PONENCIA III

Revista Investigación Educativa - Vol. 8 - n.º 16 - 1990 (P. 61-76)

EVALUACIÓN DE LOS EFECTOS DE LOS PROGRAMAS DE INTERVENCIÓN

por
Prof. Arturo de la Orden
Departamento MIDE
Universidad de Complutense de Madrid

No resulta inusual en la literatura y en los foros pedagógicos la identificación de la calidad de los programas educativos con la presencia o ausencia de ciertas características de contexto, de input o entrada y de proceso, sin una referencia expresa a los resultados, efecto o producto de tales programas. Es incluso frecuente comparar escuelas, clases, o cursos sistemáticos de acción educativa en términos de equipamiento, recursos físicos y humanos, ratio profesor-alumno o, si se prefiere, tamaño de la clase, estructura curricular, estrategias didácticas, etc., asumiendo acríticamente que tales variables y otras similares están claramente relacionadas con la calidad de la intervención y que conocemos el sentido y la magnitud de tal relación. En estas circunstancias, los efectos reales de los programas no precisan ser determinados, medidos y evaluados o, quizá, no son susceptibles de medida y evaluación. Sin embargo, parece evidente que si deseamos saber si la intervención pedagógica afecta, y como afecta, a los estudiantes, la vía más directa consiste en analizar los resultados, más que el presupuesto, las características de profesores y alumnos y los modos de interacción, aunque la consideración de tales variables nos ayude a explicar el sentido y la magnitud de los efectos. En general, los estudios de evaluación sumativa se justifican por su contribución crítica en el proceso de decisión para elegir una entre varias alternativas, en nuestro caso, de programas de intervención educativa. La decisión implica la elección del mejor tratamiento (programa) en términos de los criterios del decisor o, en palabras de Cahan y otros (1987), de la función de utilidad, uno de cuyos componentes es, obviamente, la eficacia relativa de los diversos tratamientos considerados. En otras palabras, siendo similares otros factores (consideraciones presupuestarias y logísticas, por ejemplo) el programa elegido será el más eficaz, es decir, aquel cuyos efectos se acerquen más a los

objetivos previstos. En consecuencia, la información cuantitativa sobre los resultados o efectos de cada programa es un factor de máxima relevancia para la elección del más adecuado, y la estimación de estos efectos se constituye en el propósito fundamental de todo estudio de evaluación sumativa.

Es necesario señalar en este punto que la medida y evaluación de los efectos de la intervención educativa no es una tarea simple, o más simple que otros procesos de búsqueda disciplinada de conocimiento. Se trata de una tarea de investigación en sentido estricto y no es más directa o lineal que otras tareas de investigación. Dar respuesta a las cuestiones de evaluación sin ambigüedades requiere la misma competencia y talento creativo que contrastar hipótesis científicas. Además, la evolución debe hacer frente a una millada de problemas prácticos que incrementan la complejidad del proceso. Tales problemas pueden ser controlados en los experimentos de laboratorio, pero resultan difíciles de tratar en las situaciones reales en que se mueve el evaluador.

MARCO DE REFERENCIA PARA LA DETERMINACIÓN DE LOS EFECTOS DE UN PROGRAMA

En esta ponencia sobre los efectos de los programas educativos y su evaluación, partimos de una concepción de la educación como «proceso sistemático de intervención tendente a optimizar la conducta humana con referencia a modelos considerados valiosos y deseables» (De la Orden, 1981 y 1985). Se trata, pues, de un progreso intencional que puede ser contemplado en una perspectiva sistémica y tecnológica (De la Orden, 1981 y 1985). Al caracterizar a la educación con atributos de sistema estamos postulando no sólo la posibilidad de intervenir y manipular las entradas (input) para regular el proceso y conformar los resultados (output) de acuerdo con el modelo de comportamiento valioso, implicado en las metas previstas, sino también la necesidad y viabilidad de evaluación y control del producto educativo bien mediante su contraste con los objetivos (eficacia de la intervención), bien mediante la comparación de los cambios (varianza) en los output con los cambios (varianza) en las entradas, cuyas relaciones pueden ser expresadas a través de una función de producción».

El producto educativo se identifica con lo que acontece al individuo como resultado del proceso de intervención sistemática que llamamos educación. Los procesos de intervención educativa suelen integrarse en «programas» o cursos sistemáticos de acción en función de unas metas u objetivos. Programa es un término genérico para designar un conjunto de tratamientos educativos coherentes, para su administración a diferentes unidades (individuos o grupos de individuos) de un sistema determinado en función de sus características o de forma contingente a su conducta. En el caso más simple, el programa puede consistir en un tratamiento único que ha de ser aplicado a todas las unidades. Cualquiera que sea la complejidad de un programa su aplicación se justifica por sus potenciales efectos, es decir, por

los cambios previsibles en las unidades del sistema, cuya evaluación y medida es un prerrequisito para tomar decisiones acerca de su continuación, sustitución, modificación o supresión.

Aplicando los criterios apuntados por Cahan (1987) para definir los efectos de una determinada política social o educativa (en términos prácticos identificable con un programa) a nuestro problema, podríamos decir que la determinación de los efectos de un programa de intervención requiere la especificación de tres elementos:

— *Los límites del sistema.* La especificación del sistema está relacionada con la naturaleza de las decisiones a tomar. En el caso más general el programa se aplica a todas las unidades del sistema nacional y, en consecuencia, sus efectos se definen para todo el sistema. En otros casos, los efectos se asocian a decisiones que sólo afectan a una zona, a una escuela e incluso a una clase y, por tanto, han de ser definidos para estos subsistemas.

— El *producto relevante.* Dado que el mismo programa puede afectar a diferentes resultados (variables dependientes) en el mismo sistema, la definición de sus efectos exige necesariamente la especificación del producto relevante para su evaluación. Por tanto, cada programa está asociado a un conjunto de efectos, cada uno de ellos definido en función de un producto diferente. Los resultados intentados por un programa educativo a nivel del sistema contemplado se identifican normalmente con los parámetros de las distribuciones de variables definidas a un nivel inferior de agregación.
Así, por ejemplo, el resultado de un programa dado, a nivel de escuela puede representarse por la media o la varianza del centro de una variable a nivel de clase o de individuo, como el rendimiento escolar.

— *Un punto de referencia.* La conceptualización del efecto de un programa requiere también la especificación de un punto de referencia. Normalmente este punto de identifica con el valor del producto bajo una condición de control (otro programa, el no tratamiento, un placebo, etc.). Por tanto, los efectos del programa serán diferentes con respecto a diferentes condiciones de control.

En este marco y siguiendo a Cahan (1987), se puede proceder a la definición formalizada de los efectos de un programa de intervención educativa en los siguientes términos: Dado un conjunto de J programas alternativos (que incluye, obviamente, una condición de control), el efecto A (P_j) del programa P_j (j = 1, 2,... J-1) en el resultado o producto O en un sistema dado (sistema nacional, subsistema regional, ciudad, escuela, etc.), con respecto a la condición de control C, se define como la diferencia entre dos valores del producto en dos condiciones determinadas: O/P_j, representando el valor de O tras la aplicación del programa P_j, y O/C representando el valor de O en el mismo sistema tras la aplicación del programa C de control. Esta diferencia se expresa así:

$$A\ (P_j) = O/P_j - O/C$$

es decir, la ganancia, o pérdida, en el producto O asociada con la aplicación del programa P_j, en relación con el valor del producto asociado con la aplicación del programa de control.

En esta perspectiva, el producto final de la evaluación de programas puede ser representado por una matriz. Las filas de la matriz corresponden a los diversos programas alternativos y las columnas a los distintos productos. A cada combinación de programa y producto corresponde un efecto y, dado que los efectos están expresados en la métrica del producto, la comparación directa sólo es posible entre los efectos de distintos programas con respecto al mismo producto, mientras que los efectos del mismo programa en diferentes productos no son directamente comparables. En el citado trabajo de Cahan (1987) se sugieren algunos vías de acercamiento a la solución de este problema técnico cuyos detalles rebasan los objetivos de esta ponencia.

EFECTOS DE LOS PROGRAMAS EDUCATIVOS RELEVANTES EN LA PERSPECTIVA DE SU EVALUACIÓN

Parece claro que la primera decisión, una vez establecida la necesidad de evaluar los efectos de un programa educativo, es determinar con precisión el objeto de la evaluación, es decir, cuales son los efectos de interés en la perspectiva evaluadora adoptada. Si, por ejemplo, la evaluación intenta obtener información relevante como base para tomar decisiones en el contexto de una política educativa, parece que la selección de las medidas de producto debería realizarse en el marco del diálogo entre evaluador y usuario de los resultados evaluativos. En todo caso, la valoración de la relevancia de los productos del programa como objeto de evaluación debe realizarse considerándolos en la perspectiva de un continuo medios-fines (cada producto es medio para conseguir otro y, asimismo, es un fin de los que le preceden) que culmina en una meta terminal concebida como finalidad deseada y prevista.

Según Cooley (1974), los productos educativos más importantes serán, en general, aquellos que se considera afectan o pueden afectar directa o indirectamente al éxito y la satisfacción en la vida adulta. El problema es que, como afirmaba en otra parte (De la Orden, en prensa), resulta extraordinariamente difícil obtener evidencia de las relaciones entre diferentes atributos del rendimiento educativo y la conducta posterior del sujeto en la vida. La identificación de predictores válidos de la conducta posteducativa como adulto permitiría su consideración como criterios para evaluar los efectos de los programas educativos, sin necesidad de recurrir a largos, complejos y costosos diseños longitudinales.

Un segundo criterio en la selección de variables adecuadas para la evaluación de los efectos educativos es la consideración de alguna base teórica y/o empírica justificativa de la presunción de que las medidas de ciertos productos pueden ser afectadas por las prácticas educativas implicadas en el programa de intervención.

Finalmente, es preciso seleccionar las variables de forma tal que se evite o se reduzca al mínimo la redundancia en orden a facilitar la descripción e interpretación de los resultados del estudio.

La aplicación de estos tres principios de selección de variables para la evaluación de los efectos de programas educativos, es decir, la validez predictiva, la justificación teórica o empírica y la parsimonia, llevaron a Cooley (1974) a sugerir como el criterio más importante el desarrollo intelectual general. Sin duda, Cooley se está refiriendo fundamentalmente a los programas regulares de educación integrados en el sistema formal. Concibe este desarrollo intelectual general, como un factor derivado de diferentes medidas cognitivas. Este factor medido en un determinado nivel educativo es el mejor predictor de realización académica en los niveles siguientes. Lo que un individuo es capaz de aprender en un momento es fundamentalmente una función de lo que ha aprendido previamente. Asimismo, este factor es el mejor predictor del tipo de vida postescolar de los alumnos. Por ejemplo, es el mejor predictor singular de la calidad de los logros profesionales. Por otra parte, y esta es una consideración crítica para su elección como indicador de producto educativo, el desarrollo intelectual general es parcialmente una función de la práctica educativa. Aunque el 50% de su varianza puede ser atribuido al desarrollo intelectual previo a la intervención educativa, el 50% restante puede vincularse a diferentes formas de práctica pedagógica.

Obviamente, hay otros factores de interés como expresión de resultados educativos. Para evitar, de una parte, redundancias y, de otra, las ambiguedades interpretativas de medidas de producto altamente correlacionadas, habría que llegar a un sistema de factores generales ortogonales. Quizá, la necesidad crítica en la evaluación de los efectos de la educación sigue siendo un esquema adecuado para representar el amplio y diverso espectro de resultados educativos y patentizar su valor de transferencia a la vida en general.

Además de los efectos educativos generales, la evaluación de programas de intervención no puede prescindir de los productos y efectos específicos de cada uno que, en ocasiones, pueden derrivarse del sistema general de factores representativos de los resultados de la educación formal.

Uno de los descubrimientos de la reciente investigación evaluativa es que los procesos reales de intervención educativa no son uniformes al aplicar los programas en distintas situaciones (escuelas, clases, estudiantes). En efecto, las variaciones en la aplicación de un determinado programa pueden ser tan grandes que lleguen, en algunos casos, a superponerse con las propias del programa alternativo, lo que dejaría carente de sentido el contraste de los efectos de los dos programas. Este hecho apunta a la necesidad de observar directamente los procesos de intervención implicados en los programas y de representarlos en esquemas multidimensionales, como apuntaba para los efectos mismos. El esquema de representación de los procesos en dimensiones generales debería ser la derivación natural de una teoría de la enseñanza o, si se prefiere, de la instrucción (Cooley y Lohnes, 1976). De este modo, al evaluar un programa estaríamos evaluando simultáneamente un modelo de

intervención educativa cuyas virtualidades de transferencia son patentes. Cooley y Lohnes (1976) en la obra citada proponen un modelo de aprendizaje escolar como guía para la determinación de los procesos de intervención educativa. Este modelo identifica cuatro dimensiones generales derivadas del conocido modelo de Carroll (1963): Oportunidad que el programa ofrece al alumno para aprender; grado en que el ambiente instructivo favorece la motivación para aprender; calidad de la estructura del currículum y eficacia de los acontecimientos o episodios instructivos. Considerando estas cuatro dimensiones en combinación con las destrezas y motivos con que los alumnos entran a la experiencia educativa, se explica la mayor parte de la varianza en los efectos del programa objeto de evaluación.

DISEÑOS PARA LA EVALUACIÓN DE LOS EFECTOS DE PROGRAMAS DE INTERVENCIÓN EDUCATIVA

Dados los tres dominios multidimensionales que sintetizan la varianza en el estado inicial de los alumnos, los procesos implicados en el programa y los resultados o producto del mismo, la evaluación de los efectos supone la elección de un diseño adecuado de investigación que permita obtener, analizar e interpretar la información relevante para formular un juicio de valor y tomar las decisiones apropiadas. Los diseños para la evaluación de programas pueden variar desde el más riguroso experimento de campo a la descripción naturalista, sin olvidar, claro está, la diversidad de diseños cuasi-experimentales o los correlacionales. El diseño determina la posibilidad de detectar y comprender los procesos y el impacto de un programa educativo, dada la multiplicidad de factores que influencian su operación y resultados.

Por otra parte, según Boruch y Wortman (1979), se ha contrastado reiteradamente que los efectos de programas innovadores en educación, aunque notables en ocasiones, generalmente no superan en mucho a los programas ordinarios. Parece, en consecuencia, necesario diseñar la evaluación de modo que puedan detectarse efectos de pequeña magnitud. En esta perspectiva, ofrecen la máxima garantía los diseños de experimentos aleatorizados. Sin embargo, la selección de un diseño apropiado a las características del programa y, al tipo de información necesaria para tomar decisiones sobre el mismo, debe apoyarse en el examen escrupuloso de las alternativas disponibles. Con esta premisa, al considerar el contexto y los condicionamientos prácticos de un programa en funcionamiento, la probabilidad de elegir un diseño experimental aleatorizado se convierte prácticamente en la excepción. Las condiciones en que la mayor parte de las evaluaciones se llevan a cabo, fuerzan con gran frecuencia a renunciar a los diseños ideales en favor de otras aproximaciones más viables. Esto ha permitido hablar de pluralismo en el proceso de diseño evaluativo. En este sentido, lo que teóricamente es un diseño, es decir, un conjunto de decisiones técnicas que vinculan los procedimientos de investigación a la función de dar respuesta a las cuestiones evaluativas, ha adquirido

importantes dimensiones sociales y políticas. La determinación de un diseño de evaluación del producto educativo no se limita, pues, a seleccionar una «receta de investigación» apropiada. Es necesario comparar alternativas, combinarlas según sus posibilidades y limitaciones, considerar los intereses de las diversas audiencias y clientes y la disponibilidad de recursos físicos y humanos. En cierto modo, igual que en el diseño arquitectónico, en el diseño de evaluación juegan un papel relevante los elementos artísticos.

Por otra parte, los tipos de programas, y los ambientes y contextos en que se desarrollan, condicionan las cuestiones y los métodos evaluativos, así como los factores organizativos y los recursos que, a su vez, son determinantes del diseño y de los procesos evaluativos. Sólo los programas experimentales, es decir, las experiencias piloto puestas en operación con la finalidad explícita de generalizar un tipo de intervención si se producen los efectos deseados, exigen, por definición, un diseño experimental para su evaluación. Como es bien sabido, los experimentos científicos son fundamentalmente intentos de identificar causas y evaluar sus efectos. Esto normalmente se consigue introduciendo cambios sistemáticos en uno o más factores —variables independientes—, en nuestro caso, diferentes modos de intervención educativa representados en diversos programas o variaciones de un programa, y observando si se producen cambios que covaríen con los primeros en otros factores —variables dependientes—, en nuestro caso, diversas dimensiones del producto educacional. Lo que es esencial en todo experimento es la acción de comparar. Los diseños experimentales implican la comparación de los valores de la variable dependiente asociados a la presencia de distintos valores de otra u otras variables. En el caso extremo, muy frecuente en la evaluación de programas, la variable manipulada (programa de intervención) toma valores dicotómicos (presencia-ausencia). La cuestión capital, claro está, radica en la confirmación de que los diferentes valores de la variable dependiente (producto) están asociados con diferentes valores de las variables independientes (dimensiones del programa). La comparación, en cualquier caso, se consigue a través de dos estrategias experimentales básicas: diseños intra-sujetos y diseños entre sujetos.

En los diseños intra-sujetos, la comparación se hace entre las modificaciones que se producen a lo largo de un período de tiempo dentro de un sujeto o un grupo de sujetos. Es decir, la variable dependiente se mide, al menos, después de actuar la variable independiente y, quizá, antes o durante la intervención experimental. Aunque los diseños intra-sujetos están muy vinculados a los estudios «N = 1» (Greenwald, 1976), también se utilizan con un grupo numeroso de individuos. No es, pues, necesario restringir su aplicación a unos pocos sujetos.

Los diseños entre sujetos también intentan comparar valores de la variable dependiente (resultados del programa) asociados con la presencia de diferentes valores de la variable independiente (modos de intervención pedagógica). Sin embargo, en estos diseños, la estrategia primaria es observar diferencias entre sujetos sometidos a diversos niveles de la variable independiente, es decir, entre los que han sido sometidos a un programa y los que no han sido expuestos al mismo. General-

mente, las comparaciones se hacen entre grupos y no entre individuos singulares, por lo que estos diseños se inscriben en la llamada investigación de grupo.

Existen muchos tipos de diseños tanto intrasujetos como entre sujetos. Para una descripción y valoración de los primeros puede consultarse el trabajo de Hersen y Barlow (1976) y para los segundos, especialmente para cuasi-experimentales, más adecuados en evaluación, sigue siendo una buena guía la obra de Cook y Campbell (1979) y la síntesis de Mahoney (1978).

Obviamente, además de los diseños experimentales y cuasi-experimentales, la evaluación de los efectos de programas de intervención educativa debe recurrir a otros planes de investigación que se ajusten al contexto y a las exigencias de las audiencias. Ésta es la situación más frecuente al intentar evaluar los programas reguladores de enseñanza, donde prácticamente todos los diseños evaluativos excepto el experimental sensu stricto pueden ser útiles. Los diseños correlacionales son los más frecuentes en estos casos. Sabemos que la covariación sistemática es un prerrequisito para la predicción —una función básica de la ciencia y de la política—; sin embargo, no es suficiente para llegar a inferencias de relaciones causa-efecto, ya que la covariación, aún la más sistemáticamente observada, nunca puede demostrar causación. Por esta razón, la investigación correlacional es considerada como no experimental. Este tipo de investigación proporciona, no obstante, información valiosa para la evaluación de los efectos de los programas educativos. Su diferencia con la experimentación no es tan radicalmente dicotómica como pudiera derivarse de un análisis mecánico y superficial. Ambos tipos de investigación varían en la fuerza relativa de las conclusiones que fundamentan. Los estudios correlacionales pueden desconfirmar algunas relaciones hipotetizadas y, en este sentido, pueden utilizarse para corroborar hipótesis, aunque no para su contraste y confirmación. Ciencias como la Astronomía y la Meteorología constituyen ejemplos claros del valor de los diseños correlacionales —casi exclusivos en estos campos— ya que su progreso parece indudable.

Otros diseños ex-post-facto como las encuestas por muestreo, los estudios comparativos de productos educacionales, o los meramente descriptivos, como la llamada evaluación cualitativa, pueden contribuir a contestar preguntas evaluativas importantes como ¿en qué medida se logran los objetivos previstos?, ¿qué otros resultados educativos se han producido?, etc. Es necesario señalar con Boruch (1979) que los estudios puntuales (single time o single shot), como las encuestas realizadas una sola vez, resultan insuficientes para estimar los efectos de un programa educativo, aunque contribuyen a la evaluación como aproximación explorativa que facilitará el diseño posterior de estudios de campo.

Cooley (1974) recomienda el diseño de estudios longitudicortos como la mejor vía para la evaluación de los efectos de un programa educativo, considerando que tal evaluación exige la medida del estado inicial de los alumnos, las dimensiones de proceso (características del programa) y el estado de los alumnos al finalizar la intervención. Este diseño, ciertamente, es menos controlado que el experimento aleatorizado: se identifica como un cuasi-experimento con mayor grado de control

que los estudios de campo de carácter correlacional. Se pretende controlar hasta donde sea posible los procesos implicados en la ejecución del programa y utilizar este control para reducir la correlación entre las variables independientes (características del ambiente educativo) a niveles sensiblemente inferiores a los normales en los estudios de campo meramente descriptivos.

LA MEDIDA DE LOS EFECTOS DEL PROGRAMA

Sea cual fuere el tipo de diseño elegido (experimento de campo, cuasi-experimento, encuesta, estudio de casos), la evaluación de los efectos de un programa de intervencion educativa exige la medida de los resultados potenciales de tal intervención (variable o variables dependientes en el diseño). El problema fundamental en esta perspectiva es garantizar la adecuación de la medida a las variables que representan el producto educativo. Para satisfacer el criterio de adecuación, el proceso evaluador deberá:

— Especificar las variables que se intentan medir
— Seleccionar y/o elaborar los instrumentos pertinentes
— Juzgar su adecuación para medir tales variables

La especificación de los productos del programa supone, en primer lugar, la identificación de los resultados esperados a partir de las formulaciones de metas y propósitos de la intervención. No es infrecuente obtener información conflictiva procedente de distintas fuentes. Los objetivos del programa pueden diferir según a quien y que se pregunte. Por otra parte, la evaluación exige también la identificación de productos no previstos ni intentados. En todo caso, identificados los resultados en forma de objetivos, un análisis no necesariamente profundo, puede poner de manifiesto que las grandes metas —horizontes a largo plazo— pueden ser en realidad razones justificativas del programa y no conductas terminales, lo que supone la imposibilidad de recoger información sobre ellas y de establecer relaciones precisas entre estas metas y los objetivos más específicos, alcanzables a corto plazo. En ocasiones las metas y objetivos se identifican con generalizaciones (conceptos o constructos) que es necesario operacionalizar para medir.

La selección de los instrumentos de medida es una tarea compleja. En ocasiones no existen realmente pruebas apropiadas: otras veces, existen muchas que parecen responder a las exigencias de la variable representativa del producto. Por otra parte, hay una gran diversidad de fuentes, algunas de las cuales son poco conocidas. En consecuencia, es necesario proceder a la búsqueda sistemática para identificar un conjunto de pruebas potencialmente útiles.

Al valorar la adecuación de los instrumentos de medida, consideraremos las siguientes facetas esenciales:

— La sensibilidad para detectar los efectos de la intervención.
— Relevancia curricular
— Validez instructiva
— Sesgos de medida

Al utilizar un instrumento de medida como base para la evaluación de los efectos de un programa, se trata operacionalmente de ver si, de hecho, el programa, o mejor su aplicación, afecta a la medida. En caso negativo, podemos suponer o que el programa no produce los resultados previstos, o que la medida no los detecta, es insensible a tales efectos. Cuando elegimos un determinado instrumento suponemos que existe, al menos, la posibilidad de que detecte los efectos del programa sobre los alumnos.

La tarea de identificar o construir medidas sensibles a los efectos de un programa educativo puede verse afectada, de una parte, por el grado en que las metas se refieren a cambios en una capacidad general o en una destreza (o serie de destrezas) específica, y de otra, por el tipo de metas elegidas para la evaluación respecto a su horizonte temporal, es decir, metas últimas, alcanzables a largo plazo, o metas próximas o inmediatas, alcanzables a corto plazo. Por ejemplo, una prueba que midiera una capacidad global resultante de diversos tipos de condiciones de aprendizaje —producto mediato de la educación— probablemente no detectaría los efectos de un programa de corta duración centrado en el desarrollo de un conjunto de destrezas específicas implicadas en la capacidad general. Para la evaluación de los efectos del programa, en este caso, lo importante sería la sensibilidad del test para determinar los resultados específicos de la intervención.

Otra faceta importante en la medida del producto educativo es la relevancia curricular de la prueba, es decir, la correspondencia del test (contenidos, destrezas, formato de las cuestiones) con un campo instructivo bien definido. La relevancia curricular supone que los items de la prueba constituyen una muestra representativa del dominio de conductas definido por las metas del programa. En este sentido, las pruebas comerciales y, en general, los instrumentos externos, no construidos ad hoc para la medida de una variable en el contexto de la evaluación de un programa, aunque se refieran a las conductas del dominio definido por la meta, quizá no cubran totalmente el dominio o no se corresponda el énfasis puesto en cada conducta con el asignado en el programa.

La validez instructiva de las pruebas hace referencia al grado en que mide aquello que realmente se ha enseñado. Asumiendo relevancia curricular, la prueba tendrá validez instructiva si el programa garantiza la oportunidad de aprender las metas previstas.

Finalmente, un problema capital en la medida del producto educativo es la posibilidad de utilizar instrumentos sesgados. Un test o un item se dice que es sesgado si mide diferentes rasgos en diferentes grupos. Se trata, pues, de una interacción itemgrupo. El sesgo supone una discriminación sistemática en los resultados de algún grupo en función del sexo, raza, lengua, cultura u otra característica.

Por ejemplo, grupos con bajo nivel de vocabulario, de lectura o de comprensión oral serán discriminados en pruebas que no miden estas variables pero que, de facto, exigen niveles altos en ellas para contestar correctamente a los items. El sesgo, pues, afecta a la validez de contenido y de constructo y a la fiabilidad de la prueba. Existen diferentes procedimientos para detectar el sesgo tanto de las pruebas en su conjunto como para cada item. A este respecto pueden consultarse los trabajos de Jornet (1988) y Berk (1980). En cuanto a las vías para eliminar o reducir los sesgos resultan útiles los siguientes (Berk, 1980):

- Uso de paneles de expertos para juzgar si existen o no sesgos aparentes en la prueba.
- Conducción de estudios estadísticos de análisis de items que permitan identificar sesgos, de acuerdo con la metodología ya contrastada (Jornet, 1988; y Berk, 1980).
- Limitar la interpretación a las medidas que presentan una fuerte evidencia de validez, cuidando mucho las distinciones interpretativas de justificación del contenido y del constructo.

Como síntesis de las consideraciones precedentes sobre la medida de los efectos de programas, presentamos la guía para tomar decisiones sobre instrumentos de medida que ofrece Berk (1980).

GUÍA PARA TOMAR DECISIONES SOBRE INSTRUMENTACIÓN EN EVALUACIÓN

1. **Especificación de objetivos**

   Decisiones: ¿Qué resultados o productos deben ser medidos? ¿Serán detectados los resultados no pretendidos? ¿Cuáles son los resultados apropiados?, ¿a largo plazo?, ¿a corto plazo?

2. **Sensibilidad a los efectos del programa**

   Decisiones: ¿Implican las metas destrezas globales o específicas? ¿Son las medidas sensibles a los potenciales efectos del programa?

3. **Factores que afectan los resultados de las medidas**

   Decisiones: ¿Son las medidas congruentes con el dominio instructivo? ¿Ha existido la oportunidad de aprender? ¿Estimulan los procedimientos la realización máxima?

4. **Sesgos de la medida y sus efectos en la evaluación**

Decisiones: ¿Se limitan las interpretaciones de las puntuaciones a aquellas permitidas por la evidencia de validez? ¿Existen datos empíricos sobre sesgos? ¿Han sido revisadas las medidas para asegurar que están libres de sesgos aparentes?

5. **Pruebas sobre nivel y fuera de nivel**

Decisiones: ¿Qué nivel de prueba está más relacionado con el nivel de programa? ¿Pueden superarse los problemas prácticos, sociales y políticos de las pruebas fuera de nivel?

6. **Tests comerciales versus tests ad hoc**

Decisiones: ¿Satisfacen los tests comerciales las exigencias de los objetivos del programa específico? ¿Es viable desarrollar localmente los instrumentos necesarios?

EL ANÁLISIS DE LOS DATOS EN LA EVALUACIÓN DE LOS EFECTOS DE PROGRAMAS EDUCATIVOS

Como afirma Pedhazur (1975), resulta patente la falta de acuerdo entre investigadores acerca de los métodos analíticos apropiados para el estudio de los efectos educativos. Aplicando diferentes métodos de análisis a los mismos datos frecuentemente se obtienen conclusiones diferentes que, a veces, llegan a ser contradictorias. Los resultados conflictivos en la evaluación de programas anulan la justificación del estudio como base para la toma de decisiones. Por otra parte, Levin (1977) señala que los problemas analíticos, en especial los artificios de la regresión múltiple, son elementos críticos en la producción de resultados que pueden confundirnos.

La calidad del análisis depende, aparte de la competencia e independencia del analista, de la calidad del diseño evaluativo y de su ejecución. Campbell y Boruch (1975) al estudiar los sesgos estadísticos en la estimación de los efectos de un programa, afirman que los métodos de análisis no garantizan resultados aceptables cuando los diseños y datos son inadecuados. Wiley (1976), en esta línea, alerta sobre el uso de diseños ex-post-facto en la determinación de los efectos de los programas educativos.

El proceso de análisis técnicamente no parece presentar graves problemas. Se elige la técnica adecuada en función del diseño, las variables y sus medidas, se aplica y se interpretan los resultados. Sin embargo, este esquema raramente es utilizable en los estudios evaluativos. Los manuales, ciertamente, nos proporcionan el conocimiento de una serie de procedimientos analíticos poderosos —ANOVA,

ANCOVA, regresión multiple y otros— pero que presentan ciertas limitaciones a la hora de su aplicación a la evaluación de programas educativos específicos. La razón fundamental es, quizá, que han sido desarrollados en conexión con los diseños experimentales de investigación que prefiguran estudios altamente controlados, con manipulación de variables y aleatorización completa.

Estos modelos y su aplicación estandarizada son ciertamente útiles, pero resultan insuficientes para tratar problemas evaluativos, siempre inmersos en el mundo real y con grandes dificultades de control. Además de estos modelos, necesitamos una guía para su utilización eficaz cuando tratamos con situaciones «contaminadas». En este sentido, Porter y Chibucos (1974) ofrecieron unas directrices útiles cuando la evaluación de los grandes programas, iniciados en los años 60 en USA, comenzó a poner de manifiesto las discrepancias de los investigadores respecto al uso estándar de los modelos estadísticos citados. Estos autores ofrecen una serie de comparaciones entre diversos procedimientos estadísticos en relación con los diferentes tipos genéricos de diseños. Por ejemplo, cuando los diseños son aleatorizados acuden a las relaciones entre el valor esperado de las medias cuadráticas de error en ANOVA $\rightarrow \sigma_y^2$ y ANCOVA $\sigma_{y \cdot x}^2$. Así $\sigma_{y \cdot x}^2 = \sigma_y^2 (1-r_{xy}^2)$, siendo r_{xy} la correlación intracasillas entre pre y post-test.

Cuanto mayor sea r_{xy} menor será la media cuadrática de error y mayor probabilidad de obtener una razón F significativa con ANCOVA.

Y cuanto menor sea r_{xy} menor será la precisión del uso de una u otra prueba. Si $r_{xy} \cdot 30$, parece aconsejable usar ANOVA.

Obviamente, si $r_{xy} = 0$, la media cuadrática de error es igual en ambos casos.

Cuando los diseños no suponen la aleatorización, como ocurre en la mayor parte de los estudios evaluativos, en que se recurre al bloqueo, porque se han utilizado grupos intactos, o a grupos de control no equivalente, la aplicación de ANOVA en el primer caso, tras eliminar la varianza correspondiente a los bloqueos, y de ANCOVA a las estimaciones de las puntuaciones verdaderas, en el segundo, constituyen soluciones parciales bien conocidas.

En general, los problemas vinculados a la aplicación de métodos analíticos en los estudios evaluativos no experimentales son tratados con detalle en los trabajos ya clásicos de Pedhazur (1975) y Boruch y Wortman (1979). Básicamente, los modelos considerados no se excluyen entre sí. De una u otra forma, todos tienen como sustrato la regresión múltiple, aunque se diferencian, según Pedhazur (1975) en el modo en que usan este modelo de análisis y en el énfasis sobre diferentes estadísticos obtenidos a través del procedimiento. Para su critica. Pedhazur considera como enfoques importantes de la regresión múltiple en el análisis de datos para la evaluación de los efectos de programas educativos la partición de varianza y el análisis de los efectos, con una referencia a los modelos causales.

No parece pertinente repetir aqui los argumentos de Pedhazur (1975) pero creo que no resultará ocioso aludir a ciertas pinceladas críticas de los distintos métodos. En la partición de varianza se reseña el estudio del incremento de la proporción de varianza por efecto de la introducción de cada variable independiente, de la varianza

residual y del análisis de comunalidad. Todas estas variantes se apoyan en las diversas fórmulas de R^2 (el cuadrado de la correlación múltiple entre las variables dependiente e independientes) del estudio a cuyos elementos parece querer dotarse de significado sustantivo. Estos intentos conducen muy frecuentemente a interpretaciones que son altamente cuestionables. Tukey (1954) hace cerca de cuarenta años, discutiendo las complejidades de la partición de la varianza, recuerda que la proporción de varianza explicada depende, entre otros factores, de la variabilidad de cada variable en la población específicamente estudiada. Y concluye, «es probablemente imprudente intentar asignar determinaciones a variables determinantes correlacionadas».

En cuanto al análisis de los efectos, Pedhazur (1975) ha tomado como índices los coeficientes estandarizados y no estandarizados de regresión. Parece que la más útil definición de efecto es el coeficiente no estandarizado de regresión, que constituye al mismo tiempo la más prometedora herramienta para la toma de decisiones respecto al programa. Su aplicación válida, sin embargo, exige el cumplimiento de varias condiciones: medidas con alta validez y fiabilidad y unidades de análisis significativas. Evidentemente la interpretación válida de un coeficiente precisa una ecuación de regresión correctamente especificada. Esto significa que el uso de una plétora de variables,simplemente porque se tiene acceso a su observación o medida, carece de sentido: igual que la teorización post hoc o la utilización superficial y acritica de complejos y sofisticados métodos de análisis.

Los modelos causales pueden ser útiles cuando los grupos se suponen comparables (clases formadas aleatoriamente o con agrupamientos heterogéneos) mientras que los diseños discontinuos de regresión ofrecen ciertas ventajas cuando los grupos no son comparables.

En síntesis, el debate metodológico en torno al estudio de los efectos de programas de intervención educativa, aunque parezca recurrente y perturbador del proceso de desarrollo de la evaluación, ha puesto de manifiesto, entre otras cosas, las deficiencias de los modelos analíticos en este campo y la necesidad en este punto de proceder con prudencia, recurriendo al reanálisis permanente de los datos evaluativos con procedimientos alternativos y a la comparación de resultados, como hicieron Cook y otros (1975) con el programa Sesame Street.

En esta línea y, como recomendación de esta ponencia, podría considerarse, como una medida de seguridad en el análisis de los datos de los estudios sobre los efectos de un programa educativo, seguir el siguiente proceso:

Primer paso. Análisis exploratorio de los datos, sometidos los resultados a los procedimientos rutinarios de descripción:

- Distribución de frecuencias para las variables
- Estadísticos univariados (medias, medianas, modas, varianzas, desviaciones típicas, etc.).
- Nubes de puntos para las variables que se tenga interés en correlacionar.

Segundo paso. Análisis primario, partiendo de los resultados del paso anterior, que permitirán tomar decisiones muy ajustadas respecto al tipo de análisis que se puede y debe aplicar. Es decir, se trata de decidir el modo de analizar formalmente los datos (ANOVA, ANCOVA, regresión múltiple en sus diferentes formas, etc.).
Se procederá al análisis de una variable cada vez, que coincidirá con un objetivo del programa. Esto permitirá determinar cuantos objetivos se han logrado.
En evaluación no parece aconsejable el analisis multivariante sensu stricto, es decir, el considerar más de una variable dependiente simultáneamente.
El análisis deberá mostrar sensibilidad a las relaciones no lineales con la inclusión de términos cuadrados o de mayor exponente (Cohen, 1978).

Tercer paso. Análisis secundario o reanálisis comparando diversos procedimientos.

Cuarto paso. Interpretación que implica la comparación de resultados entre los diversos análisis, determinación del significado y coherencia interna de los resultados (tamaño del efecto) y del significado y coherencia externa.

Como nota final podría decirse que la progresiva toma de conciencia acerca de las defiencias en la definición del producto educativo y de las limitaciones metodológicas en su evaluación, supone, en mi opinión, un cambio de rumbo que puede ser el comienzo de un nuevo camino hacia una evaluación de los efectos de programas de intervención cientificamente válida y socialmente útil.

REFERENCIAS

BORUCH, R. F. y WORTMAN, P. M. (1979): Implications of educational evaluation for evaluation policy. *RRE*, 7, 309-373.

CAHAN, S. y otros (1987): The definition and interpretation of effects in decision oriented evaluation studies. En R.Wolf (Edd.) Educational evaluation.

CAMPBELL, D. T. y BORUCH, R. F. (1975): Making the case for randomized assignment to treatment by considering the alternatives: Six way in which quasi-experimental evaluation in compensatory education tend to under stimate effects. En Bennett y Lumsdaine, «Evaluation and experiment», Academic Press, New York.

CARROL, J. B. (1963): A model of school learning. *Teachers College Record*, 64, 723-733

COHEN, J. (1978): Partialed products are interactions, partialed power are curve components. *Psych. Bull.*, 85, 758-766.

COOK, T. D. y otros (1975): Sesame street revisited. Russell Sage, New York.

COOK, T. y CAMPBELL, D. T. (1979): Quasi-experimentation. Design and analysis for field settings. Houghton Mifflin, Boston.

COOLEY, W. W. (1974): Assessment of educational effects. *Educational Psychologist*, 11 (1), 29-35.

COOLEY, W. W. y LOHNES, P. R. (1976): Evaluation research in education. Irvington, New York.

DE LA ORDEN, A. (1981): ¿Qué pretende ser la tecnología educativa? *Bordón*, 258, 235-243.

DE LA ORDEN, A. (1985): Hacia una conceptualización del producto educativo. *Revista Investigación Educativa*, Vol. 3, n.º 6, 271-283.

DE LA ORDEN, A. (En prensa): El éxito escolar. *Revista Complutense de Educación*.

GREENWALD, A. G. (1976): Within-subject designs: to use or not to use? *Psychological Bull.*, 83, 314-320.

HERSEN, M. y BARLOW, D. H. (1976). Single case experimental designs. Pergamon Press, New York.

LEVIN, H. M. (1977): A decade of policy developments in improving education and training for low income populations. En R.H. Haveman (Ed.), «A decade of federal antipoverty programs: Achievement, failures and lessons». Academic Press, New York.

MAHONEY, M. J. (1978): Experimental Methods and outcome evaluation. *Journal of Consulting and Clinical Psychology*, 46 (4), 66-672.

PEDHAZUR, E. J. (1975): Analitic methods in studies of educational effects. *RRE*, 3, 243-286.

PORTER, A. C. y CHIBUCOS, T. R. (1974). Selecting analytic strategies. En Borich (Ed.), «Evaluating educational programs». Educational Technology Publications, Englewood Cliffs, New Jersey.

TUKEY, J. W. (1954). Causation, regression and path analysis. En O. Kempthorne y otros (Eds.), «Statistics and mathematics in biology». Ames: Iowa State College Press.

WILEY, D. E. (1976): Another hour, another day: Quantity of schooling a potent path for policy. En W. H. Sewall y otros, «Schooling and achievement in American Society». Academic Press, New York.