

# *Evaluación de Programas de Reinserción Social*

Violeta Cardenal, Javier Corbalán Berná,  
M. Rosa Esteve Zarazaga, Vicente Gradillas Regodón,  
Blanca Moreno-Mitjana, Pilar Moreno,  
María Navarro Pacheco,  
Margarita Ortiz-Tallo Alarcón,  
Luis Valero Aguayo



MIGUEL GÓMEZ EDICIONES

1ª Edición: mayo 1994

© 1994 de los autores de sus respectivos capítulos.

© de la presente edición Miguel Gómez Ediciones.

Pº Calvo Sotelo, 28. 29016 Málaga.

ISBN: 84-88326-08-4

Depósito Legal: MA-0000/1994

Impreso en España.

Imprime: Gráficas San Pancraccio, S. L.

Pol. I. San Luis, C. La Orotava, 17. 29006 Málaga.

Diseño, maqueta y composición: gp Fotocomposición.

Arenal, 11, 2º C. 29016 Málaga. Tel. (95) 260 28 73.

1.4. Técnicas de recogida de datos.	127
1.4.1. Observación cualitativa	127
1.4.2. Observación participante	127
1.4.3. Entrevistas	127
2. Características generales del centro	128
2.1. Financiación	128
2.2. Personal laboral	128
2.3. Filosofía del programa	128
3. Etapas del programa Proyecto Hombre	130
3.1. Etapa de acogida	130
3.1.1. Características y aspectos del funcionamiento.	130
3.1.2. Categorías centrales	135
3.1.3. Comentarios de los evaluadores	140
3.2. Etapa de comunidad	141
3.2.1. Características de la comunidad y comentarios del evaluador	142
3.3. Fase de reinserción	146
3.3.1. Características	146
3.3.2. Categorías centrales	147
4. Conclusiones finales	154

## Apéndice

### El manual integrado 'Program evaluation kit'

*Luis Valero Aguayo*

1. Orientación a la medición de programas	160
2. Cómo diseñar una evaluación de programas	164
2.1. Elementos de un diseño	166
2.2. Tipos de diseños	169
3. Definición de metas y objetivos	172
4. Medición de programas	175
5. Técnicas de evaluación	178
5.1. Registros	178
5.2. Observación	179
5.3. Autoinformes	182
6. Fiabilidad y validez de las medidas	183
7. Informes de aplicación	184
Referencias	185

## Apéndice

### EL MANUAL INTEGRADO "PROGRAM EVALUATION KIT"

Luis Valero Aguayo

**N**uestro objetivo en este capítulo es presentar un manual completo para la valoración de programas, que es poco conocido en nuestro país y cuya divulgación podría servir para asentar algunas bases metodológicas en este nuevo campo. Forma un conjunto de ocho libros editados bajo el nombre general de la serie como "*Program Evaluation Kit*" (*Manual de Valoración de Programas*), y que abarca todas las etapas prácticas en la planificación, desarrollo y evaluación de programas sociales, aunque la mayor parte de los ejemplos y estudios que presentan se circunscriben al ámbito educativo. Sin embargo, dado que son un conjunto de manuales prácticos, sobre cómo hacer la valoración, podrían ser aplicados sin problemas al estudio de cualquier tipo de programa de tipo comunitario, que abarque gran cantidad de personas u objetivos amplios, en los que deban producirse cambios de comportamientos, de actitudes, de satisfacción de usuarios, etc. o bien sirva como remodelación del programa en curso.

Estos manuales fueron desarrollados por un grupo de investigadores, al frente de L.L. Morris y C.T. Fitz-Gibbon (1978) en el *Center for the Study of Evaluation* de la Universidad de California en Los Angeles, en un programa de desarrollo iniciado ya en 1966, que ha publicado la mayor parte de esos manuales en 1978, y de los cuales se han llegado a realizar una veintena de reediciones. Esos manuales incluyen un libro general del evaluador (*Evaluator's Handbook*) con recomendaciones, orientaciones y resúmenes de los aspectos fundamentales de los programas de evaluación cuantitativa. Además incluyen otros manuales sobre cómo diseñar un programa, cómo definir metas y objetivos, o cómo llegar a cabo un programa. Añaden también otros manuales prácticos sobre los métodos de medición de esos objetivos, la medición de actitudes, los cálculos estadísticos y gráficas realizables con los datos generados, e incluso cómo redactar un informe de la valoración en función de la audiencia y los objetivos de la investigación.

Presentamos, pues, a continuación un extracto de esos libros con los aspectos metodológicos que consideramos fundamentales. Algunos puntos están resumidos y otros comentados o parafraseados, por lo que remitimos a los manuales originales en inglés a aquel lector que desee utilizar esta metodología cuantitativa en sus estudios de valoración de programas.

## 1. Orientación a la medición de programas

En general, al describir la aplicación de un programa se considera como sinónimo de medición atenta de procesos objetivos o determinación del cumplimiento de medios y objetivos-metas, aunque engloba todos esos términos. Se asume por quien lee un informe de evaluación que los resultados descritos por el evaluador han sido obtenidos por un grupo completo de materiales y actividades que están directamente relacionadas con la conducta de los sujetos, profesores, terapeutas, agente sociales, padres, administradores o directivos, que han intervenido en el programa. Habría que responder a dos preguntas fundamentales: 1. ¿Han cumplido con su cometido, con los objetivos propuestos, la amalgama de materiales, medios, actividades, personas, etc., comprendidas en la descripción del programa?. 2. ¿Qué específicamente ha funcionado en el programa?.

La audiencia es un punto importante en la evaluación. Aunque se escriba un informe que nadie leerá, cada evaluador lo realiza para al menos una audiencia. Muchas evaluaciones tienen varios tipos de audiencia. Una audiencia es una persona o un grupo de personas que necesitan la información de la evaluación para un propósito distinto, y por tanto, cada audiencia necesita diferentes informaciones, cada uno mantiene diferentes criterios por los que la evaluación tendrá informaciones creíbles o fiables. De esta forma, un informe detallado de una evaluación realizada por personas no familiarizadas con ello, debería incluir y prestar atención especial a los siguientes puntos:

1. Descripción de las características del programa:
  - Información fundamental y contextual, características del programa.
  - Hechos críticos para la realización del programa.
2. Datos para apoyar la descripción realizada:
  - Medidas a llevar a cabo.
  - Discusión de la ejecución del programa.

Una evaluación debería dar respuesta a aquellos otros centros o profesionales que, antes de la aplicación de un programa concreto, responden: "No gracias, hazlo primero y dime qué paso". Una evaluación no concluyente o poco clara,

no permite responder de forma efectiva si merece la pena volver a aplicar el programa, en función de su eficacia probada.

Antes de planificar la forma de recoger los datos sobre el programa a evaluar, deben tomarse dos decisiones:

1. ¿Qué hechos del programa son los más críticos o válidos para su evaluación?
2. ¿Cómo y qué tipo de recopilación de datos son necesarios para describir con precisión cada uno de los componentes del programa?

Responder a los propósitos con los que se lleva a cabo la evaluación de un programa, supone adoptar un papel de evaluador como documentalista (una evaluación sumativa), o bien un papel de evaluador monitor.

La documentación de un programa la constituye su descripción oficial, subrayando los hechos críticos del programa en su aplicación, así como las variaciones efectuadas. La documentación naturalmente ha de ser bien definida y sólida, y por tanto debe realizarse cuando ha transcurrido un tiempo suficiente desde su comienzo, y han podido aparecer problemas o cuestiones en su aplicación. La evaluación como monitorización supone un sistema más activo y menos fijo del evaluador. Ha de tener un papel de vigilante, consejero, solventador de problemas, durante la aplicación del proyecto y no sólo describirlo. Tiene un papel más formativo puesto que se pueden aislar partes del proyecto más o menos útiles, o reemplazar partes completas de ese programa, de forma que permita modificaciones en su aplicación continua.

Parte de *la tarea del evaluador sumativo o documentalista* es recoger, por medios externos, una descripción oficial de lo que constituye el programa, con el propósito de:

1. Resumir y reunir datos, muchas veces dispersos, no recopilados o inacabados, que no permiten concluir sobre un programa, y cuantificar los datos obtenidos o los objetivos alcanzados en cada caso.
2. Proporcionar una última descripción del programa, una vez aplicado en la situación real, con suficientes detalles que permitan su nueva aplicación o rectificaciones posteriores.
3. Proporcionar una lista de posibles causas de los efectos del programa, con un diseño fiable y medidas válidas de los resultados, con lo que se constituye prácticamente en un estudio experimental. Proporcionando a los administradores pistas sobre la efectividad de diversos programas que utilizan medios, materiales o procedimientos semejantes. La información recogida se centraliza sobre aquellas actividades, cambios administrativos, financiación, materiales, etc., que constituyen los cambios específicos diseñados o propuestos por los creadores del programa como las variables efectivas de acción en la aplicación de ese programa. La cantidad de detalles sobre estas características dependerá del grado de precisión en la descripción proporcionada por esos planificadores, y como mínimo se han de describir con detalle aquellos hechos que el grupo de planificadores del programa considera como fundamentales

o cruciales en los posibles efectos de ese programa. Y se han de incluir suficientes datos sobre esos cambios y esas características como para convencer a los más escépticos. En algunos casos, los datos pueden estar redactados en plan informal en función de la audiencia, pero en investigación y donde es posible la controversia sobre las conclusiones, los datos suministrados han de ser sistemáticos y la descripción el programa es fundamental.

*El evaluador con un papel de monitor o formativo* supone mayor cantidad de responsabilidad en el program ,asumiendo no sólo actividades de reporte y recogida de información, sino también toma parte en la planificación del programa, en su desarrollo y en su aplicación. Su misión abarca:

1. Asegurar, siguiendo el desarrollo de un programa, que la descripción oficial de ese programa es seguido tal cual, reflejando exactamente cómo se aplica el programa. A veces, su primera labor es la propia clarificación del plan del programa.
2. Ayudar al grupo de planificadores y profesionales implicados a añadir, corregir o mejorar el programa en cuestión y su desarrollo. Ha de ayudar al grupo a reflejar periódicamente el curso del programa y envolverse en el proceso tal cual va ocurriendo. Ha de diseñar en algunos casos estudios-piloto tentativos, sobre partes específicas del programa, y tomar decisiones constantes sobre la marcha del mismo.

Respecto a la descripción de las características del programa, se puede comenzar subrayando los elementos del contexto del programa: por ejemplo, la clase, la escuela, el distrito escolar, el equipo de profesionales, maestros, padres, voluntarios, personal administrativo, los recursos utilizados incluyendo el material construido o comprado, el equipo didáctico, el equipo de utilización exclusiva del proyecto, los sujetos o alumnos que reciben el programa, sus características, número, y sus niveles de habilidades o competencias al comienzo del programa. Este tipo de datos constituye el esqueleto del proyecto, y deben ser incluidos en cualquier informe realizado. Informa de las actividades más importantes realizadas, y reflejarlas cuando ocurren, es importante porque resultan más difíciles de reconocer y valorar una vez terminado el programa, ya que otros tipos de materiales permanentes (exámenes, material didáctico, aulas, etc.) son más fáciles de recopilar, mientras que las modificaciones introducidas durante el programa difícilmente estarán reflejadas en ningún lado. Se han de elegir exactamente qué actividades reportar, y elegir las en función de los objetivos del programa y la audiencia a la que interesa la evaluación de ese programa. Además otras fuentes de información ayudarán a decidir qué examinar durante el programa, entre ellas podemos destacar:

1. **El plan o propósito del programa.** Algunos de los programas suelen especificar cuáles son los hechos o actividades más relevantes y por qué, otros

no lo hacen en absoluto. En general, aquellas actividades más repetidas suelen ser las más importantes en el propósito del programa.

El costo suele ser otro índice de los hechos cruciales de éste, generalmente, los mayores gastos se suponen dedicados a las actividades más importantes del programa. Una evaluación sobre la aplicación de un programa puede basarse en la medición de hasta qué punto los hechos o actividades cruciales de la planificación han sido llevadas a cabo con exactitud o tal como fueron planificadas. Esta recopilación de datos puede ser realizada también mediante entrevistas a los planificadores y profesionales implicados. A veces, es posible encontrar que el programa no ha sido planificado en absoluto, o lo ha sido de una forma muy general. En estos casos, bien se toman datos del personal, tal como cuenta que es o debería ser el programa, o se actúa como un observador y se toman datos de la situación real tal cual se lleva a cabo.

**2. La recopilación de opiniones y consejos de los expertos,** personal del programa, consultores y el evaluador mismo. A veces, muchas actividades cruciales no se ven reflejadas en la planificación del programa, pero pueden ser observadas o analizadas por el evaluador, otros profesionales u otras personas que observan la aplicación del programa. Y de cara a una clarificación metodológica la variable importante en un programa es la que efectivamente se aplica en la situación, esté o no planificada con anterioridad.

Cooley y Lohnes (1976) elaboraron unas pautas de evaluación de programas educativos, intentando eliminar la cuestión de la teoría implícita en el plan educativo propuesto, de forma que cualquier programa pudiese evaluarse con unas pautas y datos comunes, independientemente de la teoría, filosofía o bases iniciales del plan. Basan su modelo en la evaluación de cuatro cuestiones fundamentales:

- a. La oportunidad: El tiempo y lugar de aplicación, la práctica, recursos, formas de nueva información, contacto del alumno con el plan educativo, etc.
- b. Motivación: Actividades, técnicas y recursos motivacionales utilizados en el plan educativo.
- c. Estructura: Forma de presentar las actividades, la información, las técnicas, etc., su delimitación y ordenación respecto a los objetivos propuestos en cada caso.
- d. Eventos instruccionales: Eventos específicos descritos tal cual ocurren al llevar a cabo el plan educativo, número de alumnos, atención del profesor, material empleado, duración de las clases, ayudas suministradas, etc.

**3. Las propias observaciones del evaluador.** A veces es posible para un evaluador observar la ejecución de un programa con relativamente pocas preconcepciones o decisiones acerca de qué observar, y suele ser la mejor

forma de describir un programa. No siempre es posible evitar las preconcepciones y prejuicios sobre lo que se está observando. La evaluación "naturalista/responsiva", propuesta por Stake (1975), utiliza los métodos naturalistas para examinar y evaluar un programa, evitando al tiempo preconcepciones sobre lo que el evaluador debe finalmente describir. Esta aproximación es aún más necesaria cuando no existe un proyecto previamente definido o consistente de las características del programa. Y muchas veces se encuentra que las variaciones de un mismo programa son enormes de un lugar de aplicación a otro, y que los hechos o actividades importantes no lo parecen en un principio. En estos casos, existen dos opciones:

- a. Recoger la mayor cantidad de datos posible, generalmente mediante la observación directa, y esperar a recopilarlos y organizarlos al final.
- b. Describir exactamente lo que se observa, sin comparar en absoluto con lo planificado o con lo que debería haber ocurrido.

Este método naturalista sigue las siguientes fases en la evaluación de un programa:

- a. Elige la situación concreta de evaluación donde se lleva a cabo el programa. Si existen varios escoge uno de ellos para sus observaciones.
- b. El evaluador observa las situaciones naturales en la situación escogida, tratando de no influenciar la rutina del propio programa.
- c. Recoge datos que pueden tener forma de registros codificados u otras técnicas de registros observacionales, a veces simples notas de campo.
- d. El evaluador contrasta los datos recogidos con los datos formales u oficiales del programa, y conversaciones o entrevistas con el personal en cuestión. Produciendo generalmente un informe descriptivo, progresivamente más completo, a veces con gráficas, tablas, u otros datos cuantificados si se han recogido.

**4. Las variaciones durante la aplicación del programa.** A veces el elegir unas características concretas a describir depende de la cantidad de variaciones introducidas, y no planificadas, que ocurren en el programa de situación a situación, espacial y temporalmente. Desafortunadamente, la evaluación final de un programa una vez han ocurrido multitud de variaciones no permite dilucidar las razones del funcionamiento o no de dicho programa.

## 2. Cómo diseñar una evaluación de programas

Un diseño es un plan que dicta cuándo, cómo y de qué se realizarán las mediciones en el curso de una evaluación. También constituye una forma de obtener información comparativa de los resultados de un programa, y puede ayudar al evaluador a fijar qué casos deberían haber ocurrido durante el programa y no lo hicieron, y qué ocurrió en su lugar. Generalmente, un diseño emplea la comparación de las mismas medidas o instrumentos administrados

en otro grupo que no recibe el programa; y sus resultados son comparados con los que se producen en el programa a evaluar. Otras veces, también resulta posible averiguar por medios estadísticos qué hubiera pasado si no se hubiese aplicado el programa, y comparar así los resultados con o sin programa sin necesidad de un grupo control.

En la mayoría de las situaciones de evaluación, siempre es mejor alguna información comparativa que ninguna. La elección de un diseño quizás determine si la información producida es creíble y será utilizada por la audiencia que lo reciba, o bien si será eliminada porque puedan crearse otras muchas interpretaciones de los mismos resultados por no haber tenido una especial atención al control de otras variables en el diseño.

Los diseños aquí presentados no son los únicos disponibles en cualquier investigación práctica, sino que se muestran los más usuales y que en su mayoría son más potentes y fáciles de entender por la audiencia a la que se presentarán los resultados. La información obtenida en un estudio bien diseñado es difícilmente refutable, y sus resultados no son fácilmente ignorables.

El trabajo de un evaluador de programas puede seguir dos vías fundamentalmente, según su implicación con el programa: una evaluación sumativa (pica); también puede relajar los requisitos de aplicación de un diseño, realizándolo con menos sujetos, con asignación semi-aleatoria, etc., pero siempre teniendo en cuenta tales modificaciones a la hora de realizar las conclusiones finales. El objetivo es proporcionar la mayor cantidad de oportunidades posibles para contrastar y controlar los resultados que se vayan obteniendo. Y otra alternativa, consiste en planificar experimentos cortos, estudios piloto, en los que se verifique sólo un componente del programa.

En aquellos casos que por las características del programa, o de los sujetos a los que se aplica, no sea fácil utilizar un diseño con grupo control incluido, se pueden adoptar otras estrategias, por ejemplo, utilizar un grupo control no-equivalente (quizás con otro programa en marcha, o sin ninguno en especial, en otra población, o de otro nivel socio-económico, etc.); también comparar los componentes aislados del programa (materiales, profesores, horarios, organización, etc.), o compararlos en términos de satisfacción, o con referencia a personas externas al programa (padres, grupos sociales, trabajadores, etc.); tener en cuenta los criterios sociales de aplicación de ese programa y contrastarlos con las opiniones de otras personas fuera de la aplicación.

Quizás en un evaluador formativo la función de los diseños sea más importante aún, aunque necesite cambios continuos con el desarrollo del programa. De esta forma, una de sus funciones sería proporcionar información al grupo de planificación sobre la marcha del programa, sus cambios o mejoras, e introducir siempre pequeños estudios piloto que ayuden a delimitar las características fundamentales del éxito del programa completo. Ha de tener información sobre cursos de acción alternativos y sobre la efectividad mayor o menor de cada uno de los componentes del programa. En estos casos,

se puede establecer como "control" varias alternativas del mismo programa aplicadas en diferentes situaciones o sujetos con variaciones entre ellos (replicación sistemática); también puede relajarse los requisitos de aplicación de un diseño, realizándolo con menos sujetos, con asignación semi-aleatoria, etc., pero siempre teniendo en cuenta tales modificaciones a la hora de realizar las conclusiones finales. El objetivo es proporcionar la mayor cantidad de oportunidades posibles para contrastar y controlar los resultados que se vayan obteniendo. Y otra alternativa, consiste en planificar experimentos cortos, estudios piloto, en los que se verifique sólo un componente del programa.

En aquellos casos que por las características del programa, o de los sujetos a los que se aplica, no sea fácil utilizar un diseño con grupo control incluido, se pueden adoptar otras estrategias, por ejemplo, utilizar un grupo control no equivalente (quizás con otro programa en marcha, o sin ninguno en especial, en otra población, o de otro nivel socio-económico, etc.); también comparar los componentes aislados del programa (materiales, profesores, horarios, organización, etc.), o compararlos en términos de satisfacción, o con referencia a personas externas al programa (padres, grupos sociales, trabajadores, etc.); tener en cuenta los criterios sociales de aplicación de ese programa y contrastarlos con las opiniones de otras personas fuera de la aplicación concreta del programa; y tener en cuenta los objetivos, planteamiento y teorías que fundamentan el programa que se está evaluando.

### 2.1. Elementos de un diseño

La palabra "grupo" se identifica con "grupo de tratamiento" cuando se habla de diseños, y está reservada para aquel conjunto de personas que reciben un programa o tratamiento. Dividiendo siempre entre un "grupo experimental" que recibe el programa a evaluar y un "grupo control o comparación" que no lo recibe y tiene otras características en su programa.

Un grupo control siempre es similar dentro de lo posible al grupo experimental, y es medido de la misma forma y al mismo tiempo. El grupo de control puede ser de muy distintas formas, y no necesariamente significa que no reciben el programa de tratamiento. Se suelen dividir en dos tipos: aquellos que se hacen equivalentes al asignar aleatoriamente a los sujetos, lo que asegura una distribución no sesgada respecto a sus características; y aquellos que pueden considerarse como no equivalentes, al no realizarse una asignación completamente aleatoria.

La aleatorización constituye un medio de hacer un grupo realmente como grupo control o equivalente, ya que los resultados obtenidos no pueden atribuirse a otra cosa que a la diferencia en los tratamientos aplicados a cada grupo. La asignación aleatoria de los sujetos a un programa constituye la forma más efectiva de eliminar explicaciones alternativas. Además, la aplicación de pruebas estadísticas a los resultados requieren muchas veces que los sujetos estén distribuidos por igual en ambas muestras. Otras estrategias para no incluir un grupo de control puro, sin tratamiento, que supondría un gran

gasto social y económico en la evaluación de programa, consistiría en:

1. Introducir dos programas nuevos a la vez, de forma que no actúe como control del otro.
2. Crear un grupo control por debajo de las necesidades inmediatas del programa, de forma que a los sujetos que se suponga van a necesitar ese programa se incluyan todos en el grupo experimental.
3. Asignar la aplicación del programa unas veces a un grupo y otras veces al otro grupo, creando turnos de aplicación.
4. La aplicación demorada del programa, de forma que el grupo control actúe como tal en una primera parte y después pase a aplicarse el programa.
5. La evaluación demorada, que implica la evaluación del grupo control y experimental cuando ya se han llevado a cabo, efectuando la comparación a posteriori.

El grupo control no equivalente es un grupo seleccionado por su similaridad con el grupo experimental, pero no se forma aleatoriamente. Un grupo es no equivalente si se seleccionan los sujetos por un procedimiento especial, si los dos grupos son medidos de forma diferente, con otros instrumentos o en intervalos de tiempo diferentes; y debería demostrarse que los dos grupos son lo más parecidos posible entre sí excepto por el programa que reciben.

La mejor decisión respecto a qué grupo control adoptar consiste en tener como programa control otro programa en competencia directa con el que se pretende probar, para observar así su potencia frente al programa opuesto. Otra decisión es utilizar un programa diferente pero con los mismos objetivos que el experimental, o frente al programa antiguo si es una renovación lo que se pretende evaluar. Otra solución, aunque más débil, consiste en evaluar la no presentación de ningún programa, es decir, evaluar sus efectos frente a "no-hacer-nada". En todo caso, se podría llegar a la conclusión de que al menos el programa no es perjudicial para los sujetos, y siempre será mejor que ningún programa.

Las pruebas aplicadas a un programa pueden ser antes o después de su desarrollo, así son denominadas pretests- posttests, las mediciones realizadas. Los post-tests son las mediciones realizadas al final del programa, constituyen la VD en la que observar los resultados obtenidos con ese programa. A veces, el hecho de cuándo realizar el post-test viene determinados por la necesidad de presentar informes, o simplemente porque el curso escolar o el periodo de vacaciones da por finalizada la experiencia. Pero, generalmente, dependerá del equipo de evaluación la decisión sobre cuándo efectuar las mediciones, y siempre se han de realizar con todos los sujetos, eliminando del análisis aquéllos que por cualquier circunstancia no lo hicieran a su debido tiempo.

Las pruebas pre-tests se aplican antes de comenzar el programa. Se suelen utilizar como medidas para seleccionar sujetos, comprobar datos supuestos

antes del inicio del programa, y fundamentalmente para tener una comparación de la ganancia obtenida con el programa.

La selección de sujetos con medias pre-tests puede tener un peligro, y es la regresión a la media de esas medias. Si se selecciona un grupo de sujetos por su puntuación extrema (alta o baja), en la siguiente medición las puntuaciones tienden a su media, con lo que se podrían atribuir resultados a un efecto que no es sino debido a la medición repetida con la misma prueba. Por esta razón, es muy importante tener un grupo control cuando se está evaluando programas de tratamiento que suponen mejora o aceleración de los repertorios aprendidos en los sujetos. Una recomendación en estos casos es utilizar una segunda evaluación de pre-test, en la que probablemente las puntuaciones estén ya distribuidas siguiendo una probabilidad normal y permita la comparación con los datos finales.

En los casos de utilización de un verdadero grupo control, la necesidad de un pretest se reduce al haber asignado aleatoriamente a los sujetos, aunque siempre podría servir para comprobar que realmente los dos grupos son inicialmente similares en sus puntuaciones. Cuando se trata de comprobar con los pretests la ganancia por comparación con los posttests, ello es posible sólo cuando se utilizan "tests referidos a un criterio", en los que se compara las puntuaciones de un sujeto con un criterio predeterminado que se ha elaborado para medir el repertorio de que se trate, generalmente consisten en descripciones de objetivos de habilidades y repertorios. Por el contrario, un "test referido a una norma" compara las puntuaciones de un sujeto con las de otros sujetos para determinar el rango en que ese sujeto se encuentra respecto a la media general en esa materia o área predeterminada. La práctica general de comparar la ganancia en puntuación de los sujetos utilizando tests estandarizados dice muy poco sobre lo que realmente los sujetos han aprendido. En el caso de un verdadero grupo control las puntuaciones iniciales se suponen similares en ambos, por lo que no son tan necesarias, y las comparaciones se realizarían respecto a los obtenidos en los posttests. Si no se utiliza, entonces serán necesarias las medias pretests, y en ellas utilizar pruebas referidas a un criterio si se quiere tener datos de la ganancia real obtenida en los sujetos.

Un diseño tiene la potencia como para detectar pequeñas diferencias que permitan al evaluar explicar las causas de las variaciones en los resultados. La forma de incrementar la potencia de un diseño es incrementar su poder explicativo para detectar con las medidas apropiadas todas las variables que posiblemente afectarían a los resultados. Una forma de incrementar esa potencia es utilizar un pretest muy similar al post-test, puesto que el mejor predictor de la conducta futura de una persona es su medición actual en similares circunstancias. Así, al utilizar un pretest que es muy similar o idéntico al post-test se puede obtener información más precisa sobre la efectividad de un programa comparado con otro.

Hay que tomar, sin embargo, algunas precauciones y no se han de utilizar

medidas pretests cuando esto pudiese afectar de algún modo la conducta del sujeto, por ejemplo, determinadas preguntas en cuestionarios podrían forzar las respuestas sociales del sujeto y obtener datos sesgados por el propio instrumento utilizado; una solución es medir esas cuestiones sólo en algunos de los sujetos seleccionados al azar dentro de cada grupo. No se ha de emplear cuando no va a clarificar nada, o no va a aportar más información de la que ya se tiene sobre los sujetos. Desde luego, no se debe emplear cuando el programa ha comenzado ya, puesto que entonces no sería un pretest adecuado al estar sesgado por los posibles resultados transitorios del programa, y no podrían suponerse como el repertorio inicial de los sujetos en esas pruebas. Finalmente, no se deben utilizar cuando suponen un alto costo en tiempo y dinero antes de comenzar el programa en cuestión.

Un pretest puede ser:

1. Una prueba de actitudes, sobre todo cuando se pretende medir los cambios de actitud u opinión de las personas, y en estos casos por la reactividad a la prueba se suele sólo evaluar la mitad de los sujetos de cada grupo.
2. Una prueba de ejecución, al medir los resultados de un repertorio adquirido, entonces sí se hace necesario una comparación de esa ejecución antes del comienzo del programa.
3. Una prueba de habilidades, cuando se pretende medir también ejecución, y entonces se emplea como una medida de la significatividad educativa, o un juicio sobre los avances conseguidos en cada sujeto, por lo que también se emplean antes y después del programa.

También se pueden tomar diversas mediciones entre la aplicación del tratamiento, en este caso entre-tests, cuando se toman algunas medidas durante el tiempo en que el programa está en ejecución, y pueden ayudar a observar su transcurso a lo largo del tiempo. Los llamados "tests de retención o de seguimiento" evalúan los resultados del programa pasado un tiempo, en un intento de observar los efectos de mantenimiento que puede conseguir ese programa.

Por otra parte, las medias tomadas en pruebas repetidas a lo largo del tiempo ("series temporales") da lugar a otro tipo de diseños en los que la repetición de las pruebas de forma sistemática antes de comenzar el programa, durante y después de terminado, permiten obviar la necesidad de un grupo control. Puede dar razón de qué hubiera sucedido si no se hubiese aplicado el programa, y qué resultados se han obtenido con él.

Resumiendo, entonces, la elección de un diseño vendrá dado por la elección de qué sujetos medir: un grupo experimental, o dos grupos (con control equivalente o no), y por el tiempo en que se realizarán las mediciones: pre y posttests, sólo posttests, y series temporales.

## 2.2. Tipos de diseños

Los diseños constituyen las estrategias experimentales para asegurar que los efectos o cambios detectados son debidos única y exclusivamente al programa llevado a cabo. Dado el tipo de temas de estudio y las características de la población a la que van destinados estos programas, la mayor parte de los diseños utilizan la comparación entre grupos. Estos serían algunos de los diseños fundamentales disponibles:

### *Diseño 1. Pre-post con grupo control verdadero*

Este es un diseño clásico en el que los sujetos del programa son asignados aleatoriamente a dos grupos diferentes, uno que recibirá el Programa X y otro que no lo recibirá, o tendrá un programa estándar (alternativo) durante ese tiempo. Las puntuaciones del pre-test se utilizan para probar que los dos grupos parten de datos más o menos equivalentes. Si al final del programa, las puntuaciones del grupo experimental son significativamente más elevadas que las del grupo control, esta diferencia puede ser atribuida al Programa X frente al programa estándar.

### *Diseño 2. Post con grupo control verdadero*

Es similar al anterior, excepto que no hay medidas pre-test. Este diseño resulta más útil cuando las medidas iniciales podrían interferir con los efectos del programa, o llevaría tanto tiempo que las medidas pre-test no serían muy fiables. La asignación aleatoria de los sujetos a los grupos Experimental y Control asegura la equivalencia de ambos grupos. Con este diseño es posible esperar hasta terminar el programa y decidir qué tipos de medidas se utilizarán como post-test.

### *Diseño 3. Pre-post con grupo control no equivalente.*

Es similar al primer diseño sólo que los grupos no han sido asignados al azar, los grupos no son equivalentes. Resulta un diseño muy útil para la evaluación en contextos escolares, donde permite probar la similaridad o diferencia entre dos clases o grupos dentro de una clase, al menos con las medidas utilizadas. Se puede utilizar cuando, por diversas circunstancias, no resulta posible aleatorizar completamente los sujetos dentro de los grupos, así una clase que no entre dentro de un programa, o un grupo de pacientes que no reciban una terapia, se convierten en grupos control no equivalentes.

### *Diseño 4. Pre-post intragrupo.*

Es un programa con menor validez interna, pero permite también comparar los resultados de un programa antes y después de su aplicación. En este caso no hay grupo control para comparar, ni aleatorización de los sujetos. Las comparaciones son exclusivamente intra-grupo con las mediciones anteriores y posterior al programa a evaluar. Resultan más útiles cuando se quiere comparar con algunas puntuaciones normativas, de referencia anteriores al programa, o con algunos criterios de ejecución respecto al que se comparan los sujetos.

Para llevarlo a cabo sólo es necesario tomar una medida inicial de todos los sujetos del grupo que se va a utilizar en el estudio, documentar la ejecución del programa en curso, y una vez terminado realizar una nueva medición de los mismos sujetos.

Dado que presenta bastantes problemas, es necesario ser cauto al utilizarlo como diseño experimental, y aumentar su validez interna con algunas de las siguientes estrategias: (1) Utilizar tests referenciales en aquellos casos en que los objetivos del programa puedan ser medidos con tests, aunque con cautela al interpretar los datos comparativos con la población normativa; (2) Observar si el programa se aplica en diferentes situaciones, clases, departamentos, etc. para tomar esos ligeros cambios como grupos de control no equivalentes; (3) Estudiar el impacto diferencial del programa sobre diferentes sujetos con características diferenciales (sexo, empleo, nivel estudios, nivel económico, tiempo permanencia en el centro, etc.), y utilizar esas diferencias como una variable independiente; (4) Desarrollar y aplicar diversos instrumentos y parámetros para tener una medición intensiva del fenómeno en estudio, y así poder detectar aquellas mediciones que puedan ser más sensibles al programa evaluado; (5) Si el programa consiste en la consecución de objetivos, o hay criterios de ejecución, utilizar esos criterios como norma de referencia para las comparaciones, o realizar pruebas específicas con referencia a esos objetivos.

#### *Diseño 5. Series temporales intragrupo*

Este diseño utiliza los propios sujetos de un único grupo como control y experimental, realizando mediciones repetidas en el mismo grupo antes y después de la aplicación de un programa. Ello supone que se han de tener suficientes datos en cada fase de series temporales como para permitir la evaluación de los resultados posteriores, bien con la ruptura de la tendencia o el nivel de los datos tras la aplicación del programa.

Los pasos esenciales para llevar a cabo este diseño, suponen: (1) Preparar o seleccionar un sistema de medida que pueda usarse repetidas veces; (2) Decidir la composición del grupo experimental, que bien puede ser el mismo grupo de personas medido varias veces, un conjunto de personas escogidas aleatoriamente dentro del grupo en cada medición, o grupos sucesivos con idénticas características medidos una vez cada uno; (3) Recoger diversas medidas durante intervalos regulares antes de la aplicación del programa (al menos tres puntos de medición); (4) Comprobar la aplicación del programa; y (5) Volver a tomar diversas medidas a intervalos regulares después que el programa haya acabado. Todas las mediciones han de hacerse en las mismas condiciones, con los mismos instrumentos y el mismo sistema de registro; de otra forma sería difícil poder comparar puntuaciones seriadas.

#### *Diseño 6. Series temporales con grupo control no equivalente*

Es un diseño similar al anterior, sólo que se añade un grupo control para aumentar la validez interna del experimento. Esto supone que los dos grupos son medidos de forma regular durante un tiempo, antes y después de la

aplicación del programa. Puesto que supone una mezcla del diseño de grupos y series temporales añade mayor cantidad de comparaciones y fuerza a las conclusiones del estudio.

Los pasos necesarios para llevar a cabo este programa suponen: (1) Identificar el grupo que va a recibir el programa experimental; (2) Localizar un grupo que sea similar al anterior, y del que se puedan recoger también datos, pero que no vaya a estar bajo el programa en concreto; (3) Recoger al menos tres mediciones a intervalos regulares y de forma paralela en ambos grupos; (4) Comprobar la aplicación del programa en el grupo experimental, y que no hay cambios en el grupo de control; (5) Realizar otra toma de datos -al menos tres- una vez terminada la aplicación del programa, y de ambos grupos al mismo tiempo.

Si las condiciones fuesen posibles, este diseño mejoraría altamente al utilizar un verdadero grupo control, con población homogénea y de características idénticas al experimental, donde la asignación de los sujetos a uno u otro fuese completamente aleatoria. Puesto que en ese caso, al control de grupo de añadiría el control de la estabilidad o tendencia de los datos seriales antes y después del programa.

### 3. Definición de metas y objetivos

Los programas intentan conseguir unas determinadas metas, definidas en las características de planificación de dichos programas. De alguna forma, un programa nuevo se pone en marcha porque un grupo de personas, políticos, directivos, administradores, padres, educadores, profesionales de la salud, etc., tienen presentes una serie de metas que el programa en cuestión trata de llevar a cabo. La primera tarea, pues, como evaluador es desarrollar una lista de metas y objetivos, sin ambigüedad y con claridad, entre los intereses del grupo y las realidades donde se aplicará el programa.

Las metas han de describirse específicamente de forma tal que el público o la audiencia pueda afirmar si se han cumplido o no en el programa. Hay que tener cuidado en separar las metas que se describen en resultados, es decir, productos mensurables al final de un programa, de aquellas que definen procesos, es decir, las medidas para su cumplimiento. Por ejemplo, si se intenta llevar a cabo una evaluación focalizada sobre objetivos finales, que compara los resultados de dos programas diferentes, los objetivos comunes de ambos servirán de base para la comparación. Una descripción de una meta ha de ser clara en su significado para la gente implicada en el programa, ha de ser común a los planificadores y a las personas que apoyan económicamente el programa, ha de ser claramente identificable bien al final o durante la aplicación del programa, y ha de ser realista en cuanto a su costo en tiempo, dinero y personal que lo han de conseguir.

El evaluador necesita diseñar las formas de medir esos objetivos a alcanzar

por el programa. Los objetivos de resultados (objetivos conductuales, de ejecución o situacionales) describen la conducta que el programa tiene que haber conseguido al finalizar. El ejemplo siguiente ilustra -en un contexto educativo- las relaciones de especificación entre objetivos globales, específicos y conductuales:

1. *Global*: Los estudiantes han de desarrollar habilidades gramaticales en la escritura.
2. *Específico*: Los estudiantes han de entender y aplicar reglas básicas del uso y concordancia.
3. *Objetivo*: Dada una sentencia simple (sujeto, predicado), los estudiantes escribirán OK si el sujeto concuerda en número con el verbo y, si no, escribir una forma del mismo verbo que pueda corregir adecuadamente la sentencia.

Los objetivos generales y específicos proporcionan una orientación general en la dirección de la planificación, pero sólo los objetivos conductuales pueden ser medidos. La medición de estos objetivos proporcionan la evidencia fundamental de una evaluación. Originalmente los objetivos conductuales estaban formados para proporcionar una guía de evidencias de que las metas educativas estaban o no cumplidas. Un objetivo conductual es una afirmación que si se muestra por el sujeto indica que se ha adquirido esa habilidad, repertorio o conocimiento específico.

La creación de un test o una prueba con objetivos conductuales previamente definidos constituye una medida sensible para detectar los efectos de un programa. Una prueba basada en los propios objetivos del programa, o en combinación con los de un programa en competencia, constituye una excelente medida de hasta qué punto ha sido correctamente realizado. Por el contrario, las pruebas estandarizadas, como las que se realizan al final de un curso escolar, constituyen una medida pobre de los posibles resultados de un programa. Particularmente es deseable la creación de un test cuando se utiliza un diseño de comparación pre-post para evaluar los efectos de un programa. También el mostrar que un programa cumple en sus resultados con objetivos estandarizados proporciona un alto grado de credibilidad sobre la efectividad de un programa.

Existirían varios principios que se han de tener en cuenta para definir objetivos conductuales:

1. Los objetivos se utilizan para describir resultados finales de un programa, no medios o instrumentos, ni actividades.
2. Los objetivos para un programa deben reflejar los diferentes niveles de habilidad conseguidos que se intentan producir con él.
3. Los objetivos han de ser escritos tanto para el programa completo como para individuos.
4. Para resaltar el objetivo, el verbo ha de ser concreto, observable, denotar algún tipo de acción.
5. Los estímulos y respuestas del objetivo han de ser tan explícitos y

detallados como sea posible, incluso conteniendo algunos ejemplos o items de muestra.

En muchos casos, utilizar pruebas estandarizadas lleva al evaluador a hablar sobre objetivos y resultados en un contexto familiar para el público, el personal, los padres, la audiencia en suma. Hay dos formas de asegurar que los tests estandarizados y los objetivos del programa coincidan: fijar los objetivos al test, o analizar el test de acuerdo con los objetivos del programa.

Un test estandarizado puede resultar útil para la evaluación de un programa, siempre que contenga items que midan los objetivos delimitados por el programa, e incluya suficientes items de ese tipo como para constituir una medida válida del cumplimiento de cada uno de los objetivos particulares del programa. Cuando más generales son los objetivos de un programa, mayor probabilidad de que un test cumpla estos dos criterios.

Existen varias técnicas para el muestreo y la asignación de prioridades a los objetivos de un programa. Entre ellas aparecen:

1. Muestreo de objetivos. Se seleccionan aleatoriamente una serie de objetivos del conjunto total. Es muy recomendable cuando todos los objetivos tienen una importancia similar. Puede ser realizado sólo por el evaluador, y es más rápido y más simple que los demás métodos. Trata todos los objetivos del programa como importantes, aunque pueden perderse algunos que reduzcan la credibilidad de la evaluación.
2. Muestreo de objetivos importantes. Dos o tres jueces seleccionan unos pocos objetivos considerados como los más importantes. Muy recomendable cuando en el programa existen objetivos prioritarios y resultan más importantes unos que otros. Es rápido, da a los evaluadores o jueces un papel en la evaluación, aunque puede también que se pierdan importantes objetivos al enjuiciar su prioridad. Focaliza la evaluación en un pequeño número de objetivos y puede que haya errores en los jueces.
3. Matriz de muestreo. Se utilizan todos los objetivos que son asignados a las distintas partes de un test, y cada parte aplicada a un grupo seleccionado de sujetos. Es el único método que permite evaluar todos los objetivos. Puede ser utilizado sólo por el evaluador, para todos y cada uno de los objetivos, aunque la complejidad del proceso y los datos obtenidos a veces impidan el tratamiento estadístico estandarizado.
4. Asignación de prioridades mediante puntuaciones. Un grupo numeroso de jueces puntúan todos los objetivos en una escala 0-5, y la media obtenida determina las prioridades. Especialmente recomendada cuando se necesita la aceptación de la evaluación por varios grupos de personas. Implica conceder prioridad a los criterios de grupo, a las decisiones tomadas por los jueces, con las consiguientes puntuaciones que pueden comprobarse y dar mayor credibilidad a la valoración.
5. Asignación de prioridades en función de una jerarquía de objetivos. Se agrupan los objetivos en áreas, y se ordenan de simples a complejos. Los objetivos terminales más complejos reciben las prioridades más altas.

Recomendable cuando puede elaborarse o es necesaria esa jerarquía. Puede realizarse sólo por el evaluador y concede mayor prioridad a los items complejos, con la consiguiente dificultad de evaluación objetiva, además de requerir un tiempo mayor en su ejecución, en proporción con el número de items totales de la evaluación.

#### 4. Medición de programas

Los métodos de recolección de datos sirven a diversos propósitos en la medición de programas, y no son excluyentes entre sí. Su utilización o no va a depender de la funcionalidad de la evaluación, de la audiencia implicada y del tipo de información que se necesite. La riqueza de datos dará cuenta del grado de precisión y credibilidad dado por la audiencia al programa llevado a cabo.

*Método 1:* Incluir los registros ya especificados del propio programa, en cuanto a material, actividades, etc., y acumularlos tal cual el programa fue ocurriendo, más que reconstruirlo después. Puede suceder que esos registros resulten inadecuados, y entonces sea necesario crear un sistema propio de recogida de registros, asumiendo que se llega al programa en el momento adecuado para evaluarlo, o bien recopilar las versiones de los registros de los tintos participantes en el programa.

*Método 2:* Observar y registrar directamente las conductas implicadas en el programa mediante uno o dos observadores que visiten periódicamente los lugares de aplicación del programa, y utilice registros abiertos o predeterminados. La observación de la situación por un observador tiene una alta credibilidad, aunque cueste mayor dinero y esfuerzo, puesto que el observador registra los eventos que ocurren y cuando ocurren. Pudiendo aumentar la credibilidad de los datos con medidas de fiabilidad y consistencia de los observadores.

*Método 3:* Utilizar medidas de autoinforme, preguntando en entrevistas o cuestionarios, a las diferentes personas implicadas en un programa. Se puede reunir los datos de todos los sujetos que participan en él, una forma abreviada es recoger una muestra de esa población y tomar datos sólo de ellos (administradores, maestros, sujetos, estudiantes, padres, ayudantes, etc.). Las medidas de autoinforme, sin embargo pueden tener problemas en su credibilidad dependiendo de la situación. La información de autoinforme es una recolección de datos post-facto de la conducta de la propia persona que informa, y ello puede llevar a un buen número de sesgos. Siempre, se ha de comprobar la consistencia entre situaciones o personas, y a través de una toma directa de datos.

Otra estrategia consiste en utilizar *datos convergentes*, que requieren las medias múltiples y diversos métodos de recogida de datos, de diferentes situaciones y personas. Antes de crear un sistema propio de medición, conviene observar la posible existencia de programas de observación y cuestionarios, que ya han sido creados y descritos por otros evaluadores, para grupos, clases, unidades, programas similares, etc.

La recogida de información se ha de basar en tres importantes pilares: (1) una lista de actividades, materiales y procedimientos administrativos o técnicas en las que focalizarse; (2) una estrategia de muestreo -lista de situaciones, personas, lugares, etc., con los que contactar-, cuándo y cómo hacerlo; y (3) un plan para la recopilación de los datos y su análisis.

Una *lista detallada de actividades* incluye la prescripción de la frecuencia y duración de tales actividades, así como de su forma -quién, cómo, dónde ocurre-. Los hechos críticos serán los más frecuentemente citados o los que constituyen la parte más larga o importante del proyecto, y son los que servirán para centrar la atención de los datos. Por ejemplo, se puede resumir la información en un diagrama o cuadro comparativo con los diversos apartados, registrando en cada uno de ellos sus características diferenciales.

A veces no es necesario, o no es posible, abarcar todas las situaciones, eventos, participantes, actividades, etc., que se dan en un programa, entonces se hace necesario un muestreo adecuado, que ha de especificarse previamente, sobre todo respecto a dónde, quién, y cuándo observar y tomar registros. Una **muestra representativa** de los lugares donde ocurre el programa ha de servir como muestra de todas las situaciones posibles, y han de demostrar cómo se lleva a cabo el programa -tipo de población, localización geográfica, años, participación, tipo de personal administrativo implicado, nivel de recursos, formación de los profesionales, habilidades de los sujetos, etc.- Además, las variaciones posibles del programa también han de muestrearse para posibles comparaciones. El muestreo de personas va a depender estrechamente de las situaciones elegidas con anterioridad, así como del tipo de medición elegido: aquellas personas que son clave en la aplicación del programa -miembros del grupo o planificadores-, personal administrativo y miembros voluntarios o estudiantes que participen en él. También es fundamental la elección del tiempo de medición, cuándo preguntar o recopilar los datos, sobre todo si el proyecto tiene diferentes fases, o cambia durante su aplicación. En estos casos se ha de dividir ese tiempo total en segmentos iguales, representativos de todo el programa, y recabar información durante ese periodo de la población y situaciones previamente elegidas para muestrear.

En cuanto a *la forma de recopilación de datos*, puede haber dos razones para ello: (1) describir el programa y comentarlo en cuanto a su funcionamiento y seguimiento del plan previsto; (2) examinar las relaciones entre las características del programa, los resultados obtenidos y las características diferenciales en la aplicación concreta del programa en cada centro. Las descripciones de un programa generalmente son presentadas en forma de narraciones o tablas

---

descriptivas. Para resumir los datos, puede resultar adecuado utilizar una "hoja de resumen de datos" para cada instrumento utilizado, que pueden ayudar a identificar patrones de respuestas o características especiales en la aplicación del programa. Este tipo de hoja resumen requiere que los datos sean de respuestas cerrada, o que las respuestas estén categorizadas o codificadas, si no lo están el primer paso es intentar categorizar o construir códigos para la cuantificación de la información obtenida en situaciones naturales, entrevistas o cuestionarios abiertos.

Para realizar la cuantificación de los datos puede elegirse un sistema de conteo rápido a mano, donde todas las opciones de respuesta estén ya contenidas en la hoja y se puedan contar los datos directamente. Puede utilizarse para calcular (1) el número o porcentaje de personas que han contestado a cada ítem de una forma determinada, y (2) la media de respuestas en cada ítem o pregunta. A veces, se calculan datos de correlación, y entonces se tiene que conservar los datos brutos obtenidos en cada ítem y persona preguntada, para mantener la información individual no global. A partir de estas hojas ya es posible calcular otros datos de medias, correlaciones, descripciones o patrones de respuesta por ítem, etc. También es posible utilizar procedimientos para el procesamiento mecánico o computerizado de esos datos, aunque siempre habría que adaptar la forma o entrada de los datos a las características del programa de ordenador, y a veces es más breve y menos costoso una recopilación manual de los datos brutos.

Si lo que se ha obtenido son ítems de cuestionarios abiertos, entrevistas, registros narrativos, etc., se hace necesario manejarlos de una forma sistemática para obtener conclusiones de ellos. Así se puede: (1) realizar hojas resumen de datos; (2) seleccionar eventos específicos a registrar o entresacar de cada informe; (3) registrar los eventos característicos que se repiten en varios de los informes; (4) preparar resúmenes con las afirmaciones o frases más frecuentes.

También se pueden categorizar esas afirmaciones o ítems cualitativos asignándoles valores numéricos a diferentes tipos de respuestas, bien descritas y especificadas. Para realizar esas categorizaciones se pueden seguir los siguientes consejos: (1) elegir las categorías como componentes de una dimensión que puede variar a lo largo de una gama de valores; (2) releer los datos y muestrearlos para decidir si los datos descritos se ajustan a las categorías definidas; (3) refinar alguna categoría o englobar algunas cuando se considere necesario; (4) probar a más de una persona independiente para que compare las categorías, su adecuación y definición en función de los datos.

Cuando los datos sí son cuantitativos -observaciones, ítems descriptivos, respuestas cerradas, etc.- pueden realizarse análisis estadísticos con ellos, aunque previamente se ha de decidir cuál es la pregunta o la información que se quiere obtener de tales análisis, sobre todo si interesa obtener determinados "índices" de ítems específicos, de la relación de unos datos con otros, o del

grado de aplicación ajustada de un programa sobre el proyecto inicialmente realizado.

Con objeto de resumir los datos de ítems individuales, se podrían utilizar los datos totales, porcentajes, medias, media de grupo, etc. En algunos casos, se pueden analizar directamente, en otros se puede actuar con ellos como porcentajes, y éstos han de ser mostrados al auditorio como gráficos de distintos tipos y características en función de los datos a presentar.

## 5. Técnicas de evaluación

### 5.1. Registros

Los registros constituyen los productos permanentes de los efectos de un programa determinado, y permiten construir una descripción creíble de lo que ha sucedido durante su ejecución. Cuando estos registros están ya realizados en el propio programa se pueden utilizar como fuente de información de las actividades, así cosas como las gráficas del progreso escolar, registros de asistencia, cumplimiento de horarios, formularios rellenos, etc., constituyen un material del que obtener información. Otra opción consiste en crear la fórmula propia de registro durante el programa.

En el primer caso -registros ya existentes- se pueden seguir los siguientes pasos:

1. Construcción de una lista de las características del programa.
2. Pedir y seleccionar del material obtenido por los programadores y participantes, el que resulte válido para su registro.
3. Comparación de los datos entre ambos, especialmente en aspectos como la duración, frecuencia, la forma de las actividades, y la extensión de la población a la que afecta el programa.
4. Preparar un plan de muestreo de los registros, periodos típicos, focalizados o aleatorizados, en la duración del programa.
5. Crear un sistema de recolección de datos, y planificar la transferencia de los datos y registros obtenidos a este sistema de recopilación.
6. Planificar un sistema de acceso a la información de los registros que no interfieran o dificulten la labor de los planificadores.

En el segundo caso -cuando se crea un sistema de registro a propósito para ese programa- se pueden seguir los siguientes pasos:

1. Construir unas listas de las características del programa.
2. Entresacar los registros, explicarlos a los planificadores y ejecutores del programa, que se van a utilizar después.
3. Revisar detalladamente cada una de esas características, refinarlas y ajustarlas a las reales y efectivas en la ejecución del programa, y también en relación a las posibilidades de registrarlas en la situación prevista (tiempo de ejecución, utilidad, costo del registro, etc..) Tratar siempre de evitar solapamiento en los datos, acumular una gran cantidad de datos sin propósito o sin un plan previo.

4. Preparar un plan de muestreo de los registros, con objeto de facilitar su aplicación, seleccionando las situaciones, clases, personas y periodos de tiempo característicos en los que se van a realizar los registros.
5. Crear un sistema para transformar los datos en una hoja resumen, con objeto de calificar el programa como un todo.
6. Planificar un sistema de fácil acceso a los registros que se necesiten.

## 5.2. Observación

Los informes de observadores están basados en lo que ven directamente cuando el programa está en funcionamiento, y la mayoría de la gente da mayor credibilidad a este tipo de datos al ser obtenidos por un observador externo e independiente de los miembros del programa. Algunos evaluadores, incluso, consideran que la observación es el único método para obtener y aplicar la evaluación con la extrema complejidad de muchos programas.

El nivel formal de la observación a realizar -desde una observación casual a una muy estructurada- constituye el punto crucial para el inicio y planificación de un sistema de observación. La credibilidad de la información demandada que se realicen observaciones siguiendo una metodología estricta, de forma tal que la audiencia nunca pueda rebatir que la persona observadora no estaba preparada y no sabía qué observar, o había malinterpretado los datos, o no había observado en el lugar y tiempo adecuados. Los métodos de observación formales son cuidadosos en el momento de planificar cuándo han de mirar los observadores, qué ha de ver y cómo registrar esa información. Estos métodos pueden ser de dos tipos fundamentales: (1) observación sistemática, en la que el observador sabe qué observar y cómo registrarlo; (2) observación responsiva-natural, donde el observador llega a la situación sin una predisposición a qué observar y cómo describirlo.

Para la realización de un sistema de observación sistemática, se podrían seguir los siguientes pasos:

1. Construir una lista de las características del programa, una lista del material, actividades, personal administrativo y profesional, etc., que participan en el programa.
2. Preparar los escenarios donde ocurran los episodios típicos del programa. Es decir, pequeñas dramatizaciones con acciones específicas con objeto de definir lo que se va a observar posteriormente en la situación real del programa. Fundamentalmente, se han de dar las diferentes definiciones de las conductas, personas, situaciones, interacciones, estímulos ambientales, etc., a observar.
3. Preparar los escenarios de episodios alternativos o raros que ocurran en el programa. Una lista de cosas que podrían no funcionar bien, o hacerse de otra forma alternativa de registro no diseñada previamente. Ambos tipos de escenificaciones -con sus definiciones correspondientes- se enseñan al equipo y se redefinen o rectifican en función de los comentarios sobre lo que efectivamente se pretende que ocurra en el programa.

4. Escoger un método de observación. Los tipos fundamentales de instrumentos para el registro son:
  - 4.1. Listados específicos de eventos, realizados a propósito de una actividad, y que sirven para registrar la presencia, ausencia, frecuencia, duración, de un corto número de conductas tal como ocurren, aunque no se utilicen para registrar la forma o cualidad de esos acontecimientos como ocurren. Permiten cuantificar los eventos que se han producido, lo que facilita su manejo e interpretación posterior, y son fáciles de utilizar con observadores entrenados. Técnicas de registro de frecuencia y duración, utilización de muestreo de intervalos.
  - 4.2. Registros de categorías, sirven para obtener detalles cuando ocurren un gran número de conductas en un periodo determinado de tiempo, y permiten no sólo registrar qué eventos han ocurrido sino también su frecuencia, e incluso la secuencia que tiene lugar en ellos. Se asignan símbolos o códigos a las diferentes conductas a observar, se enseñan a los observadores y se procede a su entrenamiento, y posterior utilización en la situación real. Sin embargo, este tipo de datos resultan más difíciles de cuantificar, resumir e interpretar. Los registros de categorías obtienen un listado de secuencias de conductas tal cual ocurren, de las que se han de extraer los datos posteriormente. Resultan más recomendables cuando el interés reside en obtener datos de interacciones entre personas o secuencias de eventos, cuando se necesita registrar una gran cantidad de eventos diferentes, y cuando es posible utilizar otros registros videográficos que permitan su registro directo posterior. Las categorías lógicamente deberán estar comprendidas entre una serie limitada de símbolos que sean aprendidas y recordadas por el observador, y no produzca confusión entre ellas.
  - 4.3. Productos permanentes o registros demorados de los eventos. Este tipo de registros se obtienen siempre después que ha terminado el periodo de observación. Dada la cantidad de pérdida de información que se puede producir, además de requerir inferencias o juicios propios de los observadores, constituyen esencialmente "cuestionarios", más que técnicas de observación sistemática. Estos requieren que los observadores estén informados previamente de lo que han de observar, y las sesiones han de ser lo más formales o estructuradas posibles para su posterior comparación en un mismo tipo de datos; además de tener definido previamente qué va a ocurrir, durante cuánto tiempo y quién lo va a observar y registrar posteriormente en la sesión. Pueden ser de mayor utilidad cuando interesa obtener datos periódicamente o de escenas específicas, o bien cuando los eventos han de ser observados sin que interfiera su registro inmediato en la situación o en el sujeto observado.

5. Decidir la duración de la muestra de tiempo en la que se va a registrar. En general, el muestreo de tiempo deberá ser lo suficientemente largo como para registrar los eventos que interesan, pero si éstos ocurren muy frecuentemente la duración de la observación puede ser menor. Esta longitud vendrá dada por el número de observaciones necesarios para obtener la información deseada, especialmente que en esa duración ocurran las circunstancias o situaciones elegidas, en las personas elegidas y en tiempo y fecha en que interesa o estaba planificado en el programa.
6. Preparar un plan de muestreo para realizar las observaciones. El muestreo de situaciones, características y personas va a depender del programa planificado, de sus aspectos más relevantes. Puede realizarse una preselección de esas situaciones y personas, focalizarse sólo en algunas, o realizar un muestreo al azar, o bien un muestreo de individuos particulares en interacciones de grupo.
7. Preparar las hojas de registro para los observadores. Los observadores generalmente han de tener el sistema específico donde realizar anotaciones y registros, siguiendo el modelo y los objetivos propuestos para cada observación. Ello implica la definición de los eventos a registrar, eliminar o ajustar los items en el lenguaje, en suma hacer el instrumento simple para que realmente pueda ser útil.
8. Escoger observadores, con objeto de poder hallar fiabilidad de los datos, resulta conveniente utilizar al menos dos observadores.
9. Entrenar los observadores y ensayar el sistema de registro. Los observadores han de tener práctica y experiencia previa en la observación de los eventos que van a ocurrir y en la técnica específica que van a utilizar. Para ello, el role-playing, cintas de video y las prácticas en otro tipo de observaciones, constituyen técnicas adecuadas, que a veces permiten si es necesario rectificar las definiciones realizadas, el sistema de registro o el muestreo a efectuar.
10. Informar al equipo del programa sobre los resultados de las observaciones. Bien con entrevistas personas, en grupo, o haciendo circular informes por escrito entre los miembros del programa. Generalmente, resulta necesario la presentación previa de los observadores y dar a conocer el propósito explícito de las observaciones; en todo caso siempre es necesario el consentimiento explícito de las autoridades del programa en cuestión.
11. Realizar las observaciones. Asegurarse de que los observadores llevan a cabo la tarea tal como estaba planificada, con especial atención a las dificultades del registro, las posibles discrepancias o inconsistencias entre observadores, con objeto de su eliminación para las siguientes observaciones.
12. Cuantificación y tabulación de los datos una vez obtenidos de los registros. Transformar los datos individuales en hojas de datos globales.

Cuantificar y obtener los datos a partir de los códigos o categorías registradas, etc., mediante hojas-resumen y hojas de tabulación.

### 5.3. Autoinformes

Una buena forma de averiguar cómo transcurre un programa es preguntar a las personas directamente implicadas, se les puede entrevistar individualmente o en grupo, y también se puede crear y circular entre ellos cuestionarios preguntando sobre eventos específicos, descripciones de sus propias experiencias. Los autoinformes pueden ser de dos tipos: (1) autoinformes periódicos realizados a lo largo del programa, y (2) informes retrospectivos una vez acabado. Generalmente los primeros permitirán obtener una información más específica y fiable, pues se preguntan por las actividades cuando éstas van ocurriendo. Los informes retrospectivos pueden ser mejor utilizados cuando no hay más remedio: el programa es muy corto o su reconstrucción posterior es posible y fiable con los datos obtenidos.

Para realizar un autoinforme sobre las características de un programa se pueden seguir estos pasos:

1. Decidir cuántas veces se aplicarán los cuestionarios. Esta decisión vendrá dada por la homogeneidad de esas actividades, la tolerancia del personal del programa a las interrupciones de su actividad y su motivación para responder, y la cantidad de tiempo necesario para recopilar los datos, puntuarlos e interpretarlos.
2. Avisar a las personas o equipo del programa de los cuestionarios y su aplicación periódica, junto con aspectos motivacionales por su participación en ellos.
3. Preparar una lista de eventos críticos en el programa (material, personas, actividades, ajustes administrativos, etc.).
4. Decidir si se tomarán datos con entrevistas, cuestionarios o ambos. Probablemente si los cuestionarios son la única fuente de datos, se necesitarán también entrevistas individuales a los miembros y participantes en el programa.
5. Escribir preguntas basándose en la lista de características y actividades. También se necesita especificar el grado de participación, uso de materiales y formación de las personas a las que se pregunta. Especialmente preguntas sobre la ocurrencia de determinadas actividades, su frecuencia y duración, la forma explícita de tales actividades, y el trabajo o implicación del sujeto en el programa. Para diseñar cuestionarios se pueden utilizar formatos de preguntas cerradas o abiertas, estructuradas, con respuestas previstas, o con escalas de juicio y calificaciones.
6. Componer la forma de diferentes cuestionarios o entrevistas. Generalmente se requieren instrucciones en el mismo cuestionario sobre su objeto, vocabulario claro y descriptivo, incluso instrucciones de retorno y fecha límite del cuestionario si es entregado por correo.

7. Ensayar el instrumento, con otras personas que lo lean en voz alta, observen sus respuestas y lo comenten; y si es necesario cambiar preguntas o frases que no estén claras o bien definidas. Contrastar la información con diferentes personas.
8. Administrar el cuestionario de acuerdo con el plan de muestreo. Si es por correo especificar fecha de retorno y sujeto al que va dirigido. Recabar la colaboración y ayuda de los responsables o directores del programa, para aumentar la probabilidad de respuesta.

## 6. Fiabilidad y validez de las medidas

Los instrumentos para describir la aplicación de un programa a menudo son puntuados e interpretados ítem a ítem, en estos casos los instrumentos son de comparación individual, las preguntas simples son respondidas con datos simples, en el que quizás la media indique o represente a todos los demás datos de otros sujetos en el mismo programa. La fiabilidad y validez hacen referencia a diferentes aspectos de la credibilidad de una medición, la primera responde al hecho de si el instrumento mide consistentemente los mismos resultados, y la segunda responde a si el instrumento resulta apropiado para lo que se necesita medir. De esta forma un instrumento de observación, un cuestionario o un programa de preguntas será fiable en el momento en que obtenga los mismos resultados cuando se administra en las mismas situaciones varias veces. Pero el que un instrumento sea fiable no supone necesariamente que sea bueno para medir lo que se desea medir. Se han de responder a cuatro cuestiones: ¿Es la aplicación del programa precisa, relevante, representativa y completa?.

Una instrumento preciso ha de dar a la audiencia de la evaluación una idea lo más aproximada posible de lo que constituye el programa y sus resultados. Una medida relevante ha de tomar datos y poner de relieve los hechos más críticos o fundamentales del programa (hace referencia al concepto de validez de constructo). Una descripción representativa ha de incluir un ejemplo típico de las características del programa y de sus variaciones entre situaciones y tiempos de aplicación. Y una evaluación completa es aquella que incluye todos y cada uno de los hechos importantes y relevantes en la aplicación del programa (hacen referencia a las técnicas de muestreo de situaciones y tiempo, contrastación entre métodos, estandarización de la aplicación, índices de relevancia y juicios de expertos).

La fiabilidad hace referencia a la extensión con que los resultados de la medición están libres de error o fuentes impredecibles de variaciones, fuentes de error que pueden provenir de: fluctuaciones en el modo de respuesta, variaciones en las condiciones de aplicación, diferencias en la puntuación del observador o en las definiciones utilizadas, o efectos aleatorios introducidos por los sujetos en la aplicación concreta del instrumento. Generalmente los

métodos para contrastar la fiabilidad vienen dados por la comparación del instrumento cuando se aplica varias veces (en test-retest, en formas paralelas o en dos mitades de la misma prueba).

Con observaciones, la mejor forma de demostrar que la evaluación está mínimamente contaminada por inconsistencias de "instrumentos humanos" es utilizar más de un observador, bien en la situación presente o en vídeos o registros filmados y evaluados posteriormente.

## 7. Informes de aplicación

Tanto si se realiza una evaluación formativa como sumativa, se ha de preparar un informe de aplicación detallado, enumerando las cuestiones fundamentales y respondiendo a ellas en el grado de detalle en que se necesitaría en la evaluación y el nivel de esfuerzo necesario para realizarlo. La mayoría de las evaluaciones suponen un papel importante a los informes de aplicación al especificar tanto el diseño de la evaluación, los resultados previstos del programa, las medidas utilizadas para registrar los posibles resultados, así como los datos o mediciones específicas producidas. Las cuestiones, pues, fundamentales a incluir en este informe serían:

1. Un resumen que proporcione al lector una rápida sinopsis del informe, explicando la aplicación a realizar, los resultados globales y sus conclusiones. A veces con sólo una o dos páginas para "lectores ocupados".
2. Una descripción del contexto en que se aplicará el programa, focalizándose en las situaciones, los cambios administrativos, el personal y los recursos implicados. Si la audiencia no conoce el programa, éste ha de ser descrito exhaustivamente. Si se aplica en varias situaciones o centros, explicar las similitudes y diferencias entre ellos en el momento de aplicar el programa. Incluyendo la situación ambiental del programa (centros, locales, comunidad, personal, economía, ayudas, etc.), los orígenes del programa (cómo empezó, quién lo pidió, necesidades del grupo, padres, maestros, etc.), los objetivos, la historia y su fundamentación, los objetivos de grupo dentro del programa (edades, niveles, habilidades, participación, agrupación de los sujetos a los que se aplica, etc.), el personal del programa (tipo, número y función, padres, familia, etc.), cambios o tareas administrativas (posiciones jerárquicas, recursos, becas, soporte administrativo, etc.), y los gastos generados (material, costo, becas o recursos para comprarlo, desarrollo o utilización del material, entrenamiento del personal, etc).
3. Una descripción de los puntos de vista con los que se realiza la evaluación, bien una aproximación naturalista o una aproximación con un programa prescrito o un modelo concreto al aplicarlo. Qué racionalidad tiene el programa, qué bases teóricas, qué planificación subyacente no descrita en el programa, etc.

4. Una descripción de la aplicación de la evaluación por sí misma, una discusión sobre las actividades del programa seleccionadas y las medidas escogidas. Incidiendo sobre la focalización en la aplicación de la evaluación (audiencia, contexto del estudio, los hechos críticos o característicos, comparaciones, fuentes de datos fundamentales, etc), y el rango de mediciones y el registro de datos (instrumentos, selección y desarrollo de instrumentos, fiabilidad, validez, programa de recogida de datos, muestreo, lugares, situación y personas que registras, etc.).
  5. Resultados de las medidas globales de aplicación, y discusión de la aplicación del programa. Con consideraciones generales (el periodo de aplicación, sucesos o hechos cruciales en la aplicación, materiales utilizados, etc.), y hallazgos específicos y conclusiones para cada grupo de materiales o partes de cada programa (objetivos alcanzados, utilización efectiva del material, adecuación con el plan trazado, utilización de los recursos, adaptación, desarrollo siguiendo las pautas del programa, medidas tomadas, actividades conseguidas, instrucciones, tiempo de desarrollo de cada plan, procedimientos, progresos, conocimientos del personal, comparaciones entre grupos, etc).
- Fundamentalmente describir la extensión con que el programa ha llevado a cabo los objetivos propuestos o planificados de antemano; y la descripción de la aplicación independientemente de la planificación o los modelos o teorías propuestos, utilizando una descripción realista- naturalista de lo observado durante la aplicación del programa, con las decisiones propias del evaluador. Describir también lo encontrado en función de las variaciones en situaciones, personas, sujetos del programa, etc., incluyendo sugerencias y posibles modificaciones para el desarrollo de otros programas de evaluación.

## Referencias

- |  |   |
|--|---|
| Cooley, W.W. y Lohnes, P.R. (1976)<br>Evaluation research in education. New<br>York: Irvington Publ. | CA: SAGE Publ.  |
| Morris, L.L. y Fitz-Gibbon, C.T. (1978)<br>Program Evaluation Kit. Beverly Hills,                    | Stake, R.E. (1975) Evaluating the arts in<br>education: A responsive approach.<br>Columbus, Ohio: Charles E. Merrill<br>Publ. |