dólares (incluyendo la reducción tributaria). De una manera similar, las comparaciones coste-eficiencia pueden hacerse entre programas para los cuales es posible utilizar las mismas medidas de eficiencia (por ejemplo, el incremento porcentual medio en las pautas de lectura de los tests habituales para comparar programas alternativos que hacen uso de ayudas docentes de todo tipo). Por lo general, los métodos a utilizar en este caos para las comparaciones son los mismos que los utilizados en otros lugares, pero existen complicaciones nuevas peculiares de la naturaleza en curso de los temas de análisis que es probable que sean fuertemente intensivas en "popularidad".

Evaluación para mejorar el funcionamiento

La evaluación para mejorar el funcionamiento se hace con frecuencia internamente, puesto que su objetivo es estudiar cambios posibles en el programa con la vista puesta en la mejora de la ejecución, no para ver cómo está funcionando el programa en comparación con programas similares o en cualquier sentido absoluto. Un proyecto de evaluación concreto puede, naturalmente, ofrecer información tanto sobre el efecto global como sobre la mejora del programa. Por ejemplo, de acuerdo con Wholey (1972):

La evaluación Westinghouse del programa Head Star produjo, por lo tanto, hallazgos generalmente negativos. Estos, no obstante, no redujeron significativamente el nivel presupuestario del Head Star. Poderosos distritos electorales habrían luchado contra cualquier reducción en los fondos destinados al Head Star. Los resultados que parece haber producido la evaluación Westinghouse del Head Star son: 1) el "arraigo" del programa --no deberían buscarse niveles crecientes de financiación; 2) la desviación de algunos fondos del programa Head Star hacia un programa experimental de desarrollo infantil, "variación programada", diseñado para probar si existen mejores aproximaciones que las utilizadas en el programa nacional Head Star, y 3) la reducción de la proporción de fondos del Head Star destinadas actualmente al Head Star estival (el componente del programa con el menor valor aparente).

La decisión de que la evaluación está destinada a ayudar debe considerarse en el proyecto del análisis. Levine y Williams (1971) clarifican esto de la siguiente forma:

Por ejemplo, un proyecto de evaluación podría diseñarse para ayudar a los gestores de un programa a determinar lo que garantiza los mejores proyectos; supuesto que sea así, sería una evaluación mejorada. O bien podría trazarse para ayudar

a los gestores del programa a determinar los proyectos que deberían suspenderse; de ser así, sería una evaluación de asignación. Aunque en una descripción breve, ambas evaluaciones podrían resultar ciertamente muy similares, serían en realidad bastante diferentes si se proyectaran adecuadamente para hacer frente a las distintas necesidades de la decisión. La proyectada para mejorar debería concentrarse en la mezcla de técnicas dentro de los proyectos. El objetivo sería descubrir qué técnicas trabajan mejor bajo ciertas condiciones, con la idea de difundir el uso de las mejores técnicas en las condiciones adecuadas. El proyectado para determinar el valor relativo se concentraría principalmente en los proyectos peores, e intentaría comparar los costes de rescisión con los de seguir en funcionamiento. Sin embargo, frecuentemente, resultaría económico considerar cuidadosamente ambos conjuntos de factores con el fin de ofrecer inputs a ambas clases de decisiones.

La evaluación destinada a mejorar el funcionamiento establece ciertos requisitos para la recogida de datos: 1) los datos recopilados deben ser pertinentes para los objetivos administrativos --así, los administradores y operadores del programa tienen probablemente que gozar de una instrucción considerable al señalar los datos a reunir; 2) los datos deben procesarse, interpretarse y retroalimentarse a los responsables de las operaciones con suficiente rapidez como para resultar útiles. La calidad de una evaluación depende realmente en gran medida del diseño del programa a evaluar. Un buen diseño para fines de evaluación favorece la recogida de una información correcta y a propósito.

El tipo de datos pertinentes para mejorar el funcionamiento puede ser de bajo nivel, rutinario y de corto alcance --cosas tales como el número de admitidos y su clase, el número de retirados y el número de los que lo finalizan con éxito. Estos conceptos pueden servir como indicadores o aproximaciones de conceptos más abstractos --las cifras vinculadas a solicitudes, admisiones y asistencia pueden mostrar el atractivo del programa, la eficiencia del reclutamiento y el grado de interés. La evaluación puede cumplir aquí su papel más útil. Citando nuevamente a Wholey (1972):

En los últimos años, los miembros del grupo de evaluación del Urban Institute se han persuadido cada vez más de que la evaluación principal de los resultados (en términos de las decisiones realmente influenciadas) puede ser la realizada con suficiente detalle para descubrir los efectos de cambios operativos dentro de los programas en funcionamiento. Muchos gestores de programas quieren realmente conocer los que mejor funcionan y en qué condiciones. Existe un mercado, un uso, para esta clase de información detallada para evaluación.

A veces existen ventajas en disponer de una evaluación de esta clase realizada externamente, ya que puede poner en claro que el objetivo consiste en mejorar el programa y no en suspenderlo. Los extraños atribuyen frecuentemente más peso a la administración del programa que los propios trabajadores del mismo. también los participantes y empleados pueden sentir que es más liberador debatir los problemas y defectos de un programa con un asesor externo al proyecto que con alguien estrechamente vinculado a la dirección del mismo. Aun cuando el asesor externo puede saber más sobre los tests y la medición y el diseño de la investigación que los directores de investigación de un programa u organismo, la posición de auxiliar en la tarea puede darle una credibilidad ante el director del programa y sus superiores de la que un asesor interno carece (Haggart y Rapp, 1973).

Si se sugieren cambios en un programa, es necesario un modelo sobre la forma en que éste funciona, un medio de describir el curso del programa desde la entrada hasta el resultado, a partir del cual pueden calcularse los efectos de las variaciones. haciendo uso de un ejemplo de Weiss (1966) para ayudar a clarificar esto, si un proyecto que ofrece ayudas a las chicas identificadas como potencialmente conflictivas, como un medio de reducir el comportamiento delictivo, intenta mejorarse, es necesario conocer la cadena causal a través de la cual se espera que el input consejo alcance su meta. Se necesitan respuestas a preguntas tales como" ¿Se logra la mejora cambiando la propia imagen de la chica?, ¿ofreciendo otras oportunidades para la autoexpresión?, ¿fomentando una mayor realización en el trabajo escolar, ¿ofreciendo modelos para un comportamiento alternativo?, o ¿de qué forma? Es necesario un modelo de cómo funciona el programa para poner en claro lo que hemos de considerar. Si conocemos lo que se supone que ocurre en varias etapas intermedias de un programa, podemos descubrir dónde las cosas marchan erróneamente y poner de manifiesto la etapa en la que son necesarios los ajustes.

Experimentación

Los experimentos ofrecen una oportunidad para llevar a cabo una forma mucho mejor de evaluación, pero, al mismo tiempo, sugieren a los analistas un nuevo conjunto total de problemas. Estos problemas son de dos tipos: los asociados a las técnicas del diseño experimental y los vinculados al contexto público en que debe realizarse el experimento. Como parte de su formación, la mayoría de los analistas o, al menos, los formados en estadística, han aprendido algo acerca de cómo manejárselas con los problemas experimentales del diseño. Los problemas de diseño para experimentos sociales a gran escala, tales como el esfuerzo realizado por HEW 1973, para estudiar los planes de seguridad serial alternativos, son, sin embargo, probablemente, mucho más complejos que los que se incluyen en los textos escolares. De todos modos, son demasiado especializados para ser tratados en este

libro. En consecuencia, no discutiremos este nuevo aspecto.³ Pero los problemas del segundo tipo son también probablemente difíciles, incluso puede que más que los primeros.

Los políticos y funcionarios de programas a menudo no se dan cuenta de lo que es necesario si un experimento quiere realizarse adecuadamente. Por ejemplo, la asignación aleatoria de participantes (o comunidades) al tratamiento resulta fundamental para el diseño experimental. No obstante, un administrador de programas puede pedir enfáticamente responder al imperativo ético para servir a quienes más necesitan el servicio. ¿Qué puede tener de malo, puede argüir, elegir a la gente más necesitada de vivienda para participar en un experimento de asignación de vivienda? Los problemas jurídicos pueden también surgir por la distribución desigual de los servicios ofrecidos públicamente; los tribunales no han ofrecido todavía una guía clara sobre lo que Rivlin (1971, pág. 110) califica como "la equidad de las desigualdades intencionadamente creadas para fines experimentales".

Son necesarios grupos de control, pero este aspecto puede ser para el analista difícil de vender. La falta de proyectos adicionales puede tropezar con la consecución de cualquier grupo de control. Existen dificultades para obtener la cooperación de un público al cual no se ofrece nada. Y, naturalmente, el dinero y el personal son siempre escasos.

Sholey (1972) indica otra dificultad:

El tiempo presenta siempre un gran problema para el evaluador de programas experimentales. Tan pronto como hay suficiente apoyo legislativo para realizar una serie de experimentos, puede disponerse de suficiente apoyo para poner en funcionamiento ese programa en toda la nación. La preocupación por la aprobación de las normas antes de que los experimentos hayan tenido tiempo para producir resultados dignos de confianza puede conducir a presiones para la relajación de los primeros hallazgos, menos seguros. Los experimentos de la New Jersey Income Maintenance sufrieron esta presión. Algunos de los primeros resultados tentativos del estudio fueron relajados con disgusto y difíciles requisitos. Si la experimentación va a convertirse en un vehículo importante de la investigación política, entonces han de descubrirse vías para prever y tratar con estos tipos de presión.

³ Las referencias generales aquí son Cochran y Cox (1975) y Campbell y Stanley (1966). Para el experimento del seguro sanitario ver Newhouse (1972). Para una experimentación social en particular, ver Riecker y Boruch (1974), Brewer (1973b), y Archibald y Newhouse (1980).

Como hemos dicho, la asignación aleatoria de proyectos a los grupos experimentales y de control resulta esencial, pero los operadores de programas o las presiones políticas pueden invertir esto. En el anhelo por demostrar que el programa estudiado mediante un experimento tendrá éxito, pueden elegirse los participantes de entre los que con mayor probabilidad resulten beneficiarios del programa. Asimismo, incluso aunque el analista pueda a veces ejercer el control sobre quien participa en el programa, es probable que no tenga prácticamente ninguno sobre quienes se marchan.

A pesar de estos problemas, añadir un cierto tipo de comparación incluso si es tan solo un equivalente parcial del grupo expuesto, puede prestar ayuda poniendo al descubierto análisis de lo que sucede. Sin ningún grupo de comparación, el analista debe estudiar otras explicaciones de lo que ocurre, distintas del programa, que no resulten excluidas por el diseño experimental.

Además de la posibilidad de ser injusta, la experimentación es, por su propia naturaleza, arriesgada. Muchos buenos experimentos consecuentemente fracasarían. El fracaso cuesta tiempo y dinero y, si su objetivo es el público, puede conllevar otros costes. Estos pueden extenderse a los administrativos. Rivlin (1971, pág. 112) señala:

Otra reserva sobre la desabilidad de la experimentación social se refiere a la integridad con que serán divulgados los resultados experimentales. A nadie le gusta fracasar. Correcta o equivocadamente, el administrador de un proyecto experimental de éxito recibirá más aplauso y mayores oportunidades de prosperar que el de un proyecto carente de éxito. En estas circunstancias, no habrá un aliciente suficiente para un cierto engaño, para elegir los instrumentos de medición más favorables, para "perder" los registros de los hijos que fracasan o de los pacientes que fallecen, para instruir a los participantes sobre lo que decir al evaluador o ¿cómo superar la prueba?

Un proyecto denominado demostración es, por su propia naturaleza, un experimento de clase, pero a menos que los resultados puedan evaluarse adecuadamente, todo lo que puede demostrarse es que es posible gastar fondos públicos en una forma concreta. Una demostración no es un experimento auténtico a menos que se ejerza el control sobre los inputs, participantes y parámetros, y los resultados se midan con el fin de determinar hasta que punto el programa hace lo que estaba proyectado que hiciera.

A pesar de las diferencias, la evaluación de experimentos tiene generalmente mucho en común con la evaluación de los programas en curso. En ambos casos, el analista debe recordar que la evaluación es una disciplina aplicada, que se realiza para

prestar ayuda en un problema corriente. No es, al menos principalmente, investigación científica que persigue la verdad, sino que trata de ofrecer una base para la acción. Como consecuencia de ello, el evaluador debe resistir la tentación de buscar respuestas a las preguntas que le interesan, pero que no están situadas muy arriba en la escala de preguntas que quien toma decisiones precisa responder.

Observaciones

Hasta fechas recientes, las consideraciones políticas prevalecían en los experimentos sociales a gran escala. Esto ha cambiado; ahora tenemos el experimento del impuesto negativo sobre la renta del OEO de New Jersey. Los experimentos de sostenimiento de los ingresos de HEW en Vermont, Gary, Seattly y Denver, los experimentos de OEO con los contratos de ejecución en la educación elemental y el programa más amplio y complejo de experimentos sociales de la historia nacional, el House Assistance Supply Experiment, del HUD, un ensayo por diez años de subvenciones a la vivienda para familias de bajos ingresos (Lowry, 1980).

Algunos de los problemas que existen con la evaluación son señalados por Wholey (1972):

Volvamos ahora a algunos de los problemas reales que surgen en la búsqueda de una evaluación provechosa. Desde el punto de vista de quienes toman decisiones, la evaluación es un arma peligrosa. No necesitan la evaluación si ésta ofrece respuestas "erróneas" sobre los programas en los que aquellos están interesados. (*) Por otro lado, quienes toman decisiones tienen ventaja frente a los evaluadores en la formulación de preguntas pertinentes, ofreciendo respuestas a propósito a un coste razonable. La evaluación digna de confianza, válida, es muy difícil de realizar y puede costar mucho dinero. (‡) Los

(*) En algunos casos, las presiones políticas simplemente no hacen caso de la evidencia empírica sin la formalidad de un argumento metodológico. Aquí, el único recurso abierto al evaluador es publicar los resultados y confiar en que otros mejor informados o menos presionados ante problemas similares harán uso de los resultados.

(‡) La evaluación del Stanford Research Institute de programa de la Office of Education Follow Through, por ejemplo, tiene ya un coste aproximado de 57 millones de dólares.

evaluadores tienen problemas reales a la hora de detectar las conexiones causales entre inputs y outputs --y en hacerlo de una forma lo bastante adecuada como para que sea útil a quienes toman decisiones. La estructura de un programa puede ejercer una influencia importante sobre la viabilidad técnica de aislar los efectos del mismo frente a los efectos de otras fuerzas, a menudo más eficaces, que no están bajo control del programa. En la medida en que un programa se presenta como un experimento controlado, por ejemplo, las posibilidades del evaluador de considerar separadamente las relaciones causales pueden ser mayores.

Nuestras revisiones han descubierto en los estudios típicos de evaluación de programas federales ciertas características del diseño que limitan fuertemente su seguridad y utilidad:

- a) Han sido esfuerzo aislados, únicos, cuando lo que necesitamos son evaluaciones continuas de programas.
- b) Han sido realizados en términos de programas nacionales y son muy escasos en datos procesados.
- c) Han sido estudios con muestras reducidas trabajando con promedios brutos, cuando precisamos estudios lo bastante grandes para permitir el análisis de las amplias variaciones que sabemos existen en los costes y rendimientos entre proyectos dentro de los programas.

Como en otros análisis realizados para organismos públicos, el evaluador debe conocer el proceso de toma de decisiones y las limitaciones operativas de una evaluación concreta. La estructura organizativa de la agencia puede ser un determinante importante de los servicios ofrecidos. El tema del asesor interno-externo puede ser particularmente embarazoso, especialmente cuando los resultados de la evaluación son negativos. El personal del proyecto, la organización patrocinadora y los grupos con un "interés especial" en la posibilidad de éxito pueden reaccionar hostilmente, lo cual puede transformarse fácilmente en un rechazo personal a las capacidades investigadoras del analista interno y de los intereses personales.

La evaluación puede estar llena de peligros latentes. Frecuentemente, un nuevo programa o actividad pública que forja un proyecto para ofrecer un servicio se presenta por medio de una nueva estructura organizativa o por un cambio en la antigua. Al evaluar, hemos de decidir cuándo es el servicio y cuándo la organización lo que deseamos evaluar. Un programa puede evaluarse en términos de su eficiencia y costes. No obstante, una estructura organizativa no puede evaluarse exclusivamente por su éxito en el desarrollo de un programa concreto. Su capacidad para detectar fracasos del programa y mantener el éxito a lo largo del tiempo deben también tenerse en cuenta. Por esta razón, una organización

no debería evaluarse por un fracaso inicial del programa sino por su capacidad para "aprender" a partir del fracaso y modificar las políticas del programa para mejorar el funcionamiento.

Consideremos la ejecución por contrato en materia educativa. Un distrito escolar llega a un acuerdo contractual con una empresa independiente para ofrecer todo o parte de un servicio educativo en una escuela o un conjunto escolar. El contratista paga una suma determinada por mejorar el rendimiento escolar en cierto grado o por elevar los niveles hasta un cierto punto. Si no tiene éxito no se le paga, midiéndose el éxito o el fracaso en términos de algo así como la puntuación en lectura y matemáticas mediante tests estandarizados de realización.

Esta práctica significa claramente una aproximación organizativa diferente a la educación de la escuela pública. No obstante, no siempre ha sido así evaluada. Debido a que la mayor parte de los contratistas han resaltado la innovación, haciendo amplio uso de técnicas tales como el aprendizaje programado, la instrucción informatizada y los medios audiovisuales, la aproximación del rendimiento contractual se valora a veces como una aproximación innovadora de la enseñanza, basada en la tecnología. En otros casos, se considera como un instrumento para introducir cambios, a sustituir tan pronto como los profesores aprendan lo bastante para manejar ellos mismo tales programas de tecnología intensiva. En este último caso, se tiende a juzgar el programa como un éxito si el sistema escolar adopta los métodos usados por el contratista.

Por ejemplo, la OEO evaluó el rendimiento contractual mediante la selección de seis empresas privadas. Estas empresas fueron escogidas aparentemente porque emplearon una variedad de aproximaciones educativas diferentes, incluyendo diversas "herramientas" educacionales, incentivos y métodos de formación del profesorado. Las fórmulas de pago fueron esencialmente idénticas, considerándose el rendimiento para el período de un año. Las conclusiones fueron que los estudiantes que participaban en los experimentos no lo hacían mucho mejor que los que no participaban. Como consecuencia de ello, es dudoso que esta evaluación nos diga algo sobre el ajuste organizativo y la adaptabilidad, puesto que las empresas se vieron esencialmente limitadas a usar los mismos métodos durante el año.

Otro defecto puede ser atribuir un resultado a una causa equivocada. El incentivo puede producir una diferencia. Consideremos un caso de ejecución contractual en el que el contratista va a ser retribuido sobre la base del número de estudiantes del programa que superen las normas nacionales en dos test habituales. Para maximizar sus beneficios dividió a los estudiantes en tres grupos: uno, alto, que no necesita más instrucción para superar la dividió; un grupo bajo, desahuciado, y otro medio, para el cual el tema resultad dudoso. Se centró en el

grupo intermedio, dejando a los otros dos básicamente sin instrucción. El grupo medio se comportó bien; los otros dos se hundieron tanto como para que el promedio global no aumentara. Los educadores juzgaron este caso como un fracaso para la contratación por rendimiento.

Otra posibilidad de error está en determinar el ganador real a partir del programa. Archibald (1968) señaló esto de la siguiente manera:

¿Debería prestarse alguna atención en la evaluación a lo que alguien gana? en ciertos casos, la formación puede ser poco más que un subsidio indirecto al patrón. Estoy pensando en aquellos casos en los que el patrón emplea a grupos minoritarios y ofrece generalmente formación para el empleo. La fabricación de joyería ofrece un buen ejemplo ... Los fabricantes de joyas de la ciudad de Nueva York han sido, durante largo tiempo, asalariados de Puerto Rico; realmente, algunas tiendas son de hispanohablantes. Los chicos se forman durante un cierto número de semanas o de meses ... siendo después empleados a un salario legal mínimo, puesto que los patronos arguyen que todavía tienen que dar formación considerable en tareas específicas que los chicos han de llevar a cabo. El patrono es subvencionado a través de la OJT de Nueva York, de tal forma que ni siquiera paga todo el salario mínimo legal durante el primer período de empleo. El resultado final es que los chicos gozan de cierta ayuda social y facilidades de empleo, pero el patrono obtiene una importante subvención procedente del ahorro en la formación de sus empleados y un subsidio en las nóminas. Si evaluamos solamente el porcentaje de colocación en un caso como este, la eficiencia del programa acabaría siendo groseramente valorada en exceso.

otro aspecto de los programas de formación justifica la atención prestada al proceso evaluatorio (Archibald, 1968):

No he visto a nadie preocuparse, bien a nivel operativo o de planificación, por las consecuencias que para los clientes tiene el fracaso de los programas. Si las "carreras nuevas" acaban siendo empleos muertos, si la formación profesional prueba su irrelevancia a inutilidad y el cliente ha recibido las promesas de la agencia sinceramente, la principal consecuencia será un aumento de la alienación. Los costes de estos fallos pueden ser especialmente altos si es correcto el punto de vista de que el problema para el pobre es ya el de no relacionar medios y fines en el sentido de aplazar la recompensa y dirigir el esfuerzo hacia una meta. El fracaso de un programa de formación les probará nuevamente la carencia de una relación predecible entre medios y fines institucionalizados. Una evaluación debería tener en cuenta los costes y los riesgos de fracasar. Pienso que esto

constituye una seria preocupación para un programa como New Careers.

En la evaluación, más que en la mayoría de los análisis políticos, el analista puede tener que hacer cargo de los problemas de aceptación cuando los resultados son negativos, indicando que el programa se finalice o reduzca, debido a que esto significa que todo o parte del personal perdería sus empleos, siendo de esperar la resistencia. Weiss (1966) observa que si el analista "puede hacer entender la idea de que 'nosotros' estamos valorando el proyecto" y que los resultados negativos pueden tener implicaciones constructivas para mejorar el programa, puede evitar el tener que preocuparse más tarde. Un analista debe reconocer que si su evaluación indica que el proyecto es deficiente y que no hace lo que se supone que hace, ha creado un problema para el director y su personal. desde su punto de vista, la causa más probable de ese problema y la que primero debería estudiarse e si hay algo erróneo en la propia evaluación.

Notas a modo de conclusión

En un sentido informal, los programas de evaluación pública son una actividad que casi todos nosotros aprobamos. Sin embargo, comenzamos a verificar que son necesarios intentos más sistemáticos para determinar los efectos, positivos y negativos, de tales programas por parte del público, además de que el objetivo no debería ser simplemente clasificar los programas con la perspectiva de suprimir el peor, sino, lo que es más importante, de mejorar los que merecen continuarse.

La evaluación cuenta con una literatura en rápido crecimiento; además de las referencias citadas, otras referencias podrían ser Guttentag (1971), Nay et al. (1976), Rossi y Williams (1972), Struening y Guttentag (1975), Wholey et al. (1970) y Weiss (1972). El informe de Hatry et al. (1977) es una tentativa excelente de proporcionar una guía práctica a los funcionarios locales interesados o bien en desarrollar una capacidad para evaluar programas o bien en mejorar su capacidad actual.

1
2
3

4
5
6

