REPÚBLICA DE CHILE
MINISTERIO DE EDUCACIÓN
CENTRO DE PERFECCIONAMIENTO
EXPERIMENTACIONES
INVESTIGACIONES PEDAGÓGICAS
DEPARTAMENTO DE EVALUACIÓN

CÓMO SELECCIONAR, PRESENTAR E INTERPRETAR LOS RESULTADOS DE UN DISEÑO PARA EVALUACIÓN

CENTRO DE ESTUDIOS DE EVALUACIÓN DE LA UNIVERSIDAD DE CALIFORNIA, LOS ÁNGELES UCLA (CSE)

Traducción: Marta Soto Rodríguez (CPEIP)

Santiago - Chile

CÓMO SELECCIONAR, PRESENTAR E INTERPRETAR LOS RESULTADOS DE UN DISEÑO PARA EVALUACIÓN

LYNN LYONS MORRIS, CAROL TAYLOR FITZ-GIBBON

Centro de Estudios de Evaluación
Universidad de California
Los Ángeles UCLA

Publicación del Departamento de Evaluación
Centro de Perfeccionamiento, Experimentación
e Investigaciones Pedagógicas

Traducción: Prof. Marta Soto Rodríguez.

Doc. Nº 0661, -

SANTIAGO – CHILE 1979

PREFACIO

En los momentos en que se llega a las postrimerías de la década de los años setenta, ya no se discute si se deben o no se deben realizar evaluaciones. El problema es cómo realizarlas para que respondan a los distintos requerimientos.

Mientras que para un director de establecimiento es muy importante conocer los resultados "globales" obtenidos con la aplicación de un nuevo programa, para los profesores puede resultar de gran interés detectar si los materiales de que se dispone son adecuados para el logro de determinados objetivos por parte de los alumnos. Para los padres resulta fundamental tener antecedentes sobre el impacto que el programa ha tenido en sus hijos y saber si dicho impacto ha sido positivo. Vemos así que una evaluación tiene que responder adecuadamente a los intereses e inquietudes de distintos tipos de personas para que se le considere útil y para que oriente el trabajo futuro.

No hay "único" método de evaluación para cada ocasión. Tanto las diferentes situaciones que se presentan en el contexto en que debe realizarse la evaluación -- recursos, tiempo, variables que se deseen considerar, etc. -- como las informaciones que se necesitan y como las características de las audiencias y del evaluador, han hecho necesario el desarrollo de una serie de estrategias de evaluación.

La evaluación que se desarrolla durante la ejecución de un programa con el fin de usar sus resultados para revisar y adecuar lo que se está realizando – evaluación formativa -- tiene características distintas a la evaluación que se realiza una vez que el programa ha terminado - evaluación acumulativa -- ; del mismo modo no se puede esperar que el método o estrategia de trabajo que se utilice para llevar a efecto una evaluación informal sirva para realizar una evaluación formal; ni se puede esperar que la forma de conducir una evaluación para un caso particular sea la misma que se utilice para conducir una evaluación cuyo principal propósito sea generalizar los resultados.

Lo anterior es también válido para evaluaciones que sean descriptivas o que se limiten a juzgar resultados o procesos; para evaluaciones que sean internas o evaluaciones que sean externas; para evaluaciones que sean globales o para evaluaciones que apunten a un solo aspecto o factor, de un programa, de un material, de un proceso, etc.

Sin embargo, a pesar de la variedad de posibles estrategias de evaluación, hay un aspecto que tiene particular importancia para la efectividad de ella. Dicho aspecto corresponde al Diseño que se utilice en la evaluación.

La selección de un diseño adecuado es un paso importantísimo para conseguir información de buena calidad y que sea de utilidad para los fines de la evaluación. El diseño responde, básicamente, a las dos preguntas siguientes: ¿qué personas o grupos de personas originarán la información?, y ¿cuántas veces y en qué momentos se harán las observaciones que originan la información evaluativa?

Las ventajas, limitaciones y consideraciones espaciales que deberán tenerse en cuenta al enfrentarse con la tarea de seleccionar un diseño para la evaluación, se discuten en este libro.

Este libro fue elaborado originalmente en el Centro de Estudios de Evaluación de la Universidad de California de Los Ángeles (UCLA). Gracias a la respuesta positiva que la Directora de dicho Centro, Dra. Eva L. Baker, diera a mi solicitud de seleccionar material producido en el Centro y traducirlo al español con el fin de utilizarlo en los Cursos que el Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas de Chile ofrece a los profesores latinoamericanos, y gracias al apoyo entusiasta y efectivo de la subdirectora del Centro de Estudios de Evaluación, Dra. Adrianne Bank, y de la Codirectora del Proyecto dentro del cual se escribió este libro, Dra. Lynn L. Morris, fue posible realizar la traducción del material que constituye esta publicación.

Agradezco muy sinceramente a la Dra. E. Baker, a la Dra. A. Bank y la Dra. L. Morris por el valioso apoyo prestado para que se hiciera esta publicación y a la Profesora Marta Soto Rodríguez quien al tener la oportunidad de trabajar junto a la Codirectora del Proyecto – Dra. L. Morris – cuidó con interés y dedicación cada detalle de la traducción.

Agradezco también al Profesor Javier Zabalza por el tiempo dedicado a la lectura de la traducción y a la Srta. Carmen Henríquez cuyo paciente y cuidadoso trabajo de dactilografía y composición permitió, finalmente, poner este libro a su disposición.

Si las informaciones, ideas y sugerencias incluidas en este libro le son de utilidad en la planificación y desarrollo de evaluaciones, se habrá cumplido el propósito del trabajo realizado en la traducción y se habrá recompensado el esfuerzo y el tiempo que entregué a la revisión técnica del material.

RAFAEL HERRERA RUIZ
JEFE Departamento de Evaluación
C. P. E. I. P.

ÍNDICE

CAPÍTULO 1:

LOS ELEMENTOS DEL DISEÑO

Un diseño de evaluación es un plan que establece a *quién* se mide y *cuándo* se mide, es decir, qué *grupos* de alumnos se van a medir y en qué *momento*. Este capítulo describe los dos elementos mencionados: grupos y momentos; además examina algunas consideraciones relevantes de cada elemento.
La Tabla 1 que aparece en la página 32 muestra varias de las formas en que estos elementos pueden combinarse para construir los seis diseños que se presentan en este libro.

GRUPOS

Conviene señalar inmediatamente que al hablar sobre cuáles son los "grupos" que se miden con los diseños de evaluación, la palabra "grupo" se usa en un sentido especial, como una abreviación para "grupo de tratamiento". Es decir, un "grupo" de alumnos quiere decir aquellos alumnos que toman el mismo programa. De este modo, el "grupo experimental" se compone de todos los alumnos que toman el programa experimental. De vez en cuando hay confusión con el significado de "grupo" debido a otros usos. Por ejemplo, una persona pudiera decir, "dos grupos recibieron el programa: un grupo de niños y un grupo de niñas". En términos de diseño sería mejor decir que el grupo experimental –aquellos alumnos que recibieron el programa- podrían clasificarse como niños o niñas porque el sexo fue un "factor de clasificación". La palabra "grupo" *se reserva* para un grupo definido por el "tratamiento" (programa) recibido. Ocasionalmente, "grupo" podría referirse a un grupo de cursos que se han seleccionado para un programa, más bien que a un grupo de alumnos considerados individualmente.

Al planificar una evaluación, usted siempre intentará medir el "grupo experimental", por ejemplo, el grupo de alumnos que recibe el programa que se va a evaluar. En este libro se usará con frecuencia la abreviación "grupo-E" para referirse al "grupo experimental".

Cuando *solamente* se mide el grupo experimental, la interpretación de los resultados resulta difícil y a menudo poco convincente. Sin tener un grupo de *comparación* es difícil saber cuán buenos son los resultados, si los resultados hubieran sido tan buenos con cualesquier otro programa e incluso, si el programa tuvo siquiera algún efecto sobre los resultados. Por consiguiente, es muy recomendable que usted use un grupo "control". Por grupo control se entiende un grupo que se compone de alumnos que son lo más parecido posible a los que constituyen el grupo-E y a quiénes se mise *en el mismo momento que se mide* al grupo-E, pero que no reciben el programa experimental. En este libro se usará a menudo la abreviación "grupo-C" para referirse al "grupo control". El hecho de que el grupo control no reciba el programa experimental no significa que el grupo no reciba el programa. El problema relacionado con el *tipo* de programa que recibe, exactamente, el grupo control se considera en la sección que empieza en la página 15.

Grupos Control

Los grupos control se presentan de muchas maneras. Dependiendo de las circunstancias, un grupo control podría componerse, por ejemplo, de seis alumnos, o de tres cursos diferentes, o del colegio que hay cerca, o de los niños con defectos físicos del distrito escolar vecino, etc. La situación ideal es que el grupo control y el grupo experimental sean tan parecidos como es posible que sean dos grupos de alumnos –iguales en todo sentido. En la realidad, por supuesto, los grupos control pueden ser diferentes del grupo experimental en diversos grados- algunos muy diferentes, algunos muy parecidos; tratar de descubrir, exactamente, cuán semejantes son dos grupos sería una tarea interminable. Usted podría ver si son semejantes con respecto a sexo, CI verbal, aptitud no verbal, antecedentes del hogar, conocimiento de matemática básica, perseverancia, gusto por el colegio, cooperación, etc. Una serie de características como éstas *podrían* afectar la reacción de un alumno hacia un programa. Por consiguiente, mientras la semejanza entre dos grupos podría estar en un continuo que va desde muy semejante a muy diferente, nosotros no tratamos de determinar este continuo. Simplemente dividimos los grupos control en dos clases: aquellos que se hacen *equivalentes* por asignación al azar, lo que generalmente asegura una distribución objetiva de las diversas características, y aquellos que no emplean asignación al azar y por lo tanto deben considerarse *no equivalentes*. Ahora describiremos y explicaremos cada uno de estos tipos de grupo control y daremos ejemplos de cada uno de ellos.

El grupo control equivalente o "verdadero"

Un grupo control "verdadero" es aquel grupo que se forma por asignación aleatoria o al azar. Aleatorización o asignación al azar *es la forma de hacer un grupo control equivalente o verdadero*.
El mejor diseño de evaluación que usted puede implementar, es el que emplea grupo control verdadero. Esto es porque, en general, los resultados que se obtienen con dicho diseño no pueden haber sido causados por ninguna otra cosa que no sea la diferencia en tratamiento dado a los grupos experimental y de control.

Suponga, por ejemplo, que una evaluación ha mostrado que los alumnos del programa X (el grupo experimental), en promedio, tuvieron puntajes más altos en el postest que los alumnos del programa C (los alumnos del programa C forman un grupo control). Los resultados obtenidos podrían cuestionarse con comentarios como los siguientes:

"Bien, los niños en el programa X eran más inteligentes".

"Ese curso fue mejor porque los padres de muchos de los niños que están en dicho curso son profesionales. Es el curso que forma la mayor parte de la orquesta del colegio",

"El grupo control empezó más abajo que el grupo que está recibiendo el programa X, de tal modo que es natural que ellos terminaran más abajo".

La asignación al azar de alumnos a los programas es la forma más efectiva de eliminar tales interpretaciones. Las asignación al azar evita interpretaciones alternativas haciendo probable, cuando se van a comparar dos os más programas, que los factores que pueden influir en los resultados –inteligencia, antecedentes de los

padres, nivel de rendimiento en el pretest-. Se distribuyan en forma pareja en los dos programas desde un principio. La asignación al azar de alumnos a los dos programas, por ejemplo, hará que los dos grupos sean casi igualmente inteligentes y de semejantes antecedentes del hogar.

La aleatorización de alumnos también eliminará todo tipo de otros factores que usted podría no haber considerado jamás, pero que pudieran influenciar los resultados. Por ejemplo, la asignación de alumnos al azar distribuirá en forma pareja a los niños que viven en el mismo sector. Este puede resultar ser un sector donde, sin que usted lo sepa en septiembre, habrá una epidemia de influenza el próximo mes de enero, causando muchas inasistencias prolongadas. Todas esas inasistencias a *un* programa podrían ser devastadoras en sus efectos. Si la influenza afecta *a los dos* programas su evaluación aún tiene sentido, ya que aproximadamente el mismo número de niños estará ausente de ambos programas; aún es posible hacer la comparación a fin de año. Ninguno de los programas alcanzará todo lo que pudiera haberse esperado, pero su capacidad para compararlos no se deteriorará seriamente y esta comparación es la esencia de su evaluación.

Usted tiene razón si ha pensado que la aleatorización puede ser efectiva para igualar grupos cuando el número de alumnos o equipos involucrados es grande, pero que es poco probable que sea efeciva a medida que los números se hacen más pequeños. En el caso de grupos muy pequeños (menos de 15 alumnos en cada grupo cuando usted está haciendo grupos separados), se deben hacer consideraciones especiales para las asignaciones al azar.
Estas consideraciones se discuten en el Capítulo 7 que describe procedimientos para la asignación de alumnos al azar a cada grupo.
Es deseable que se efectúe la aleatorización no sólo por el efecto que tiene en hacer los grupos equivalentes, sino también porque *los tests estadísticos* de los resultados requieren que haya habido aleatorización para su aplicación adecuada e interpretación exacta.
Como la aleatorización es la mejor manera de eliminar interpretaciones alternativas de los resultados que usted desea atribuir a un programa, la selección o asignación de alumnos al azar a los programas es la forma más deseable para diseñar una evaluación, asegurando alta credibilidad a lo que usted descubra.
Aunque la aleatorización es importante para fortalecer los resultados de la evaluación, usted pudiera encontrar resistencia al tratar de usarla en situaciones prácticas. Los párrafos siguientes sugieren algunas estrategias para vencer las objeciones más comunes.

La estrategia de dos programas nuevos. Debido a que los programas nuevos despiertan interés y entusiasmo, aunque eventualmente ellos pueden resultar pobres, usted puede pensar que es mejor usar un segundo programa nuevo como control para el programa X: asigne alumnos al azar a uno o al otro programa nuevo. Este otro programa "nuevo" fácilmente podría ser una versión del primero, al que se sustrae o aumenta algún rasgo importante o costoso del primero.
También podría ser un competidor verdadero y totalmente diferente del primero. (Un análisis más completo de lo que podría ser este programa aparece en la página 16). Al final, la evaluación mostrará cual de los dos programas nuevos fue el más efectivo, o si el rasgo que usted ha eliminado o modificado causó realmente una diferencia.

La estrategia del grupo control intermedio. Esta estrategia es adecuada cuando a los alumnos "Más necesitados" se les debe dar el programa X, tal como cuando el programa X es un programa correctivo o un programa de educación especial para un grupo con una necesidad especial. La medición de quién está "más necesitado" no es nunca perfectamente exacta. Hay casi siempre un gran grupo de alumnos "intermedio" entre los que es difícil decidir quienes son los más necesitados. Si se usan los puntajes de un *test* determinado para decidir quién toma el programa, los alumnos intermedios serán aquellos cuyos puntajes se desvían sólo unos pocos puntos del puntaje límite de separación. En tal caso los alumnos *más* necesitados, no intermedios, pueden asignarse al programa X, pero el grupo intermedio puede ser asignado al azar al programa X, o al grupo control. Esto permite una evaluación cabal del valor del programa para los alumnos intermedios, información que es en gran medida pertinente para decidir si se debe o no se debe expandir el programa.
Algunos procedimientos para formar un grupo control intermedio aparecen en las páginas 130 – 132.

La estrategia de turnos. A veces se puede dar el programa X primero a un grupo y luego a otro grupo. Por ejemplo, una máquina de lectura de alto costo puede estar a disposición de seis cursos seleccionados al azar, durante la primera mitad del año, y luego a disposición de los seis cursos restantes durante la segunda mitad del año. Debido a que los cursos podrían ser asignados al azar a su "turno", se logra un grupo control verdadero.

La estrategia de programa diferido. Aunque eventualmente a todos los alumnos se les dará el programa X, se puede obtener una evaluación poderosa de por lo menos parte de él, si se difiere la entrada de un grupo de alumnos seleccionados al azar. Es, a menudo, justificado y aconsejable, partir en pequeña escala cuando se implementa un programa nuevo.

La estrategia de evaluación diferida. Si el programa X ya está funcionando de tal modo que es demasiado tarde para formar un grupo control equivalente, conduzca una evaluación formativa esta vez y prepare las bases para realizar una evaluación sumativa la próxima vez que ensaye el programa.

El grupo control no equivalente (no elegido al azar).

Este es un grupo que se *selecciona* por ser semejante al grupo experimental. No se *forma* por asignación al azar. Usted debería siempre tratar de encontrar un grupo que sea lo más similar posible al grupo experimental y usar ese grupo para comparación. A continuación se dan algunos ejemplos de dichos grupos "no equivalentes".

Ejemplo 1. En un colegio de niños que están terminando su enseñanza básica se comenzó con un programa de matemáticas, acelerado, para un grupo de sus veinte alumnos más capacitados. Con muchos deseos de controlar los efectos de regresión (discutidos en página 21 de este libro) y de responder preguntas tales como:
"¿Progresaron los alumnos más de lo corriente?", el director pidió a un colegio vecino que cooperara permitiendo que se realizaran algunas mediciones. Conseguido el permiso, él hizo una prueba a los alumnos del colegio vecino y

seleccionó los veinte mejores *por medio del mismo procedimiento que había usado en su propio colegio.* Este grupo entonces llegó a ser el grupo control no-equivalente. Este grupo bien podría haber sido diferente, en diversos aspectos, puesto que provenía de sectores residenciales ligeramente diferentes y había asistido a un colegio diferente. Sin embargo, una comprobación de los puntajes en el pretest de matemática no mostró diferencias significativas entre los grupos en esa medición, convirtiéndolo así en un grupo bastante deseable como grupo control. Los puntajes similares en matemáticas no significaron que los grupos eran similares en todas las mediciones. Además, hubo muchas diferencias entre lo que sucedió a los grupos aparte de su programa diferente en matemáticas, es decir, "confusión" fue un problema. Sin embargo, dicho grupo control no equivalente *bien valía la pena usarlo*.

Los resultados de los pre y postests un una prueba estandarizada de matemática mostraron que los dos grupos habían progresado bastante en un año, y que no existían diferencias significativas entre las medias aritméticas. Sin embargo, el grupo experimental, además, fue capaz de escribir programas de computación bastante complicados. De este modo, estos últimos no sólo habían "continuado" con sus compañeros de experimentación el aprendizaje en el programa de matemática, sino que además sacaron provecho del programa especial. Si la evaluación hubiera mostrado solamente que los alumnos acelerados habían aprendido algo sobre programación de computador, se habría podido decir que el aprendizaje se había producido a expensas del rendimiento en matemáticas más rutinarias. La comparación con el grupo control disipó este desafío. *El grupo control no-equivalente bien valía la pena usarlo.*

Ejemplo 2. Un distrito escolar implementó una política de reubicar a los niños educacionalmente impedidos en cursos regulares, agrupados en forma heterogénea, en la escuela básica. Una meta de este programa fue "mayores actitudes positivas hacia el colegio por parte de los alumnos educacionalmente impedidos". Los administradores del distrito estaban ansiosos por saber que efecto, en realidad, tuvo este cambio de política sobre estas actitudes y pidieron una evaluación formal. El evaluador consideró los distritos escolares de los alrededores y pidió a dos de ellos que cooperaran en intercambiar información sobre alumnos EI (Educacionalmente Impedidos). Un distrito era adyacente, pero de status socioeconómico ligeramente más alto. El otro distrito se seleccionó porque, aunque mucho más lejos, los niveles de entradas y avalúo de la propiedad eran muy parecidos a aquellos de su propio distrito. Al término del año un psicólogo que no estaba informado del propósito del estudio ni de la diferencia en programas entrevistó a una muestra de alumnos EI seleccionados al azar en cada distrito. Este psicólogo calificó a los alumnos por su actitud hacia el colegio, auto concepto, etc. Sus resultados mostraron que sólo el 49% de los entrevistados en el distrito que introdujo innovaciones tenía una actitud fuertemente positiva hacia el colegio. Este resultado pudiera haber parecido desalentador; pero los resultados de los otros dos distritos (los grupos control no equivalentes) mostraron que sólo el 15% y 20% tenían actitudes positivas hacia el colegio. El uso de grupos control no-equivalentes,

otra vez, valía la pena: ellos mostraron lo que hubieran sido los resultados sin el programa.

Como usted puede ver con estos ejemplos, la reunión de información de un grupo control no-equivalente puede ser enormemente útil al evaluar el efecto de un programa y por lo tanto, es valioso intentarlo en situaciones en que un grupo control verdadero no puede formarse.

Observe estos tres puntos en relación a grupos control no-equivalentes:

Selección del grupo-control no-equivalente

1. *Si el grupo experimental se selecciona por medio de algún* procedimiento (por ejemplo, pretest de matemáticas, en el primer ejemplo, y ubicación de niños EI en el segundo ejemplo), *entonces el grupo control debería seleccionarse por medio de un procedimiento que sea lo más similar posible.*

Administración de pruebas a los dos grupos

2. *Al grupo control no-equivalente se le deberían administrar todos los tests principales que se la administran al grupo experimental.* De esta manera ellos también se acostumbran al estilo y contenido de los tests. La administración misma de la prueba puede producir una diferencia en el rendimiento de los alumnos, por esto, *a menos que los tests sean una parte integral del programa,* los grupos experimental y de control deberían recibir los mismos tests a través del programa. Esto es importante si además del pretest que se acostumbra aplicar al comienzo del programa y del postest que se administra al final, se van a administrar otros tests, con frecuencia, durante el programa.

   ¿Cómo puede la administración de los tests afectar el rendimiento? Los tests:

   a) centran la atención del alumno en partes importantes del currículo,
   b) dan a los alumnos práctica adecuada, preparándolos de este modo para que rindan en el test final,
   c) entregan información al alumno sobre cuán bien lo está haciendo, si los resultados están disponibles, motivando así a los alumnos,

d) indican al profesor las áreas débiles en la instrucción de tal forma que pueda corregirlas.

Por supuesto que tener el mismo programa de administración de tests para los dos grupos es igualmente importante cuando hay un grupo control *verdadero,* pero cuando el grupo está en otro colegio, como frecuentemente sucede con grupos control no equivalentes, este principio de igual administración de pruebas se viola muy a menudo.

Documentación de semejanzas y diferencias

3. Esté preparado para documentar semejanzas y diferencias entre los grupos control y experimentales. La credibilidad de sus conclusiones dependerá directamente de su habilidad para demostrar que los grupos control y experimental han sido lo más parecido posible, a excepción de la diferencia en los programas que recibieron.

   Piense en todas las formas en que un escéptico pudiera lanzar dudas sobre lo apropiado que es el grupo no-equivalente. Por ejemplo, podría decir, "el tamaño de los cursos era diferente, "las horas dedicadas a la asignatura eran diferentes", "el jardín infantil al que asistieron esos niños enseñaba el sistema fónico", "el grupo-E era de estatus socioeconómico más alto", etc. Estudie cualesquier problema serio, reuniendo información relevante y examinándola.

¿Qué programa debería recibir el grupo control?

Algo está, obviamente, sucediendo al grupo control (al grupo-C) cada día, durante el tiempo en que el grupo-E recibe el programa X. ¿Qué será ese algo? En esta sección se entrega una lista de las posibilidades, considerando la mejor solución primero y bajando hasta la más débil, aunque incluso la solución "más débil", esté seguro, es muchísimo mejor que no tener absolutamente ningún grupo control. El análisis que sigue sirve para ambos grupos control, el verdadero y el no-equivalente.

La mejor solución. Recordando que las evaluaciones se realizan con el propósito de entregar información para la toma de decisiones, queda claro que la mejor solución es la que entrega la información más útil para cualesquiera que sean las decisiones que tengan que tomarse. Si la evaluación del programa X se realiza con el fin de elegir entre el programa X y cualesquier otro programa, idealmente es este otro programa el "competidor más cercano", y es, por lo tanto, el que debería tomar el grupo control.

Ejemplo. Un comité de un distrito había estado examinando materiales sobre educación sexual, con la idea de adoptar un conjunto de materiales para usarlos en un programa nuevo de educación sexual. Ellos habían eliminado varios materiales llegando a preseleccionar dos conjuntos. Algunos miembros del comité pensaban que un conjunto de materiales, aunque bueno, era demasiado explícito y produciría reacciones negativas de los padres. Otros miembros del comité pensaban que los materiales explícitos eran los mejores, que no habría objeciones a ellos y que los materiales alternativos eran demasiado evasivos. Se decidió que ambos conjuntos de materiales deberían ensayarse y que se deberían evaluar las reacciones de los alumnos y de los padres. Con el objeto de probar el conjunto de materiales explícitos y el de materiales evasivos, se seleccionaron al azar cursos de quinto y sexto grado básico. Al final del ensayo se enviaron cuestionarios a los padres y también se aplicaron cuestionarios a los alumnos. Basándose en esta información, se evaluó el impacto de cada programa y luego se pudo hacer una elección informada.

Cuando quiera que se vaya a ensayar un programa nuevo es bueno ensayar no tan sólo *un* programa nuevo, sino dos o más. De esta manera usted puede obtener una buena evidencia con respecto a cual de los programas existentes es mejor para su colegio. También, teniendo dos programas y siendo ambos nuevos, se hace más fácil adoptar una selección al azar que cuando usted tiene sólo un programa nuevo para compararlo con un programa antiguo. Todos tenemos una tendencia a querer estar en el programa *nuevo*, sólo porque es nuevo. Este deseo hace que sea difícil persuadir a las personas para que tomen parte en un grupo control.

Una alternativa importante a la de comparar dos programas nuevos, es dar al grupo control una *versión más barata* de un programa nuevo. ¿Puede el programa X implementarse dejando fuera algunas partes que no son indispensables? Si es así, esto podría ser exactamente lo que el grupo control recibiera. Esto debería ser exactamente lo que el grupo control recibiera. Esto debería ayudar a las personas a decidir sobre la forma como debería implementarse el programa, en forma cara o barata. Después de todo, si usted obtiene los mismos resultados con una versión barata ¿por qué usar la versión cara? Sin embargo, puede ser que usted quisiera *estar seguro* de que el programa nuevo es mejor que el antiguo, antes que suponer que es así e implementar dos programas nuevos. En este caso, usted debería tener más de un grupo control. Quizás usted podría comparar el programa X con sus dos competidores más cercanos *y* con una versión más barata del mismo.

La segunda mejor solución. Si usted no puede implementar lo que considera que es un verdadero programa alternativo para el grupo control, entonces, por lo menos, sería bueno que el grupo control recibiera un programa que tuviera metas y objetivos similares a los del programa X. Este programa control podría ser el programa antiguo, aunque usted no espere continuar con dicho programa. Quizás, por ejemplo, usted piensa solamente continuar buscando un programa *mejor*. Aún vale la pena mantener el programa antiguo funcionando un año o dos mientras ensaya algunos nuevos. ¿De

qué otra manera puede usted saber si el programa antiguo fue o no fue, en verdad, *mejor* que cualesquiera de los programas nuevos que usted prueba?

En otras situaciones el programa control podría ser un programa que usted nunca ha ensayado.

Ejemplo. Usted va a usar un grupo de alumnos de un colegio secundario vecino como grupo control no-equivalente para el programa X, un programa nuevo en matemática que enfatiza las destrezas de solución de problemas. Usted averigua con los profesores de matemática del colegio vecino y descubre que ellos también están apuntando a habi-lidades matemáticas de solución de problemas en sus programas. Sus alumnos forman un grupo control no-equivalente adecuado, aunque usted no esté intentando implementar el programa que ellos están recibiendo.

Una solución débil. A veces usted podrá solamente comparar el grupo experimental que ha recibido un programa especial, programa X, con un grupo control que no recibió ningún programa. Esto pudiera ser todo lo que es posible hacer aun cuando "ningún programa" no es una forma de acción considerada. ¿Vale la pena hacer una comparación entre el programa X y ningún programa? *Es*una prueba del programa, aunque débil, es decir, una prueba en la que el programa es poco probable que falle. Le permite decidir si el programa es mejor que no tener programa alguno.

Ejemplo. A un grupo de alumnos seleccionados al azar (el grupo-E) se le dieron materiales de matemática programados, para que trabajaran con ellos en casa. Al cabo de seis semanas, a estos alumnos se les hizo una prueba sobre sus conocimientos del material que aparece en los libros programados. Junto con ellos, a un número igual de otros alumnos (al grupo-C) también se les dio una prueba. No se esperaba que los alumnos del grupo control rindieran tan bien como el grupo-E, puesto que ellos no habían tenido programa especial. Sin embargo, sus puntajes proporcionaron una *línea base* a partir de la cual se podía juzgar los puntajes del grupo-E. Si los puntajes del grupo-E fueran sólo ligeramente más altos que los puntajes del grupo-C, esto no indicaría que enviar materiales programados a la casa fuera un programa muy efectivo.

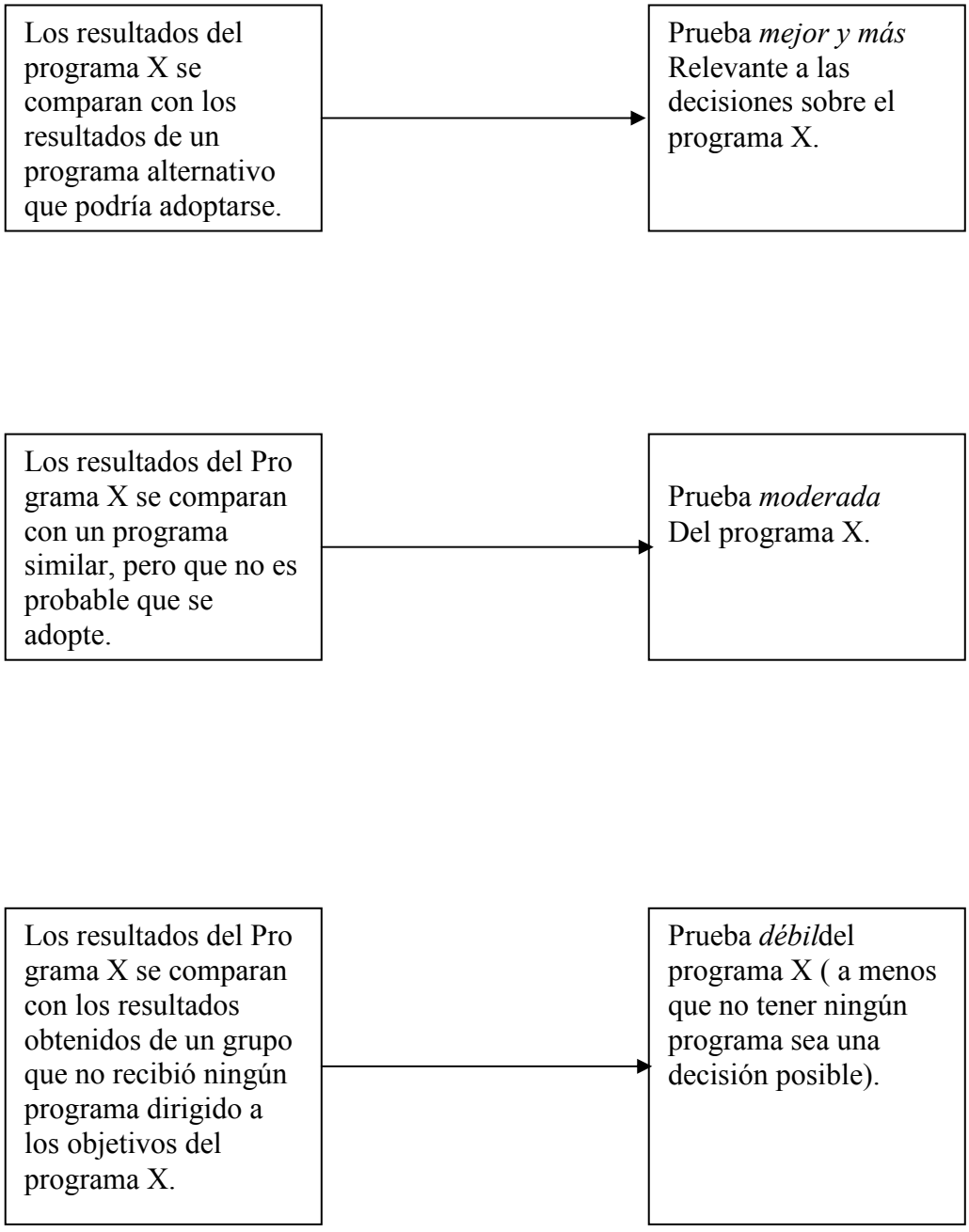
¿Vale la pena hacer una comparación con un grupo que *no* recibe programa? El ejemplo anterior ilustra como el grupo sin programa proporciona una línea base a partir de la cual se pueden juzgar los resultados del programa. Una razón importante de por qué vale la pena hacer esa comparación es que la comparación "controla" los efectos de *maduración*. Es decir, toma en cuenta el simple hecho que los alumnos llegan a ser mayores y más capaces y recogen información simplemente con el pasar

del tiempo. Al final de muchos programas, especialmente de programas de un año de duración, usted no sabe si los alumnos están obteniendo puntajes más altos de los que obtuvieron al comienzo, debido al programa o porque son mayores. El crecimiento rápido durante los años de la escuela básica hace que este problema sea muy general al evaluar programas de este nivel.

El escéptico podría argumentar que los niños en el programa están rindiendo más simplemente porque ellos han madurado y por lo tanto habrían aprendido de todos modos. Ellos habrían logrado lo mismo sin el programa. Midiendo otro grupo que no recibió el programa, usted tiene una idea de cuánto del aumento en puntajes se puede atribuir a este crecimiento de los alumnos. Un buen programa debería producir más aumento que simplemente el aumento natural debido al crecimiento, es decir, debido a madurez.

Un comentario más queda por hacer con respecto al programa que recibe el grupo control. Se refiere a la interpretación de comparaciones hechas entre los resultados del grupo-E y los resultados del grupo-C. Si ambos grupos reciben programas que tienen *exactamente las mismas* metas y objetivos, cada programa está sujeto a un test muy estricto: ¿se compara favorablemente contra un competidor verdadero? Sí, sin embargo, el grupo-C recibe un programa que tiene un énfasis algo *diferente* del programa que el grupo-E recibe (por ejemplo, el programa de matemática C podría no gastar tanto tiempo en habilidades básicas de computación como el programa X, prefiriendo introducir conceptos más abstractos) entonces la interpretación de resultados diferentes es más difícil. Si el programa X recibe puntajes sobre el contenido que él mismo enfatiza, los que *no son más altos* que los del programa control, el programa X no parece estar haciéndolo muy bien –no puede ganar en su propio terreno, por así decirlo. Es una buena idea hacer esta comparación; la actuación del otro programa proporciona un tipo de límite inferior para resultados aceptables. Esto es suficiente si usted se está centrando en la evaluación del programa X. Sin embargo, si usted quiere analizar el valor de ambos programas se requerirán más esfuerzos. Dos soluciones posibles son; (a) proporcionar tres conjuntos de resultados, un "conjunto" para las metas principales y claramente establecidas de cada programa y un "conjunto" para metas que tienen en común los dos programas o (b) pedir a un experto independiente que seleccione los ítems más importantes del test y usar esta selección como la medida de resultado básica para ambos programas.

Para resumir las ideas con respecto al programa que el grupo control debería recibir: El programa X puede estar sujeto a un test fuerte o débil, dependiendo del tipo de programa control con el que se compara. Mientras más cercano está el programa control a la o las alternativa(s) que se consideran, más útil es la evaluación para decisiones futuras.



Se dijo a comienzos de este capítulo que un diseño es un plan que establece a quién se mide y cuándo se mide. Hasta ahora hemos considerado la primera parte, a *quién* medir, "grupos", la sección siguiente trata sobre el otro componente de los diseños: *cuándo* se hacen las mediciones.

MOMENTOS EN QUE SE HACEN LAS MEDICIONES

Postests

Los tests que se administran antes que comience un programa se llaman *pretests*, una horma breve para decir tests previos al programa; así mismo, los tests que se administran al final de un programa se llaman *postests*. Los tests también se pueden administrar varias veces: antes, durante o después de un programa. A continuación se analizará cada una de estas posibilidades.

Los postests son mediciones que se hacen al final de (o a veces durante) un programa y es en los puntajes del postest donde deberían aparecer los resultados del programa. En términos de análisis de datos el postest es la "variable dependiente", los resultados del postest *dependen*, en parte al menos de lo que sucedió en el programa. Debido a que los postests miden los resultados de un programa, los diseños nunca omiten totalmente los postests, aunque ocasionalmente sólo se administren a parte de un grupo. Hay una pregunta práctica, sin embargo, en el sentido de saber exactamente *cuando* debería administrarse un postest. A veces la fecha de un postest está determinada por la fecha en que debe entregarse un informe sobre el programa. En este caso, asegúrese de fijar la fecha para el postest de tal manera que se deje dos veces el tiempo que piensa que necesitará para preparar el informe. Sin embargo, si el informe no debe entregarse hasta bastante tiempo después que termine el programa, usted puede ser más flexible en la elección de la fecha para administrar el postest.

En un programa que abarca un año escolar, puede ser que usted quiera darle bastante tiempo al programa para que funcione y obtener así los resultados máximos, pero también considere la inquietud que comienza a crearse antes de las vacaciones de verano. Esta inquietud de fin de año podría significar que usted puede obtener una medición mejor del programa si la realiza un mes antes de que termine el año escolar, en lugar de esperar hasta el último momento. Su juicio y el de su equipo de trabajo serán su mejor guía para determinar exactamente cuándo calendarizar el postest. Otro aspecto que debe considerarse al calendarizar es la necesidad de "pruebas de recuperación. Los alumnos ausentes el día del postest deberían tomarlo inmediatamente después que regresan al colegio, a *menos* que usted y su equipo de trabajo sientan que en ese momento el contenido del test ha llegado a ser de conocimiento general, y esto pudiera afectar en forma injusta los resultados en un "test de recuperación". En tal caso los alumnos que no se someten al postest, simplemente no deberían contarse en el análisis de los datos. Estas omisiones deberían informarse como "mortalidad" del postest. Si se *va* a dar "test de recuperación", el postest debe calendarizarse con la suficiente anticipación como para permitir que haya oportunidades posteriores para tests de recuperación.

Pretests

Cualesquier puntaje de test o medición que se reunió antes de que los alumnos recibieran el programa puede llamarse pretest. A continuación se entrega una lista de las razones principales por las que se puede querer usar algún tipo de pretest:

1. seleccionar alumnos para el programa

2. verificar supuestos que se han hecho al planificar el programa
3. investigar o asegurar el grado de comparación de los grupos
4. verificar los progresos alcanzados durante un programa
5. aumentar la sensibilidad de la prueba de un programa.

1. Seleccionar alumnos para el programa. A veces los alumnos se seleccionan para un programa, o se consideran elegibles para un programa, en base a los puntajes que han obtenido en un test.

Ejemplo. Los alumnos serían seleccionados para el programa de recuperación en lectura si durante más de un año sus puntajes en el test de lectura estuvieran bajo el nivel del curso.

En el ejemplo anterior, el test de lectura fue la "medida de selección". Tener un puntaje más bajo del nivel real del curso durante un año, fue el criterio para la selección.

Es tentador usar el test de selección como un pretest en análisis subsiguientes de los resultados del programa; pero tenga cuidado, ya que si un grupo de alumnos se seleccionó debido a puntajes extremos (puntajes altos o bajos) en una prueba, la media aritmética de los puntajes de ese grupo "regresará hacia la media" en un test posterior. Esto significa que la media aritmética del grupo de alumnos que se seleccionaron porque fueron *bajos*, será más *alta* en un "re-test", incluso sin que intervenga enseñanza o aprendizaje. La media aritmética "mejoraría" de esta manera porque algunos alumnos, especialmente aquellos cercanos al punto límite de separación para la selección, estaban en el grupo seleccionado debido a mala suerte: sus adivinanzas eran equivocadas, mientras otras eran correctas, o ellos se sentían enfermos ese día, o mal entendieron una serie de instrucciones. En un "re-test" estos alumnos probablemente no tendrán la suerte en su contra otra vez y algunos de sus puntajes subirán, incluso en algunos casos puede haber saltos considerables. En general la media aritmética del grupo mejorará. Cuando usted selecciona el extremo inferior de un grupo, ellos no tienen donde ir sino hacia arriba. Los especialistas en estadística lo llaman "regresión" o "regresión hacia la media aritmética".

En el caso de alumnos seleccionados sobre la base de puntajes *altos* si se les diera el test otra vez su media aritmética probablemente sería *inferior*. Otra vez, hay regresión hacia la media aritmética. En este caso los alumnos seleccionados por altos puntajes incluían alumnos que obtuvieron dichos puntajes altos debido a errores en su favor, tales como adivinanzas que resultaron correctas. Al no tener dicha suerte en un "re-test", sus puntajes bajarán. Queda claro que los efectos de regresión pueden afectar la evaluación de un programa. Así como los programas de recuperación pueden aparecer como falsamente exitosos, también los programas "avanzados" o "acelerados" pueden aparecer como falsamente fracasados, simplemente debido a efectos de regresión. Por esta sola razón es muy importante tener un grupo control de algún tipo cuando se

evalúan programas de recuperación o acelerados, o cualesquier programa para el cual se seleccionaron alumnos considerando los puntajes extremos.

Recomendación Cuando se usa una prueba con el fin de seleccionar alumnos para un programa sobre la base de puntajes altos o bajos, es mejor administrar un segundo test como pretest. El segundo test no estará sujeto a los mismos efectos de regresión, y los puntajes en él probablemente estarán normalmente distribuidos de tal modo que los tests estadísticos serán adecuados.

2. Verificar supuestos que se han hecho al planificar el programa. Un pretest es a veces deseable como una verificación de la adecuada implementación de un programa. Una pregunta al implementar un programa es, "¿son estos los verdaderos participantes que especificó el plan del programa? Un pretest puede usarse para verificar casos tales como el rendimiento del pre-programa, o actitudes de los alumnos, profesores, etc. que van a estar involucrados.

Ejemplo l. Se podría haber planificado un programa bilingüe suponiendo que la mayoría de los alumnos de apellido español hablaban español en su casa. Un test podría indicar que esto no es así y luego señalar algunas modificaciones necesarias en el programa.

Ejemplo 2. Se planificó un programa de matemática para la escuela básica, suponiendo que los niños que sabían las tablas de multiplicar podían también sumar números con facilidad. Pero un pretest, analizado por objetivos, mostró que algunos niños que hicieron bien los ítemes relacionados con multiplicación eran, no obstante, aún muy deficientes en adición. El orden jerárquico de instrucción que se había planificado tenía que modificarse. Algunos niños aparentemente no aprenden cosas en la secuencia esperada.

---

3. Investigar, o asegurar el grado de comparación de los grupos. Si los puntajes de dos grupos en el postest se van a comparar, la primera pregunta que salta a la mente es, "¿eran los grupos semejantes *al comenzar*, es decir, antes que un grupo recibiera el programa X y otro recibiera el programa C?"

Si usted está usando un grupo control verdadero, es decir, si los dos grupos se formaron por aleatorización, no habría necesidad de dicha verificación puesto que el número de alumnos involucrados es más bien grande (digamos de 25 alumnos o más por grupo si los alumnos fueran bastante heterogéneos, con un gran rango de habilidad u otra característica relevante: 15 por grupo si los alumnos fueran un rango más bien homogéneo al comenzar). La aleatorización debería haber funcionado para hacer los grupos razonablemente equivalentes. Sin embargo, si el número de alumnos para asignar fuera pequeño, la aleatorización podría tener menos éxito en igualar los grupos. En este caso, uno podría usar un pretest para ver si la aleatorización había producido grupos que no mostraran diferencias apreciables en el pretest. Aun es mejor, sin embargo, la práctica de *usar un pretest para ayudar a la aleatorización.* Se da un pretest a todos los alumnos, se forman bloques de alumnos con altos, medianos y bajos puntajes, y alumnos de cada bloque se asignan, entonces, al azar al programa

experimental o de control. (Ver la sección sobre "cómo formar bloques", en el Apéndice). Si usted está usando un grupo control no equivalente (no al azar, es *esencial* un pretest para ver si el grupo control es comparable inicialmente al grupo experimental.

Recomendación En el caso de un grupo control verdadero, la asignación al azar reduce la necesidad de un pretest, pero use uno para asegurar grupos equivalentes si el número involucrado es pequeño (unos 15 por grupo o menos) o si la variabilidad es muy grande (por ejemplo, rango de C.1. desde 70 a 120). Siempre use un pretest si el grupo control es no-equivalente (no al azar). El pretest debería ser una forma equivalente del postest o, si esto no es posible, debería medir una habilidad conocida, para relacionarla con la medición del postest.

4. Verificar los progresos alcanzados durante un programa. Usted no sabe cuán lejos ha llegado si no sabe donde estaba al comienzo. Del mismo modo, sin tener una medida de cómo eran las cosas antes que comenzara un programa, usted no podrá, cuando el programa termine, señalar las ganancias que se han logrado. Si se necesita evidencia sobre qué es lo que exactamente se ha aprendido en el programa, entonces ésta es una buena razón para usar un pretest. Dicha evidencia es susceptible de ser obtenida, sin embargo, solamente cuando se usan tests referidos a criterio. Cuando el contenido real de un test no va a ser considerado, como por ejemplo cuando se usan los puntajes de los tests estandarizados, no es necesariamente esencial un pretest y el uso de puntajes de progreso debería evitarse. La práctica, desafortunadamente común, de usar en evaluación puntajes de progreso en tests referidos a normas, le dice muy poco sobre lo que los alumnos aprendieron.

Ejemplo. Se adoptó un programa de destrezas básicas de escritura para los alumnos de octavo grado del año anterior. Se identificaron las destrezas básicas y se agruparon como objetivos. A comienzos de año, veinte alumnos fueron sometidos a un test que cubría 20 objetivos. Siguió después la instrucción en pequeños grupos compuestos por alumnos que aún no habían dominado los objetivos identificados para el grupo. Al final de año, todos los alumnos fueron sometidos nuevamente al test que cubría los 20 objetivos. Se informó a los padres sobre las ganancias de cada alumno, pues ellos habían indicado que deseaban tener esa información.

En caso de que usted esté usando un grupo control verdadero, generalmente se puede suponer que los puntajes del pretest han sido equivalentes debido a la aleatorización. La comparación que debe hacerse es entre los puntajes de postest de los grupos E y C. En tal caso, sólo se necesitan los puntajes del postest para responder a la pregunta: ¿Dio el programa como resultado mayores logros que otro?
Si usted está usando un grupo control no-equivalente, será necesario un pretest. En el caso de que no haya grupo control es esencial, un pretest, referido a criterio (basado en objetivos) de tal manera que los logros puedan documentarse.

5. Aumentar la sensibilidad de la prueba de un programa. Si usted sospecha que el programa que está evaluando podría no producir una diferencia muy dramática en resultados, pero quiere estar seguro que podrá mostrar que existe, entonces lo que

usted necesita es un test que sea sensible al programa. Usted necesita un diseño "poderoso". Tener un pretest que esté muy estrechamente relacionado al postest agrega poder a su diseño. Puesto que el poder de un diseño es un concepto importante, se discute en la sección siguiente.

El Poder de un Diseño.

"Poder" se usa aquí como término técnico. La palabra probablemente, se eligió para estadística teniendo en mente la analogía de un lente de aumento. Mientras más poderoso es el lente más pequeñas son las cosas que usted puede ver con él. Es algo parecido lo que sucede con los diseños: mientras más poderoso es un diseño, más pequeñas son las diferencias que usted puede detectar entre dos conjuntos de resultados. Con un lente de aumento poderoso usted puede detectar una diferencia entre dos hormigas, lo que no podría detectar con un lente menos poderoso. Con un diseño poderoso usted puede detectar una diferencia entre dos programas, la que podría haber parecido insignificante con un diseño menos poderoso.

Un diseño tiene más poder para detectar diferencias sutiles cuando permite al evaluador explicar más sobre la variación en resultados, variación causada por diferencias de alumno, errores, etc. Cuanto más se "explica", más claramente podemos ver lo que se deja atrás. Las diferencias causadas por los diferentes programas pueden así ser sensitivamente ubicadas.

Ejemplo. Cuarenta niños que asistían regularmente a un grupo de juego después de clase, solicitaron ser aceptados en un campamento para el cual había solamente 20 plazas. El director del campamento seleccionó al azar veinte alumnos para que fueran al campamento por dos semanas. Al final de las dos semanas él pensó que los niños parecían haber perdido peso. Cuando todos regresaron al programa que se ofrecía después de clase, él pesó a los que fueron al campamento y a los que no fueron. Casi no hubo diferencia entre las medias aritméticas, pero él pensó que esto se debía al hecho de que algunos niños muy pesados habían ido al campamento. Él se dijo, "Si yo tuviera una medida exacta del peso de cada niño antes de ir al campamento, entonces yo podría saber si el hecho de asistir al campamento produjo o no diferencia, incluso una pequeña diferencia en los pesos de los niños" (Él había preguntado a los niños cuánto pesaban antes de ir al campamento; pero no podía confiar en los resultados de la memoria y de las balanzas de baño, de exactitud dudosa, y en todo caso muchos niños no tenían idea). El director del campamento tenía mucha razón. El podía haber tenido una prueba más poderosa de sus hipótesis si hubiera podido saber más, si hubiera podido explicar más las variaciones en pesos finales como debida a diferencias iniciales en peso. Entonces él podría haber visto si los niños que fueron a campamento habían en verdad perdido peso. El también tenía razón al considerar que una medida con una serie de errores en ella –como los resultados de preguntar a los niños cuál era el peso que creían tener antes del campamento- no aumentaría mucho su poder.

En cuanto a los pretest se refiere, la implicación es que el mayor poder puede obtenerse teniendo un pretest lo más parecido posible al postest. El mejor predictor de cualesquier tipo de conducta futura de una persona es su conducta anterior en circunstancias similares. Aplicando este principio muy sensible de psicología a la medición mediante tests, la mejor predicción de lo que un niño capaz de hacer en el futuro en un cierto tipo de test es una medida de cómo puede él desempeñarse en ese test ahora. Por lo tanto, mientras más exactamente podamos predecir los puntajes en un postest a partir de los puntajes de un pretest, más exactamente podremos detectar desviaciones debidas posiblemente a los efectos del o de los programas que se evalúan. De este modo, usar un pretest que es una forma equivalente (o exactamente igual) del postest (por ejemplo, ambos test miden las mismas destrezas matemáticas o ambos tests miden ortografía de palabras de igual frecuencia o dificultad) da la información más precisa sobre la efectividad de un programa comparado con otro. Aumenta el poder del diseño de evaluación. Puede hacer un diseño de un grupo control verdadero virtualmente indiscutible, y puede aumentar la credibilidad de, incluso, un diseño que no incluye siquiera un grupo control.

En esta sección se han analizado cinco razones de por qué usted pudiera querer usar un pretest. Como usted habrá notado, el uso de un pretest puede ser innecesario cuando un grupo control verdadero es parte del diseño. A continuación hay algunas razones de por qué pudiera ser que usted no quisiera usar un pretest.

Algunas razones que indican por qué pudiera ser que usted no quisiera usar un pretest. Primero, pudiera ser que usted no quisiera usar un pretest debido a que tomar un pretest probablemente alteraría a los alumnos en alguna forma no medible. Este podría ser el caso si el postest va a ser una medida de actitudes y usted quiere evaluar las mismas actitudes antes del programa. Hay evidencia de que el hecho de formular las propias actitudes (como uno a veces está forzado a hacerlo, por ejemplo, cuando contesta un cuestionario) puede afectar aquellas actitudes. Puede "ubicarlas" o hacer que el alumno esté más vigilante sobre ellas, haciendo menos probable cambios posteriores. Además, el pretest de actitudes podría alertar a los alumnos hacia los objetivos de un programa y producir subjetividad en sus respuestas al programa.

A pesar de estas razones para NO medir actitudes con anterioridad al programa, a uno siempre le gustaría saber cuáles eran las actitudes existentes y aquí hay una solución muy efectiva si usted tiene números bastante grandes (al menos 30 personas que respondan en cada grupo). Usted puede medir las actitudes pre-programa que tenía una mitad de los alumnos, seleccionados al azar que reciben el programa. Las actitudes post-programa de la mitad de los alumnos que no fueron sometidos al pretest no habrá sido influenciada por ninguna tendencia inducida por el pretest. Aquí hay un ejemplo de este procedimiento.

Ejemplo. Los profesores de matemática esperaban que un programa de matemática nuevo que estaba siendo implementado en todo un colegio secundario mejorara la afición de los alumnos por las matemáticas, así como también mejorara sus habilidades. Antes que comenzara el programa, ellos prepararon dos tipos de cuestionarios. Uno preguntaba sobre matemáticas (¿Te gustan? ¿Son entretenidas?) y el otro preguntaba sobre lectura. Los

cuestionarios se mezclaron en forma equilibrada en un montón y luego fueron entregados al azar a los alumnos de tal modo que más o menos la mitad de los alumnos respondió el cuestionario de matemática y la otra mitad respondió el cuestionario de lectura. Hacia el término del programa se prepararon los mismos cuestionarios para ser usados nuevamente. Los nombres de quiénes habían contestado los cuestionarios de lectura se anotaron en cuestionarios de matemática en blanco. Los otros nombres se pusieron en los cuestionarios de lectura. De esta manera, cuando se entregaron los cuestionarios con nombres a los alumnos a quienes anteriormente se les había interrogado sobre matemática ahora se les interrogaba sobre lectura y viceversa. Este procedimiento proporcionó una medida de pretest y postest, tanto para matemática como para lectura. Los resultados de matemáticas se analizaron para verificar el impacto del nuevo programa de matemática (los resultados de lectura proporcionaron incidentalmente una medida de tendencia general en otra área de contenido, información que podría ayudar en la interpretación de resultados).

Otra razón de porqué usted pudiera omitir un pretest es que en algunos casos usar un pretest podía no tener sentido. Por ejemplo, si usted estuviera listo para introducir un nuevo programa de enseñanza en francés, no habría razón de dar un pretest de francés a un grupo de niños de habla inglesa. Sus puntajes presumiblemente se deberían simplemente a la adivinanza; el pretest no proporcionaría una información significativa. En cambio, podría usarse como pretest un test de aptitud lingüística, o algún otro test que se espera que correlacione con las medidas de resultados.

Tercero, puede ser que usted no desee usar un pretest debido a que el programa ya se está desarrollando y no se dio un pretest. ¿Cree usted que simplemente porque no se dio un pretest tiene que aceptar esta situación como inmutable? Puesto que se supone que las medidas de habilidad no se afectan mucho por los programas escolares, usted podría usar una medida de habilidad como un pretest "retrospectivo". Aunque se administrara una medida de habilidad al grupo E y C en medio del programa, usted podría suponer que esta medida representa perfectamente bien la posición relativa de los dos grupos tal como fue antes que el programa comenzara. Este, por supuesto, sería más probable que fuera válido si la medida de habilidad fuera un test de habilidad ya sea "no-verbal" o no influenciado culturalmente, más bien que un test de habilidad convencional que contenga cantidades considerables de lectura y de aritmética, habilidades que el programa podría haber influenciado.

Finalmente, puede ser que usted no quiera usar un pretest debido al costo en tiempo o dinero. En general, sin embargo, las ventajas del pretest casi siempre sobrepasan las desventajas.

Resumen de las observaciones sobre pretests. La sección siguiente resume las ideas que se han presentado con respecto a los pretests. Éste se hace presentando la misma información en un formato diferente. El lector habrá notado las siguientes posibilidades para los pretests:

- Un pretest puede ser un test de actitudes
- Un pretest puede ser un test de rendimiento
- Un pretest puede ser un test de habilidad.

Un pretest que es un test de actitud. Éste sería muy adecuado si el programa estuviera apuntando a cambios de actitud. Debido a la posibilidad de que el pretest pueda en sí mismo influenciar actitudes o afectar la forma en que los alumnos responden al programa y/o al postest posterior, se recomienda a veces que, dentro de cada grupo, sólo a la mitad elegida al azar se le administre el test de actitud.

Un pretest que es un test de rendimiento. Si el postest va a ser un test de rendimiento, entonces el uso de un pretest que es el mismo o similar al postest da una información valiosa. Las ganancias pueden identificarse en varias partes del contenido cubierto por los tests. Además, el pretest de rendimiento es una medida muy relevante para considerar si dos grupos son o no son "equivalentes". Una verificación de la equivalencia de los grupos será especialmente importante, si existe una de las siguientes situaciones:

- hay un grupo control no equivalente (administre siempre un pretest si el grupo control no se formó por medio de asignación al azar)

- número pequeño, digamos memos de 15 por grupo

- gran variabilidad en la "población" que se muestrea (por ejemplo, si los grupos que se seleccionaron pudieran contener un gran rango de habilidad).

El pretest que es un test de habilidad. Cuando el postest va a ser una medida de rendimiento, se usa menudo una medida de habilidad como pretest. Esto permite que los puntajes del postest se relacionen a los niveles de habilidad de los alumnos. Dicha comparación es a menudo un camino excelente para obtener un juicio de la significación educacional de las diferencias en los puntajes del postest. El efecto del programa puede compararse con los efectos de diferentes niveles de habilidad.

Ejemplo. Al final de un experimento para evaluar la efectividad da la tutoría de los padres en las tablas de multiplicar, se descubrió que el grupo experimental obtuvo diez puntos más, en promedio, que el grupo control. Nadie estaba muy seguro sobre cuán bueno era este resultado., hasta que un profesor señaló que, en el grupo control, la diferencia en puntajes entre alumnos bajo la habilidad promedio (C.I. 90 promedio) y alumnos sobre la habilidad promedio (C.I. 115) era de alrededor de 10 puntos. Esto dio información adicional para ayudar a interpretar los efectos del programa. Los alumnos bajo la habilidad promedio que habían recibido el programa estaban actuando como alumnos sobre la habilidad promedio que no había recibido el programa.

El uso de una medida de habilidad como pretest no dará por lo general tanto poder como el que da el uso de un pretest que es equivalente al postest. Esto es porque un test de habilidad no estará, generalmente, tan íntimamente relacionado a los resultados del postest. Por ejemplo, usted podría hacer un pronóstico mejor del percentil en aritmética de un alumno para el final de un año, si en lugar de conocer solamente su percentil C.I., usted conociera su percentil en aritmética a comienzos del año. Pero su valor al interpretar cuan impresionante fueron los efectos de un programa hace que su

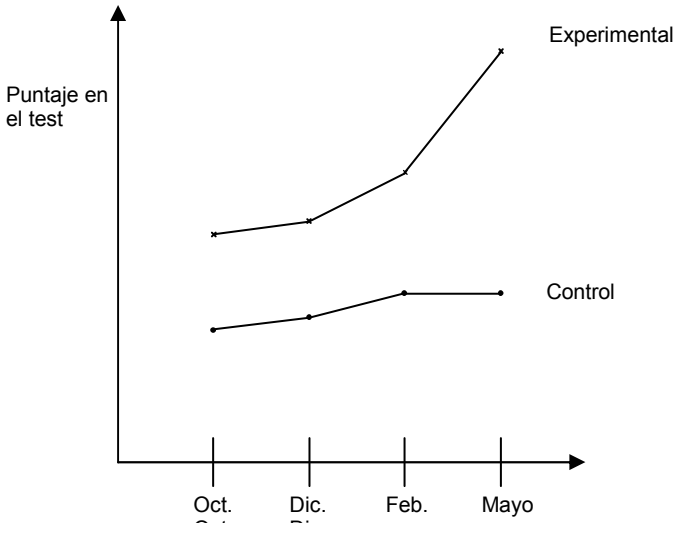
uso sea deseable. Una medida de habilidad como pretest es deseable en varias situaciones:

- cuando el hecho de saber cuál es la diferencia que producen los niveles de habilidad en los puntajes del postest le ayudará a interpretar el tamaño de la diferencia que un programa produce en los puntajes del postest. (Vea el último ejemplo dado anteriormente).
- cuando un pretest de asignatura no es posible debido al hecho de que la asignatura será totalmente nueva para la mayoría de los alumnos. Esto podría ocurrir, por ejemplo, con un curso para principiantes en lenguas extranjeras o con un curso de lógica.
- cuando el postest es una medida de actitud y usted cree que las actitudes pudieran diferir entre los niveles de habilidad (generalmente difieren),
- cuando usted se encuentra en una situación en que el programa comenzó sin haber dado ningún pretest a los grupos control. Una medida de habilidad puede usarse como un pretest retrospectivo.

Otros test que no son ni pretest ni postest.

Es posible que usted quiera hacer algunas mediciones durante el desarrollo del programa. Los tests administrados durante el desarrollo del programa o tests intermedios pueden indicar, por ejemplo, el impacto del programa a través del tiempo, como se muestra en los ejemplos siguientes.

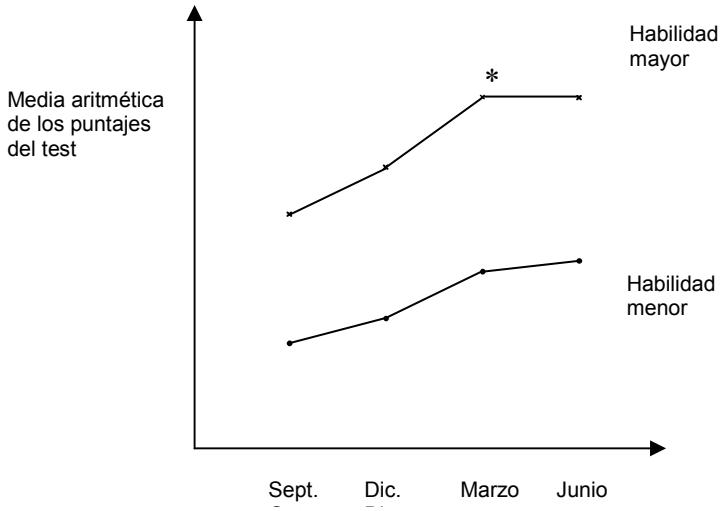
Ejemplo A. Puntajes en el test de matemáticas de octubre a mayo



Interpretaciones:

Después de un comienzo lento, el programa experimental mejoró cada vez más en lo que se refiere a ganancias en los puntajes del test.

Ejemplo B.



* Este puntaje indica que los alumnos de mayor habilidad habían recibido un beneficio máximo del programa en marzo y probablemente podían emprender otros estudios sólo con algún apoyo práctico en el programa para mantener su rendimiento. Los alumnos de menor habilidad continuaron haciendo progresos hasta el final del año.

Los tests intermedios pueden ser particularmente útiles cuando el diseño de su evaluación no tiene grupo control. En tal situación, usted tiene que concentrarse en examinar el progreso del programa que se evalúa. Una forma de lograr esto es observar de cerca el impacto en diferentes subgrupos, tales como los alumnos de mayor y menor habilidad en el ejemplo B.

Tests de Retención. "Ellos olvidan todo en las vacaciones de verano". Este es un comentario conocido y plantea todo el problema de retención. En una prueba de fin de año, un programa que aparentemente enseña bastante a los alumnos y en forma muy rápida, podría parecer bueno frente a un programa en el que se avanza lentamente; pero al final de las vacaciones de verano ambos grupos de alumnos podrían no distinguirse entre ellos en términos de lo que pueden recordar. Aún esto es una simplificación puesto que un grupo podría hacer progresos más rápidos cuando volviera a trabajar.

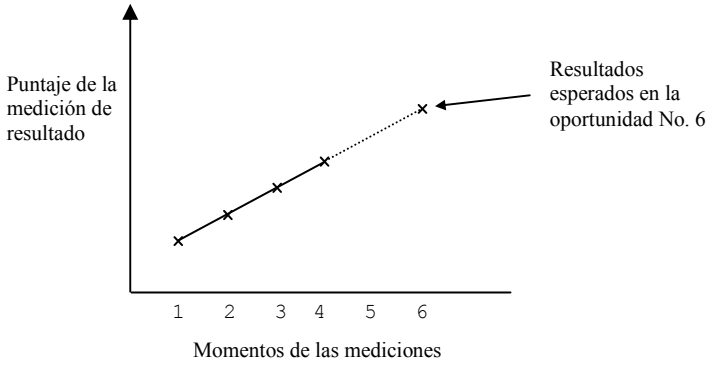
Recomendación. Si el programa E parece mayor (en términos de puntajes significativamente más altos de los alumnos) que el programa C al fin del año escolar, verifique los puntajes *otra vez* en septiembre volviendo a administrar el postest después de las vacaciones de verano. Si aún parecen mejor, usted tiene un programa realmente bueno. Si la diferencia significativa entre los puntajes no aparece más, no se pronuncie.

Tests de "Series Cronológicas". Cuando usted empieza a pensar en los tests de retención como opuestos simplemente a un "postest inmediato", queda claro que los tests únicos no son totalmente satisfactorios. Una serie de tests administrados después que un programa termina podría mostrar los efectos a largo plazo y verificar si los

beneficios del programa son duraderos o transitorios. Una serie de tests, generalmente dados a intervalos iguales de tiempo, antes y después del programa, se llama *tests de series cronológicas*.

Una serie de tests administrados antes que comience realmente un programa puede eliminar, en cierta medida, la necesidad de un grupo control. Un problema del diseño antes - después, un diseño que no tiene grupo control, es que es imposible saber cuáles pudieron haber sido los resultados sin el programa.

Una *serie de mediciones hechas antes que el programa comience* puede usarse para proyectar los resultados que serían esperados si las cosas continuaran sin modificación. Por ejemplo, dados los resultados que se incluyen en el gráfico que se muestra a continuación para las oportunidades 1, 2, 3 y 4, se podrían "predecir" los resultados en la oportunidad 6, por la extrapolación indicada por la línea de puntos. Note que *al menos tres* mediciones son deseables con el fin de trazar la línea de tendencia que es extrapolada. Estas tres mediciones deben ser con el mismo instrumento; por ejemplo, el mismo test o cuestionario o recuento de frecuencia.



Si se pueden reunir varias mediciones, antes y después de un programa, estas mediciones serán de ayuda considerable al interpretar los efectos de un programa.

En este capítulo se han considerado los dos elementos principales del diseño *a quién* se mide y *cuándo se mide*. Los diseños en los capítulos siguientes representan con mayor detalle las diversas combinaciones de posibles elecciones del grupo que se mide (qué) y el tiempo de medición (cuándo). La tabla siguiente muestra como se combinan estos elementos en los seis diseños que se analizan en los capítulos.

Hay tres elecciones de grupos a medir:

- Grupo experimental solamente
- Grupo experimental y un grupo control verdadero (seleccionado al azar)
- Grupo experimental y grupo control no-equivalente (no seleccionado al azar)

Y hay tres elecciones para el tiempo de medición:

- Pretest y Postest
- Postest solamente

- Series Cronológicas, una serie de al menos tres mediciones antes que se implemente el programa y tres después que termine.

Combinaciones de estas elecciones, de quién se mide y cuándo, forman los seis diseños, tal como se muestra en la Tabla 1.

TABLA 1

DISEÑOS PARA EVALUACIÓN SUMATIVA

		A QUIÉN SE MIDE		
		Grupo experimental solamente. Estos diseños pueden sólo responder preguntas relacionadas a la forma en que funciona un programa.	Más de un Grupo. Estos diseños pueden responder preguntas comparando los efectos del programa con alguna alternativa.	
			Control verdadero	Control no equivalente
Cuándo se hacen las mediciones	Pretest y Postest	DISEÑO 6	DISEÑO 1	DISEÑO 3
	Postest solamente	No recomendado	DISEÑO 2	No recomendado
	Series Cronológicas	DISEÑO 4	Bueno, pero poco frecuente	DISEÑO 5

Títulos de los seis Diseños

Diseño 1: El Diseño Grupo Control Verdadero, Pretest-Postest
Diseño 2: El Diseño Grupo Control Verdadero, Postest Solamente
Diseño 3: El Diseño del Grupo Control No-Equivalente
Diseño 4: El Diseño de series Cronológicas
Diseño 5: El Diseño de Series Cronológicas con un Grupo Control No- Equivalente
Diseño 6: El Diseño Antes-Después.

SELECCIÓN DE UN DISEÑO

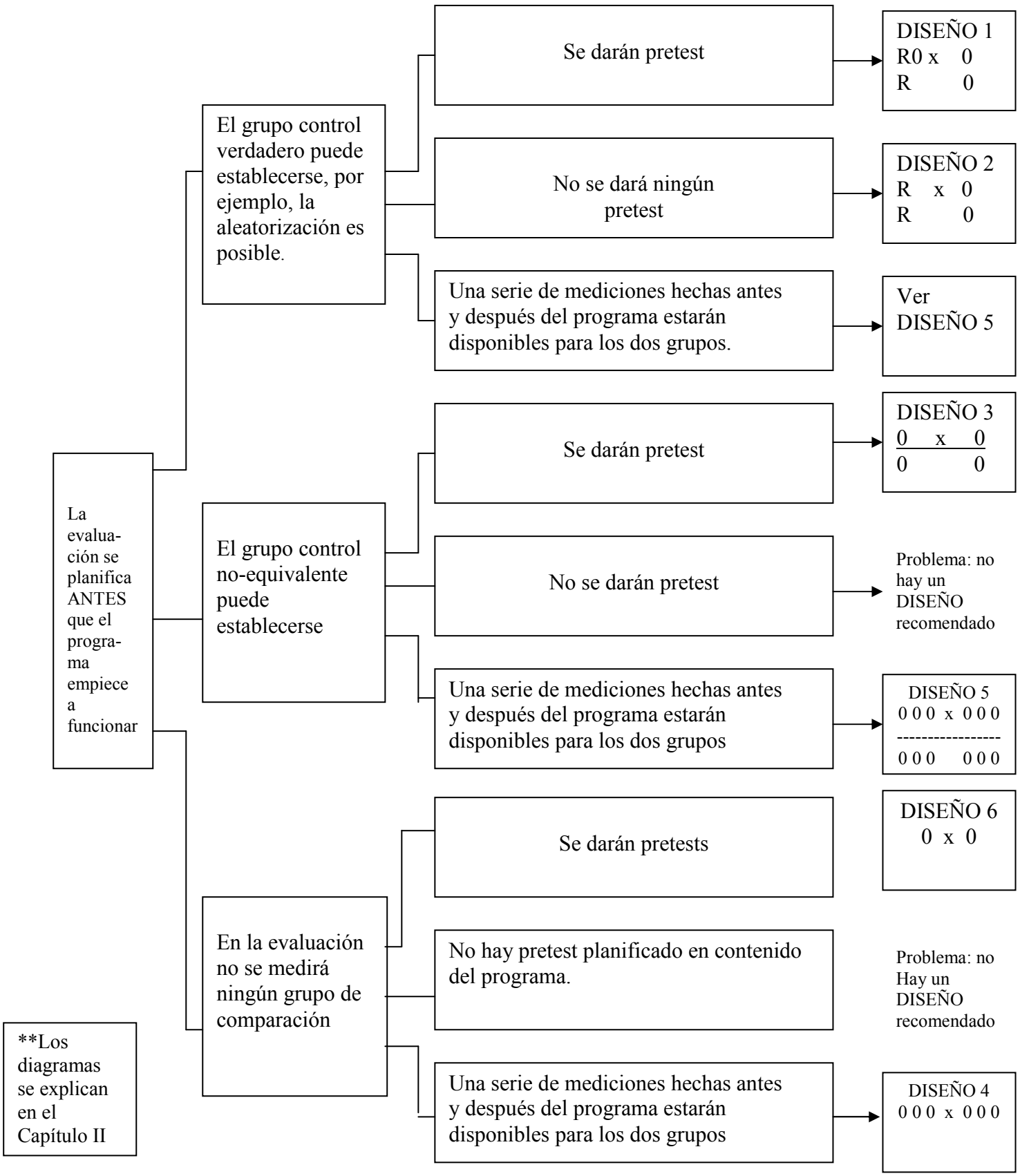
Quizás usted ahora puede seleccionar el diseño o diseños que usará para su evaluación simplemente examinando la Tabla 1 *¡Observe que usted puede seleccionar más de un diseño para la misma evaluación!* Cada instrumento podría tener un diseño propio; por ejemplo, usted puede querer hacer un diseño postest solamente para una medición de actitudes, un diseño pretest – postest para una medida de rendimiento estandarizada y un diseño de series cronológicas para una medida de rendimiento referida a criterio. Del mismo modo, usted podría pensar que puede encontrar un grupo de comparación apropiado para parte de los resultados de su programa (por ejemplo, una medida estándar de la historia de los Estados Unidos), pero esta otra parte de su programa (antropología para alumnos de 4º grado o lógica simbólica para alumnos de 8º grado) es única y no sería significativo tratar de medir un grupo de comparación en esa parte del programa.

Aunque el uso de varios diseños es una posibilidad muy clara, si usted está conduciendo una evaluación por primera vez, sería muy aconsejable que la dejara tan

pura y simple como fuera posible, seleccionando sólo un diseño apropiado. Si usted ha podido seleccionar el o los diseños que necesita, ahora puede encontrar una breve descripción del o los diseños en el Capítulo 3, y luego seguir un análisis más largo de cada diseño en el Capítulo 4 o 5. Si aún no ha seleccionado un diseño, los cuadros que se encuentran en las páginas siguientes pueden ayudarlo.

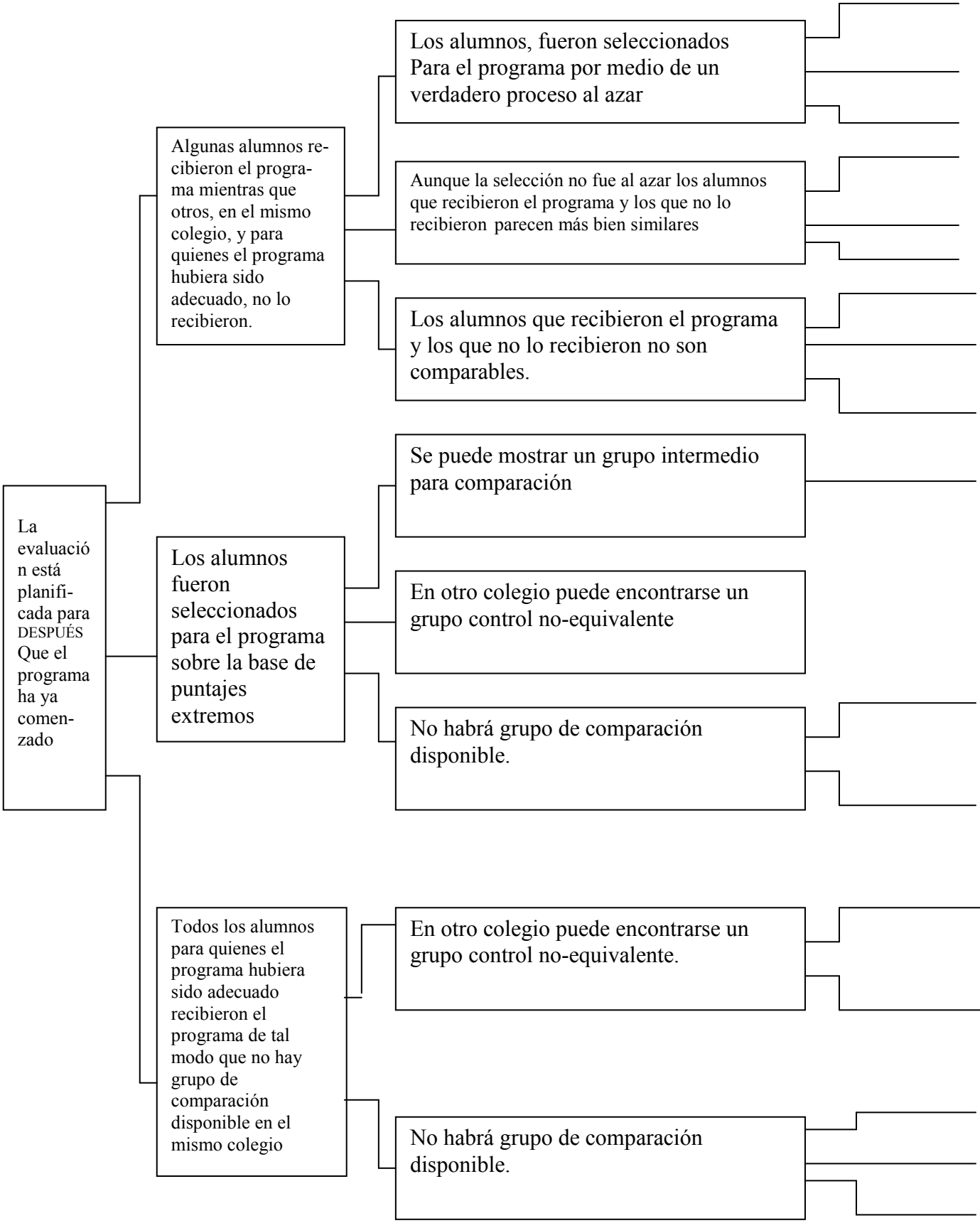
Use el cuadro 1 si usted está planificando la evaluación para antes que comience el ciclo del programa. En esta situación, usted elegirá un diseño sobre la base de cuál es el que piensa que será capaz de implementar. Si piensa que podría decidirse por elegir un gran diseño –uno con un grupo control verdadero (al azar), considere todas las posibilidades que se analizan en las páginas 11 y 12. Si aún no sable como llegar a un grupo control verdadero, haga la prueba con un grupo control no-equivalente que valga la pena usar. Si usted tampoco encuentra un grupo control no-equivalente o si tiene dudas en cuanto a encontrar uno bueno, considere un diseño de series cronológicas. *Solamente elija un grupo único, diseño antes – después, si no tiene otra alternativa.*

CUADRO 1
SELECCIÓN DE UN DISEÑO
ANTES QUE COMIENCE EL PROGRAMA

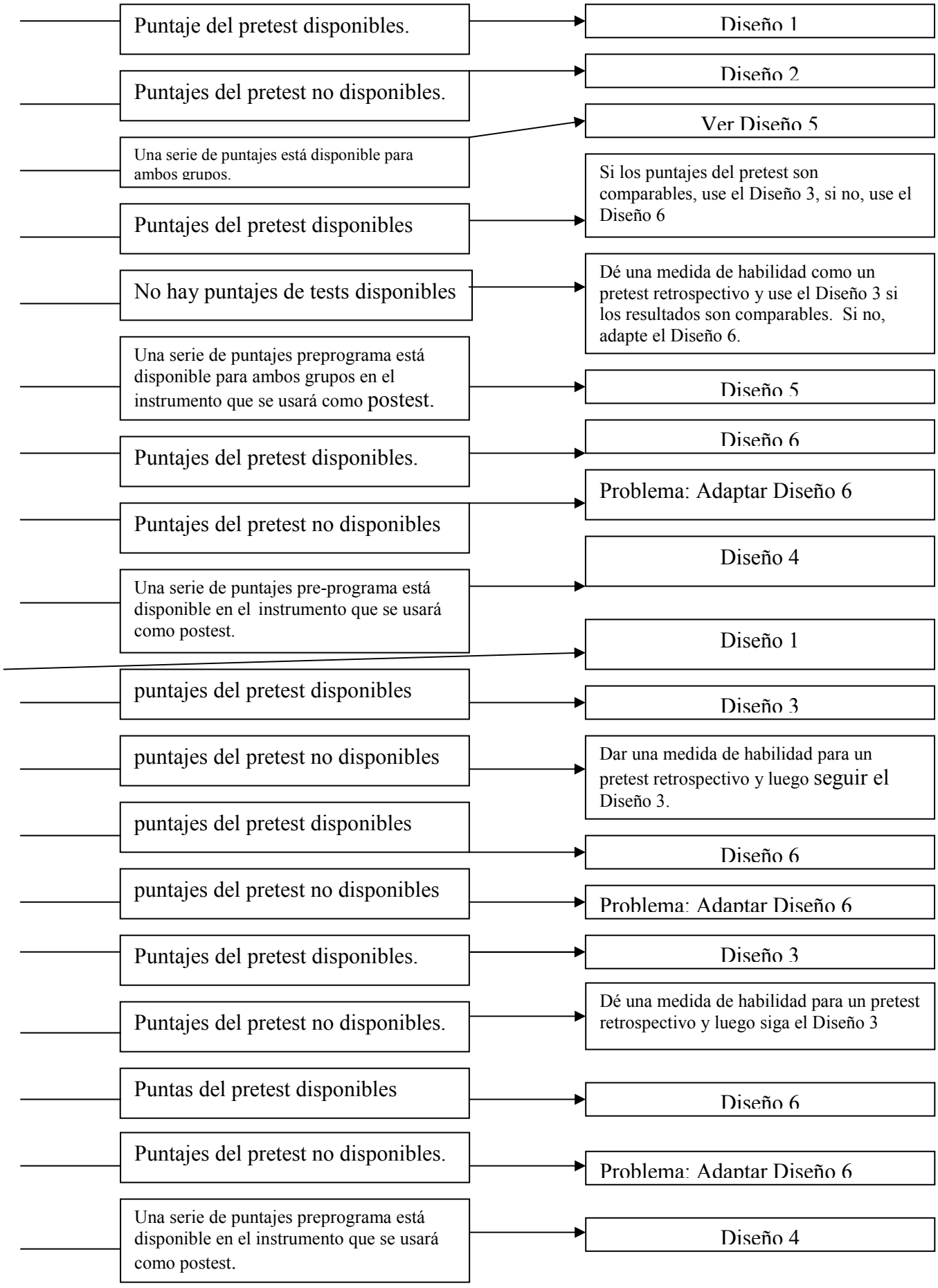


Use el cuadro 2 si usted está planificando realizar la evaluación en la mitad del ciclo del programa, durante el cual el programa debe evaluarse. En este caso, usted estará dando una mirada hacia atrás, antes que nada, para ver como se seleccionaron los alumnos para el programa. Puede ser que tenga que buscar registros o hacer preguntas al personal para descubrir exactamente cuál fue el proceso que se usó. Si a usted se le dice que la selección fue más o menos al azar, siga preguntando para descubrir cuál fue la parte "menos al azar". Por ejemplo, quizás hubo selección al azar, excepto en un alumno que formaba parte de la orquesta del colegio por razones de horario no pudo estar en el programa. En tal caso usted debería tener cuidado, si analizara la información, como si fuera un diseño de grupo control verdadero, de incluir *solamente* mediciones de aquellos alumnos que realmente eran parte de la selección al azar.

CUADRO 2


La evaluación está planificada para DESPUÉS Que el programa ha ya comenzado
Algunas alumnos recibieron el programa mientras que otros, en el mismo colegio, y para quienes el programa hubiera sido adecuado, no lo recibieron.
Los alumnos, fueron seleccionados Para el programa por medio de un verdadero proceso al azar
Aunque la selección no fue al azar los alumnos que recibieron el programa y los que no lo recibieron parecen más bien similares
Los alumnos que recibieron el programa y los que no lo recibieron no son comparables.
Los alumnos fueron seleccionados para el programa sobre la base de puntajes extremos
Se puede mostrar un grupo intermedio para comparación
En otro colegio puede encontrarse un grupo control no-equivalente
No habrá grupo de comparación disponible.
Todos los alumnos para quienes el programa hubiera sido adecuado recibieron el programa de tal modo que no hay grupo de comparación disponible en el mismo colegio
En otro colegio puede encontrarse un grupo control no-equivalente.
No habrá grupo de comparación disponible.

SELECCIÓN DE UN DISEÑO EN MEDIO DEL PROGRAMA



CAPÍTULO 2:

DISEÑOS: *UNA DESCRIPCIÓN BREVE*

Probablemente después de leer el Capítulo 1 usted llegó a una decisión en cuanto a qué tipo de *grupos* formar y medir, y en. Qué *momentos* recoger la información; usted estaba en condiciones de seleccionar de la Tabla 1 el diseño que le interesa implementar. Si es éste el caso, ahora querrá probablemente sólo leer la breve descripción que aparece en este capítulo sobre el diseño que seleccionó y luego deseará pasar directamente a las páginas de los Capítulos 3, 4 o 5 donde se aborda el diseño en detalle. Sin embargo, si usted aún no ha seleccionado un diseño, debería leer en este capítulo las descripciones breves que se hacen de cada uno de los seis diseños, tratando de seleccionar el o los diseños que usted implementará. Recuerde que usted puede seleccionar diseños diferentes para instrumentos diferentes.

La Simbología Que Se Usará Al Diagramar Cada Diseño

Se usará la notación simple que usa Campbell y Stanley en su artículo clásico. [*] Se usan los siguientes símbolos al diagramar los diseños:

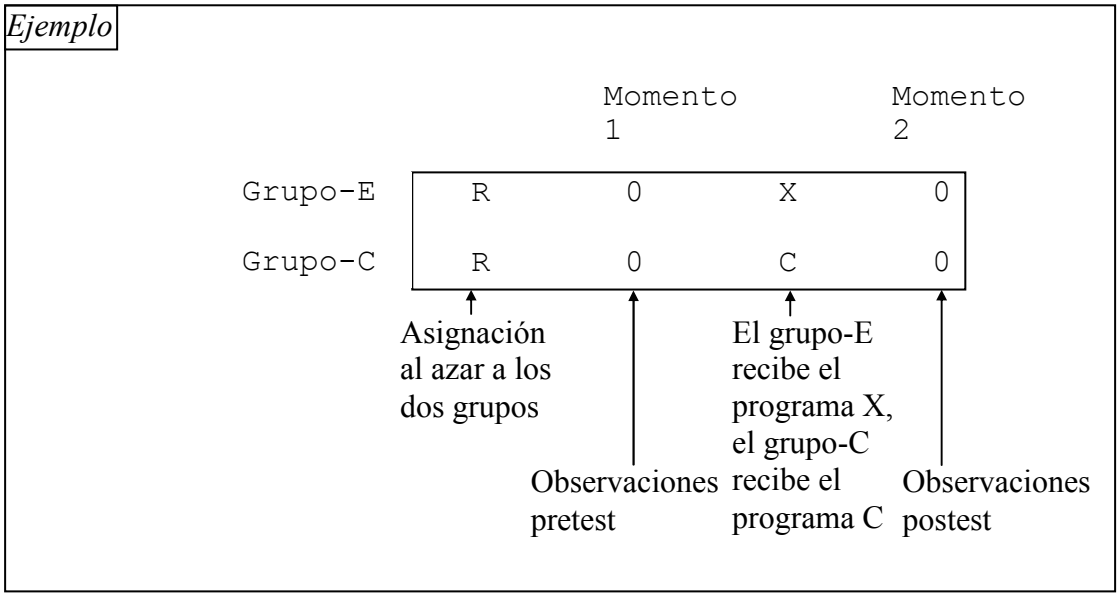
"R" = aleatorización o asignación al azar

"-----" = dibujado en forma horizontal, separando los dos grupos, indica que los grupos no fueron asignados al azar, por ejemplo, los grupos son "no equivalentes"

"0" = una medición de algún tipo, una "Observación"

"X" = el programa que se va a evaluar; el programa experimental.

Ejemplo



Esta simbología se usará en la presentación de los seis diseños que se hace a continuación. Estos seis diseños son aquellos que aparecen en las celdas de la Tabla 1, en la página 32. Observe que en los diagramas que siguen la "C", que representa el

[*] Campbell, D.T. and Stanley, J.C. *Experimental and Quasi-experimental Designs for Research.* Chicago: Rand McNally, 1966

programa que recibió el grupo control en el momento en que el grupo experimental recibió el programa X, se omite con el fin de mantener el diagrama lo más claro y preciso posible. Debe recordarse, sin embargo, que *no es necesario que el grupo control no reciba ningún programa.* En verdad, el grupo control idealmente debería recibir un programa alternativo que le permita competir con el programa X.

En lo que queda de este capítulo usted encontrará un diagrama y una breve descripción de cada uno de los seis diseños.

DISEÑO 1: EL DISEÑO DE GRUPO CONTROL VERDADERO, PRETEST-POSTEST

	1 (pre)			2 (post)
Grupo-E	R	0	X	0
Grupo-C	R	0		0

Este es el experimento clásico verdadero. Los alumnos para quienes el programa X es adecuado se asignan *al azar* para formar dos grupos, uno que *recibirá* el programa X y un grupo que *no lo recibirá.* El grupo que no recibe el programa X puede que no reciba ningún programa o puede recibir un programa alternativo (programa C). Los puntajes del pretest pueden usarse para comprobar que los grupos comenzaron siendo más o menos equivalentes. Si al final del programa la media aritmética de los puntajes del postest del grupo-E es significativamente más alta que la media aritmética de los puntajes del grupo-C, puede atribuirse esta diferencia al efecto del programa X, en oposición al efecto del programa alternativo. Este es un diseño muy bueno; permite que se haga una prueba poderosa entre el programa X y el programa alternativo.

DISEÑO 2: DISEÑO DE GRUPO CONTROL VERDADERO, POSTEST SOLAMENTE

	Tiempo		
	1 (pre)		2 (post)
Grupo-E	R	X	0
Grupo-C	R		0

Este es también un "verdadero diseño experimental". Es exactamente como el Diseño 1, excepto que no se usa "pretest". Este diseño es útil cuando un pretest podría interferir con los efectos del programa en algún sentido, o cuando no se encuentra disponible un pretest o cuando tomaría demasiado tiempo administrarlo. La asignación aleatoria de los alumnos ya sea al grupo-E o al grupo-C, generalmente asegurará grupos equivalentes. Este diseño se emplea a menudo cuando las mediciones que se van a hacer son mediciones de actitudes.

Observe que debido a la asignación aleatoria que debería haber igualado los grupos en la mayoría de las variables (especialmente si los grupos son bastante numerosos, es decir, 20 o más) es posible, en este diseño, esperar hasta el final del programa para decidir exactamente cuál será el postest. Igualmente, pueden agregarse tests que originalmente no se habían planificado. El Diseño 2 puede usarse fácilmente en conjunto con el Diseño 1. Usted podría administrar un pretest y un postest de rendimiento, por ejemplo, pero medir actitudes usando un postest solamente.

DISEÑO 3: DISEÑO DE GRUPO CONTROL NO-EQUIVALENTE, PRETEST-POSTEST

	Tiempo		
	1 (pre)		2 (post)
Grupo-E	0	X	0
Grupo-C	0		0

Este diseño pretest-postest con un grupo control no-equivalente (no al azar) es como el Diseño 1, excepto que el grupo control *no* se formó por asignación al azar. Para enfatizar el hecho de que el grupo control es no-equivalente (no al azar), el diagrama tiene una línea de puntos separando los grupos E y C. Este es un diseño muy útil para evaluación en los colegios. Los puntajes del pretest permiten hacer una comprobación sobre las semejanzas, al menos con respecto a la medición usada.

Este diseño puede usarse cuando usted no puede asignar al azar a los alumnos a los programas, sino que debe trabajar con cursos ya formados. Algunos cursos que no están recibiendo el programa experimental pueden formar un grupo control no-equivalente para los cursos que están en el programa experimental. (Si se dispone de un número de cursos suficiente, usted podría verificar si pueden asignarse al azar cursos más bien que alumnos originando el Diseño 1).

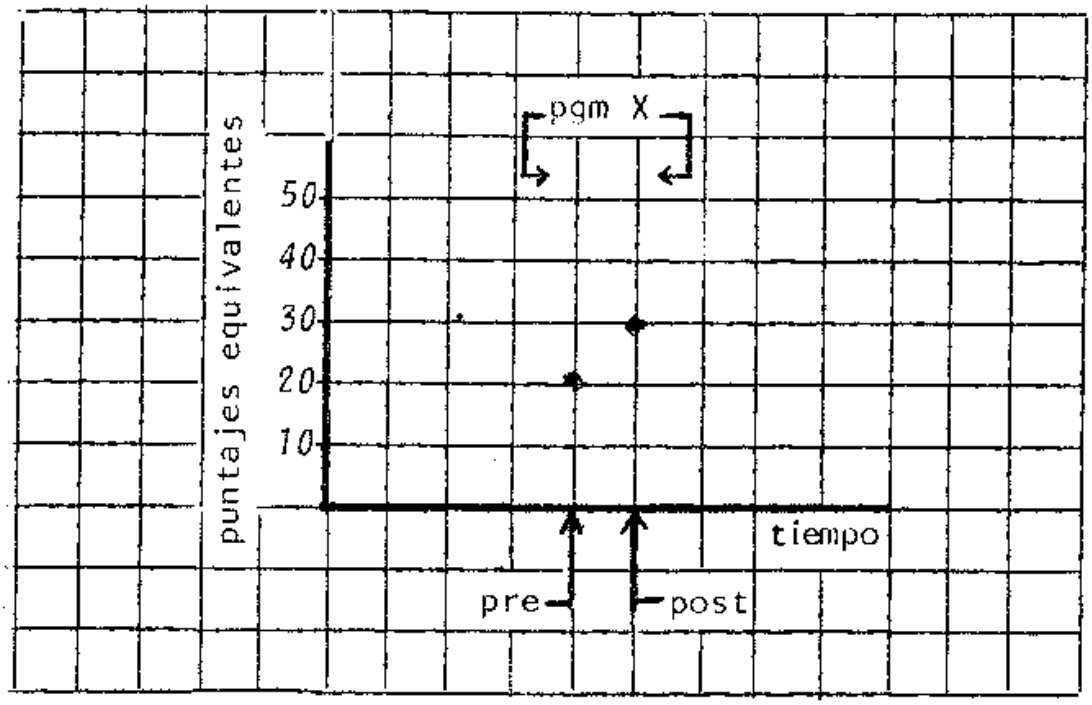
DISEÑO 4: DISEÑO DE SERIES CRONOLÓGICAS DE UN SOLO GRUPO

	Tiempo						
	Pre				Post		
	1	2	3		4	5	6
Grupo-E	0	0	0	X	0	0	0

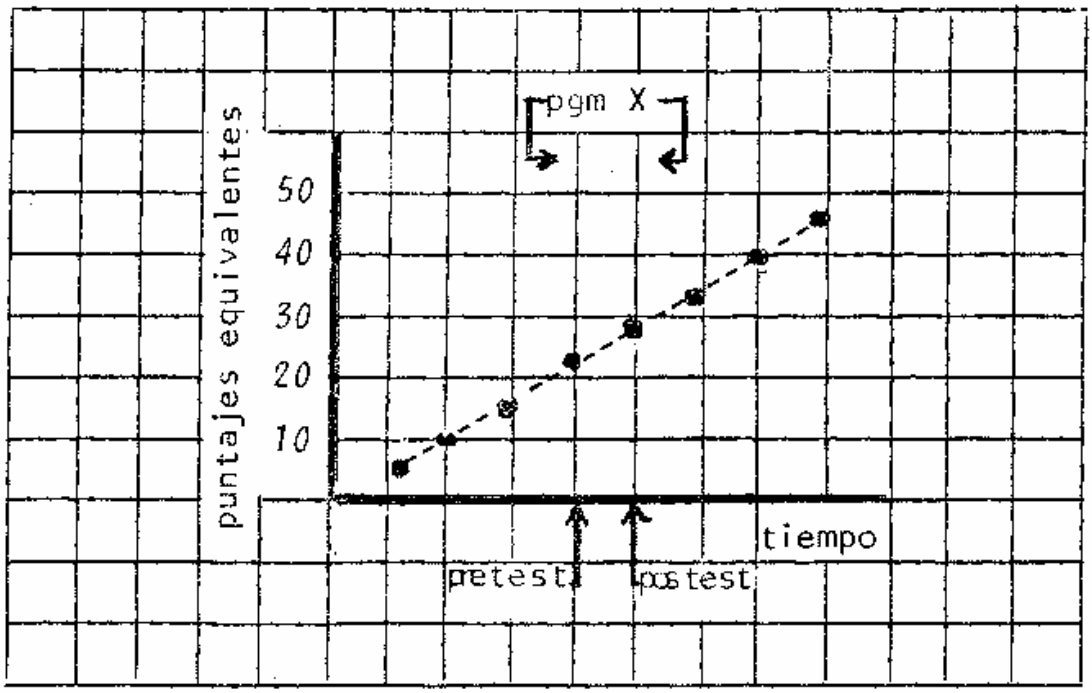
Este diseño utiliza a los alumnos que están en el programa como su propio grupo control. La *misma medición* se hace al *mismo grupo de alumnos* a *intervalos regulares de tiempo,* varias veces antes y varias veces después del programa. Viendo si el programa X parece perturbar la tendencia en los resultados, podemos determinar si el programa X parece o no parece haber tenido un impacto en la medida de los resultados.

Este diseño, como el diseño Antes-Después, requiere solamente mediciones en un grupo, pero es un gran avance en relación al diseño Antes-Después.

La forma en que en diseño de series cronológicas permite una mejor interpretación de resultados que lo que permite el diseño Antes-Después puede ilustrarse considerando las siguientes situaciones. Suponga que un diseño pretest-postest mostrara una ganancia en lectura alcanzada en cinco clases:

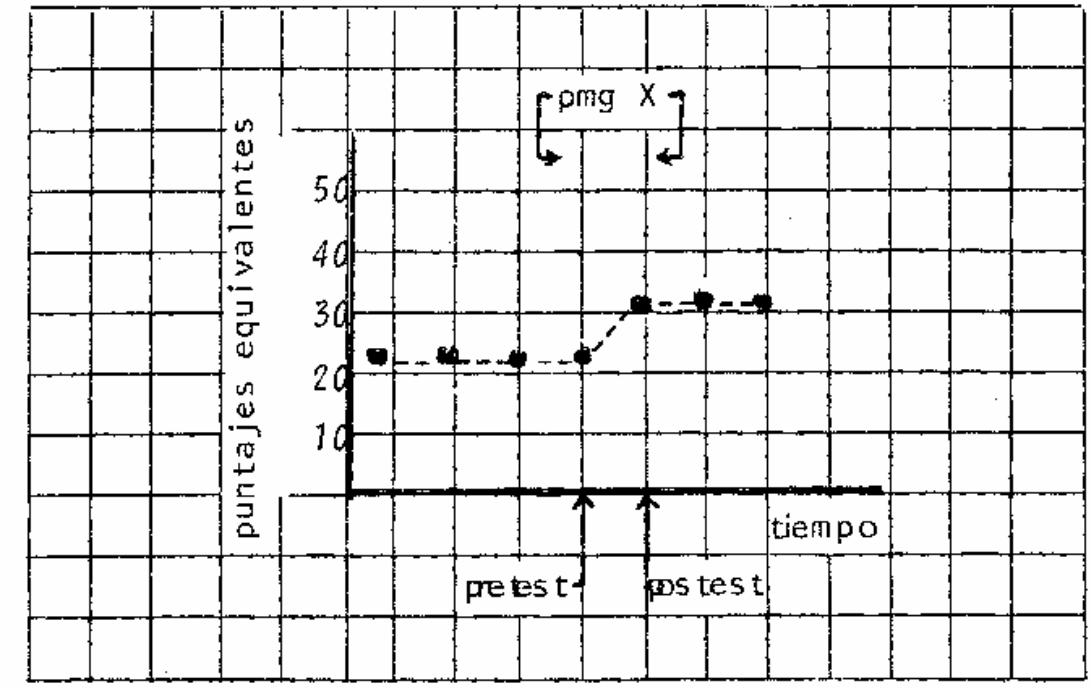


Ahora suponga que se agregaron mediciones de series cronológicas al mismo gráfico, y entonces el gráfico se viera así:



Parece que los puntajes estaban subiendo antes del programa y simplemente continuaron subiendo en la misma proporción. El programa no pareció hacer daño, pero tampoco pareció cambiar nada.

Sin embargo, supongamos que las mediciones de series cronológicas agregadas al mismo diagrama inicial lo hicieran aparecer así:



Ahora la interpretación de la misma ganancia pretest-postest es totalmente diferente. El programa parece haber causado un salto hacia arriba en puntajes de lectura y produjo la única gran ganancia registrada durante las mediciones en series cronológicas.

De este modo, la gran ventaja de un diseño de series cronológicas sobre un diseño de un solo grupo antes-después es que, obteniendo una serie de mediciones antes y después de la implementación de un programa, el efecto que tiene dicho programa puede medirse con más exactitud que cuando tiene solamente una medida antes y una medida después del programa.

El problema principal con el diseño de series cronológicas es el siguiente: incluso si usted observa un cambio claro en las observaciones que siguen a la implementación del programa X, es difícil saber si X *causó* el cambio o si X se aplicó más o menos en el momento en que las medidas hubieran cambiado de todas maneras, debido a otros sucesos. Sin embargo, usted puede frecuentemente excluir dichas coincidencias.

Un diseño de series cronológicas es una forma excelente para controlar la actuación del programa año, en una especie de evaluación permanente. Ya sea que las observaciones repetidas sean tests estandarizados, test referidos a criterios o cuestionarios, en la medida en que el test se conserve igual, una representación gráfica de los puntajes puede proporcionar un cuadro claro del progreso año a año.

Dos tipos de Diseños de Series Cronológicas

El diseño de series cronológicas se presenta en dos formas. Una supone mediciones sobre el mismo grupo de alumnos. Esta se puede llamar *tipo longitudinal* del diseño de series cronológicas para distinguirlo del otro, más común, el *tipo de grupos*

sucesivos. En un diseño de series cronológicas del tipo de grupos sucesivos se mide más bien la misma *categoría* de alumnos cada vez y no los mismos *verdaderos* alumnos. Por ejemplo, en lugar de medir y graficar puntajes en lectura para los mismos alumnos a medida que pasan de primero a segundo, a tercer grado (series cronológicas longitudinales) usted podría medir y graficar los puntajes de los alumnos de primer grado cada año, durante varios años, por ejemplo, se miden grupos sucesivos. En una serie cronológica de grupos sucesivos probablemente es difícil saber si los cambios que se observan se deben a la naturaleza del grupo determinado de alumnos o a la naturaleza del problema que reciben los alumnos.

Los dos tipos de diseños de series cronológicas, grupos longitudinales y grupos sucesivos, se analizan con más detalle en el Capítulo 4.

DISEÑO 5 : EL DISEÑO DE SERIES CRONOLÓGICAS CON UN GRUPO CONTROL NO-EQUIVALENTE

	Tiempo						
	1	2	3		4	5	6
		(Pre)				(Post)	
Grupo-E	0	0	0	X	0	0	0
Grupo-C	0	0	0		0	0	0

Este diseño es exactamente como el Diseño 4, pero con la adición de un grupo control no-equivalente. Se miden generalmente dos grupos de alumnos antes del programa y luego un grupo recibe el programa X, pero el otro no (el otro grupo podría recibir un programa alternativo o no recibir programa).

La adicción de un grupo de comparación al Diseño 4 hace un diseño mucho más poderoso. Si hubiera algún acontecimiento externo que resultara coincidir con el programa X, sus efectos deberían aparecer tanto en el grupo-C como en el grupo-E. De este modo, esta explicación específica alternativa de los resultados puede, a menudo, excluirse cuando hay un grupo control no-equivalente. Además, como se indica en el rectángulo de puntos [ ], este diseño incorpora el diseño de grupo control no-equivalente, pretest-postest (Diseño 3), y así puede darle a usted toda la información que ese diseño le da, además de la otra. Si usted está pensando usar el Diseño 5, lea sobre el Diseño 4 en las páginas 78 a 96.

DISEÑO 6 : EL DISEÑO ANTES – DESPUÉS

	Tiempo		
	(pre)		(post)
Grupo-E	0	X	0

Un grupo de alumnos toma un pretest, recibe el programa X y luego toma un postest. Los resultados podrían compararse con un grupo de "referencia" o "norma" o con lo que se esperaba o se tenía la esperanza de obtener. Existen considerables problemas cuando se trata de establecer cualesquier tipo de juicios en torno al programa X sobre la base de este diseño.

Sin embargo, en el Capítulo 6 de este libro hay varias sugerencias para sacar el mejor provecho posible de este diseño que continúa siendo común en evaluación, incluso en evaluaciones sumativas.

Principales Amenazas a la Implementación de los Diseños

Seleccionar un diseño es una cosa, conseguir establecerlo con precisión y hacerlo funcionar en forma impecable es otra cosa.

Incluso si usted puede establecer un grupo control verdadero, o un grupo control no-equivalente, aún hay problemas potenciales en la implementación real del diseño. Por ejemplo:

Diferencias entre los grupos E y C en tiempo dedicado al programa. Un factor que cada vez más se reconoce que afecta el rendimiento de los alumnos en una asignatura es *el tiempo* que el alumno dedica a la asignatura. Un programa que proporciona instrucción por un largo tiempo puede producir mejores resultados entre los alumnos, no importa cual sea la *calidad* comparativa de esa instrucción. Por lo tanto, puede ser que al analizar los resultados de un programa usted desee considerar el tiempo asignado diariamente a éste y que también desee analizar los registros de asistencia de los alumnos. Usted puede incluso decidir comparar los resultados de un grupo-E con los resultados del grupo-C considerando solamente aquellos alumnos que fueron expuestos a la misma cantidad de tiempo de instrucción.

Atrición. Los alumnos abandonan el programa del grupo experimental o del grupo control por muchas razones: enfermedad u otro tipo de ausencia del colegio, peticiones de traslado a un curso diferente, traslado del sector escolar, etc. Este abandono puede afectar los resultados de los programas. Si por ejemplo los alumnos más atrasados abandonan un programa nuevo, este programa obtendrá puntajes promedio más altos simplemente debido a la pérdida de los alumnos más atrasados. El abandono de alumnos de los grupos seleccionados originalmente para formar los grupos E y C se llama "atrición", "mortalidad", o pérdida de casos en la literatura sobre diseños.

El problema de confusión o de influencias ajenas

En un diseño de dos grupos, tales como el Diseño 1, 2 o 3, una confusión es algo ajeno al programa que le sucede a uno de los grupos (grupo-E o grupo-C), pero que no le sucede al otro grupo y que podría influir en los resultados del programa. De este modo, si el grupo-C saliera del colegio media hora antes que el grupo-E, este solo hecho podría hacer que las actitudes de los alumnos fueran más favorables hacia el programa del grupo-C. Con un diseño de un grupo tal como el Diseño 4 o 6 (el diseño Antes-Después) una influencia ajena o "confusión" es un suceso ajeno al programa en observación, la que ocurre al mismo tiempo que se desarrolla el programa y podría esperarse que influenciara las mediciones de resultados. Por ejemplo, si en un diseño de series cronológicas el grupo en estudio recibiera un profesor nuevo al mismo tiempo que comenzaron a usar el programa nuevo, sería casi imposible saber si fue el profesor nuevo lo que causó algún cambio detectado, o si fue el programa nuevo.

Trabajar con un número grande de unidades experimentales protege en cierto modo contra las influencias ajenas. Si usted estuviera trabajando con 20 cursos, no sería posible que los 20 recibieran profesores nuevos precisamente en el momento de recibir el programa nuevo. En general, al comparar los grupos-E con los grupos-C

compuestos de muchos casos, las diferencias en lo que les sucede, fuera del programa mismo, tenderán a promediarse a veces favoreciendo al grupo-E, a veces al grupo-C.

El problema de *contaminación*

La contaminación ocurre cuando los métodos o materiales de un programa nuevo que el grupo-E está recibiendo los usa, en alguna medida, el supuesto grupo control. Por ejemplo, el profesor del grupo-E puede compartir películas con su amigo que está enseñando en el grupo-C. O, en discusiones en la sala de profesores, las mejores ideas y métodos del grupo-E pueden ser tomadas por el profesor del grupo-C. Obviamente, el problema de contaminación es más agudo cuando los grupos E y C están en el mismo colegio.

Una forma de protegerse contra la filtración del programa al grupo control contra la contaminación) es discutir el problema con el personal involucrado, explicarles la dificultad de evaluar el programa nuevo si usted no puede aislarlo, y luego pedirles sugerencias y cooperación. La contaminación puede ser un problema menor cuando los grupos E y C están en colegios diferentes y/o cuando el grupo C también está recibiendo un programa nuevo, de tal modo que los profesores están suficientemente ocupados con sus propios métodos y materiales nuevos y es poco probable que estén adoptando los métodos y materiales del grupo-E.

Los cuatro problemas recién discutidos, diferente cantidad de tiempo dedicado a los programas, atrición, influencias ajenas y contaminación, son todos problemas a los que usted debe poner atención cuando documente el programa. El grado en que han ocurrido estos problemas debe considerarse antes que se interpreten los resultados.

CAPÍTULO 3

DISEÑOS 1, 2 y 3

PROCEDIMIENTOS, ANÁLISIS E INTERPRETACIÓN

En este capítulo se describen en detalle los Diseños 1, 2 y 3, ubique el diseño que ha decidido usar. Usted encontrará un diagrama de flujo que le mostrará los pasos esenciales en la implementación del diseño (estos pasos pueden parecerle decepcionantemente simples). Hay también un par de ejemplos para cada diseño. A continuación de los ejemplos se incluye una discusión sobre el análisis y presentación de la información con respecto al diseño.

DISEÑO 1

EL DISEÑO DE GRUPO CONTROL

VERDADERO, PRETEST - POSTEST

DISEÑO 1

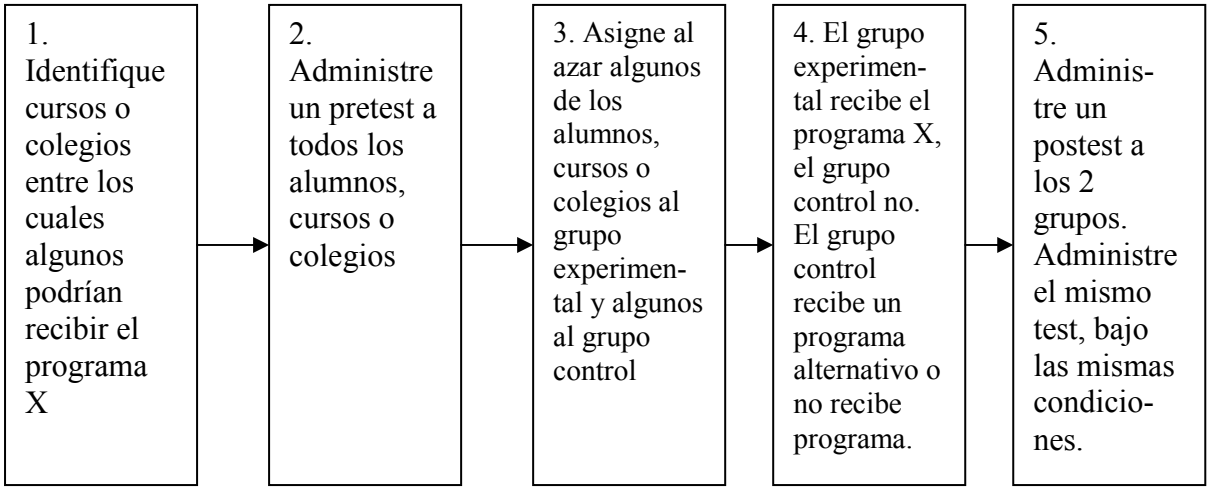
EL DISEÑO DE GRUPO CONTROL VERDADERO, PRETEST-POSTEST

DIAGRAMA

	Tiempo			
	1 (pre)			2 (post)
Grupo Experimental	R	0	X	0
Grupo Control	R	0		0

Resumen. Se forman aleatoriamente dos grupos equivalentes, pueden ser alumnos, cursos o colegios los que se seleccionan al azar. Los dos grupos se miden antes que comience el programa (el pretest) y después que el programa ha tenido tiempo para producir un efecto (el postest).

PASOS ESENCIALES AL IMPLEMENTAR EL DISEÑO 1



NOTA: *Los pasos dos y tres podrían ocurrir al revés, lo que correspondería al diagrama que muestra que "R"—aleatorización— ocurre antes que el pretest (por ejemplo, antes que el momento 1). La aleatorización tiene que seguir al pretest si se emplean bloques sobre el pretest, explicado en la página 127.

EJEMPLOS:

Ejemplo 1. El ejemplo de la Máquina de lectura

Un colegio había comprado una máquina diseñada para ayudar a los alumnos cuando estuviesen aprendiendo a leer. Algunos materiales para la máquina se habían comprado, pero puesto que representaban una cantidad significativa del dinero disponible en el colegio, había necesidad de saber cuán valiosos eran dichos materiales. Con el fin de determinar esto, el profesorado tomó las medidas para conducir un experimento verdadero. Se elaboró un test para

determinar los resultados del trabajo diario de tres semanas con los materiales que ya se habían comprado para la máquina. Los alumnos a quienes correspondía este trabajo, fueron seleccionados y divididos aleatoriamente en dos grupos. Los alumnos de un grupo (el grupo experimental) trabajaron con la máquina de lectura todos los días durante tres semanas, mientras que los alumnos del otro grupo (el grupo control) simplemente continuaron con su programa regular de lectura. El programa regular se modificó, sin embargo, para enseñar el mismo vocabulario que se enseñaba con la máquina de lectura. Al término de tres semanas, a los alumnos se les administró un test para ver si el grupo que trabajó con la máquina de lectura (grupo experimental había logrado mejores aprendizajes que el grupo control.

Ejemplo 2. Sistema de Administración basado en objetivos

Se había hecho una sugerencia a la Junta Escolar en el sentido que el distrito adquiriera e implementara un sistema de administración basado en objetivos (ABO) desarrollado para los programas de matemáticas de séptimo y octavo año. No dispuestos a comprometerse con tal gasto sin tener evidencias de la efectividad del sistema, el distrito pidió un ensayo limitado y una evaluación. La junta enfatizó que a los colegios no se les debía exigir que participaran en este ensayo.

El personal encargado de la investigación pidió a los colegios de enseñanza básica superior (7º y 8º año) que indicaran si participarían o no en el sistema ABO, en el caso de que estuviera disponible. El personal encargado de la investigación dividió a los colegios en colegios de alto, medio y bajo rendimiento sobre la base de los resultados del año anterior y luego seleccionó al azar dos colegios de alto, dos de medio y dos de bajo rendimiento para el grupo experimental y lo mismo para el grupo control. Se administraron pretest en los 12 colegios seleccionados de esta forma. El sistema ABO estuvo disponible entonces para los seis colegios del grupo experimental. Al término del año se administró un postest de rendimiento en matemáticas a los 12 colegios. En su informe a la junta el personal encargado de la investigación estuvo en condiciones de entregar evidencia clara con respecto al impacto que el sistema ABO había tenido en el rendimiento. Ellos enfatizaron que sus conclusiones eran aplicables única y exclusivamente a los colegios que voluntariamente aceptaron el sistema y que los resultados, que fueron positivos, podrían ser diferentes si el sistema fuera impuesto más bien que aceptado. La Junta felicitó a la comisión de investigación por la minuciosidad y claridad del diseño.

PRESENTACIÓN Y ANÁLISIS DE LA INFORMACIÓN PARA EL DISEÑO 1

La gran virtud de este diseño es que cuando se ha implementado adecuadamente usted puede sacar conclusiones muy poderosas sobre el efecto del programa X en los puntajes del "postest". Además le permite detectar mejoramientos, incluso pequeños, a corto plazo; es decir, puede ser una prueba sensitiva de un programa.

Hay dos tipos básicos de información que se debe entregar con el fin de usar adecuadamente el Diseño 1: información de implementación sobre el diseño e información de resultados.

Informe sobre la Implementación del Diseño

Mostrar con cuanto éxito se implementó el diseño, implica considerar las siguientes preguntas:

1. Implementación del Programa. ¿Recibió el grupo-E programa? Esta pregunta la responderá la documentación del programa.

2. Contaminación. ¿No recibió el grupo-C el programa?, por ejemplo, ¿se evitó la contaminación? Esto requiere, al menos, alguna documentación de lo que sucedió en los cursos del grupo-C. *

3. Confusión. ¿Hubo diferencias consistentes entre lo que sucedió al grupo-E y al grupo-C sin considerar el programa X, diferencias que con razón podría esperarse que subieran o bajaran los puntajes en las mediciones de resultados. Es decir, ¿hubo confusión o factores ajenos que confunden los resultados y que tendrán que tomarse en cuenta cuando se interpretan los resultados?

4. Atrición (pérdida de casos). ¿Hubo alguna diferencia entre el número de "casos" perdidos en el grupo-E y el número de casos perdidos en el grupo-C? ¿Hubo alguna diferencia en los tipos de casos (alumnos, cursos o colegios) que tuvieron que eliminarse del análisis en cada grupo? La incorporación en su informe de una tabla como la que se muestra a continuación puede constituir un resumen apropiado.*

NÚMERO DE ALUMNOS EXCLUIDOS DEL ANÁLISIS
POR DIVERSAS RAZONES

Razón	Número de alumnos excluidos del grupo-E	Número de alumnos excluidos del grupo-C
Ausente para el postest		
Ausente del colegio durante el programa.		
Trasladado del grupo a petición de los padres		
Abandonó el colegio		
Otras razones		
Número total de alumnos excluidos del análisis		

* Contaminación, confusión y atrición se discutieron en el Capítulo 2 bajo el título "Principales Amenazas a la Implementación del Diseño".

Debido a que el grupo-E y el grupo-C pueden ser de diferente tamaño, sería aconsejable poner entre paréntesis el porcentaje que representa la cantidad "número eliminado" del total de casos del grupo pretest. Por ejemplo, si un grupo-E se componía de 300 alumnos, para quienes hubo resultados de pretest, y 30 fueron excluidos del análisis porque estuvieron ausentes del colegio por más de un determinado número de días durante el programa, en la segunda fila la entrada en la columna del grupo-E sería "30 (10%)". Esto muestra que 30 alumnos se excluyeron, lo que representa el 10% del grupo-E original $\left[\frac{30}{300} \times 100 = 10\%\right]$

Cuando el programa ha sido asignado aleatoriamente a cursos o colegios, a menudo sucede que algunos cursos o colegios deciden que no pueden implementar el programa como se ha planificado, ellos abandonan en alguna etapa durante el desarrollo del programa. Cuando un curso o colegio en el grupo control rehúsa tomar parte en el postest, también causa atrición; hay un menor número de unidades disponibles para análisis que las planificadas para ello.

Interpretación de la tabla que mostró pérdida de casos. Si la tasa de deserciones para los grupos E y C fuera muy diferente, trate de explicarlo. Esto puede requerir que se hagan consultas, que se trate de descubrir lo que estaba sucediendo. Si usted sospecha que los alumnos con ciertas características abandonaron el programa, podría tabular información adecuada y examinarla.

Ejemplo. Un coordinador de proyecto indicó al evaluador que un mayor número de alumnos de menor habilidad en relación al número de alumnos de mayor habilidad habían abandonado un programa de trabajo. El notó que 12 alumnos de menor habilidad y sólo 8 de mayor habilidad habían desertado. El evaluador reunió esta información:

	Desertores		Total
	Mayor habilidad	Menor habilidad	
Grupo-E	12 (60%)	8 (40%)	20
Grupo-C	10 (59%)	7 (41%)	17
Total	22	15	37

El evaluador señaló que aunque era verdad que un mayor número de alumnos de menor habilidad que alumnos de mayor habilidad habían abandonado el grupo-E, la misma situación se presentaba también para el grupo-C. El porcentaje de alumnos que abandonaba era muy parecido en los grupos E y C.

Una buena forma de generar hipótesis razonables sobre lo que podría causar atrición es entrevistar a una muestra de alumnos elegida al azar entre aquellos que han abandonado cada uno de los grupos, el grupo E y el grupo C.

Informe de Resultados

A continuación hay una tabla simple para informar sobre los resultados del diseño 1, el diseño grupo Control Verdadero, Pretest-Postest.

TABLA 2

RESULTADOS DEL PRETEST Y POSTEST PARA LOS GRUPOS EXPERIMENTAL Y DE CONTROL

	N	Pretest			Postest		
		Media	SD	Test-t de diferencia	Media	SD	Test-t de diferencia
Grupo-E							
Grupo-C							

N es el número que se mantiene para análisis, que es igual al número de casos para los cuales hubo resultados de pretest menos el número total de casos no considerados en el análisis. Las medias aritméticas son los puntajes promedio para cada grupo en cada test y la desviación estándar (SD) relacionada a cada media aritmética se entrega en la columna adyacente. Cada test-t es un test de la significancia de la diferencia entre los puntajes de la media aritmética para los grupos control y experimental. Si usted usa, ya sea la asignación al azar simple o "por bloqueo" (ver página 127) para formar los grupos, el test apropiado para usar es el test-t para grupos no pareados, a veces llamado el test-t no correlacionado. Si usted usara un método de pareos (página 128), use el test-t para grupos pareados.

Ejemplo. (Resultados de un experimento real)

RESULTADOS DEL PRETEST Y POSTEST PARA TUTORES Y NO TUTORES

	N	Pretest (habilidad)			Postest (Aritmética)		
		Media	SD	Test-t de diferencia	Media	SD	Test-t de diferencia
Tutores	8	24.00	7.25	0.25	30.30	5.10	4.16*
No tutores	14	24.86	7.12		15.80	8.60	

* Estadísticamente significativo al nivel .05

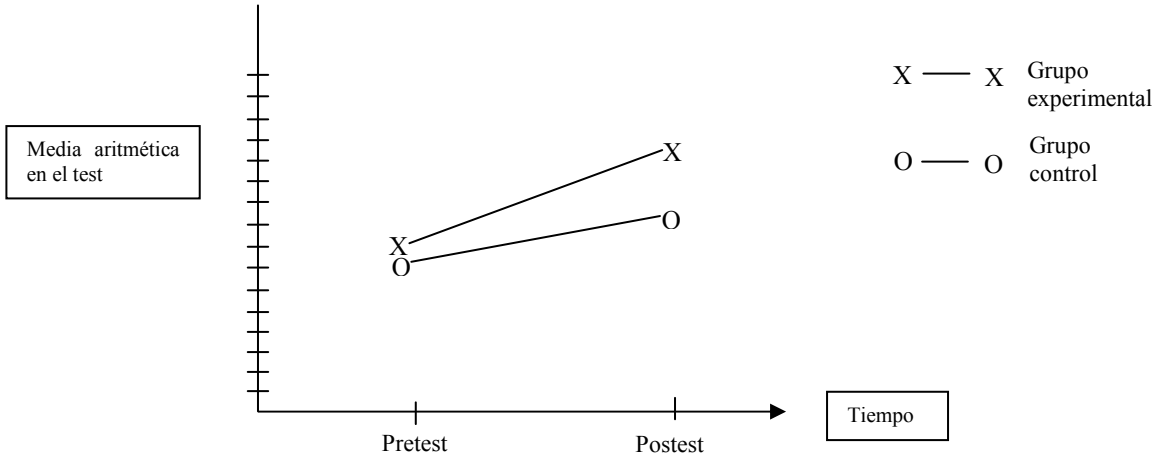
El ejemplo ilustra la forma en que estos resultados se interpretan; el test-t para los puntajes del pretest no fue significativo (esto se indica por la falta de asterisco)[*]. Esto significa que en el pretest, los dos grupos fueron más o menos equivalentes;

[*] Habiendo asignado al azar alumnos o cursos a los grupos, es muy poco probable –pero no imposible– que usted encuentre una diferencia estadísticamente significativa entre los resultados del pretest. Si esta rara mala suerte sucediera, entonces haga una de estas cosas: (1) vuelva a asignar al azar, si es posible, y use asignación al azar equiparada, o (2) siga el consejo que se da en la página 74 para abordar el problema en el Diseño 3 donde es más común.

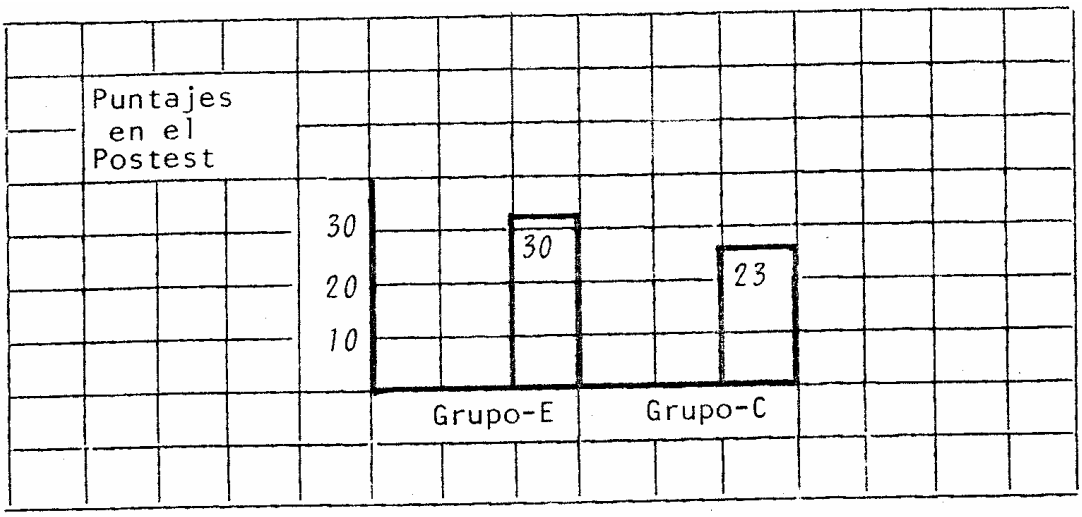
cualesquier diferencia entre la medias aritméticas no fue más de lo que uno podría esperar debido al azar. Habiendo establecido que los grupos eran equivalentes al comenzar, uno observa los resultados después del programa y compara las medias aritméticas de los dos grupos, otra vez usando un test-t. En el ejemplo que se muestra, hay una gran diferencia entre las medias aritméticas y no es sorprendente que el test-t muestre que fue significativa, es decir, más grande que lo que se podría haber esperado que ocurriera al azar. En el ejemplo que se muestra el evaluador podría establecer:

"Como se esperaba, debido al use de la selección al azar, los tutores y no tutores fueron inicialmente equivalentes en la medición de habilidad que se usó como pretest. Después del programa, sin embargo, la media aritmética de los puntajes de los tutores fue significativamente más alta que la media aritmética de los no tutores en el postest. Esta ganancia por parte de los tutores se puede atribuir a su participación en el programa de tutoría".

Antes de discutir la interpretación de los resultados en mayor detalle, es útil ver algunas formas en que los resultados pueden mostrarse gráficamente. La representación gráfica es muy deseable al informar sobre conclusiones de avaluación puesto que los gráficos se entienden e interpretan rápidamente. Además, las representaciones gráficas son casi siempre preferibles cuando se presentan resultados a una audiencia que está presente cuando se entrega el informe.
Si el pre y postest fueran los mismos test, entonces esta forma de representación gráfica siempre será adecuada:



Si el pretest fue diferente del postest (por ejemplo, el pretest pudo haber sido un test de habilidad), entonces muestre solamente los puntajes del postest. Este es el método más simple de mostrar y es perfectamente aceptable si no hubo diferencia significativa entre los puntajes del pretest.



Ahora examinaremos un poco más la interpretación de la información. Por favor, observe que la pregunta crucial es "¿fue la media aritmética del postest para un grupo, significativamente diferente de la media aritmética del postest para el otro grupo?"

No es un buen procedimiento comprobar la ganancia entre un pretest y un postest para ambos grupos. Suponga que ambos grupos lograron ganancias significativas; usted aún necesita saber si entre ellos hubo progresos diferentes.

Usted no puede usar el tamaño del valor –t para saber si un grupo fue mejor que el otro, porque el tamaño del valor-t está influenciado por el número de alumnos en el grupo. Incluso si ambos tuvieran el mismo número, usted necesitaría saber si cualesquier diferencia observada entre valores-t era una diferencia <u>significativa</u>. En resumen: compare <u>los mismos</u> tests entre los dos grupos. <u>No</u> el mismo grupo entre el pretest y el postest.

R 0 X 0 ⟵ Esta es la comparación que
R 0 0 debe hacerse

<u>Comparación de puntajes del Pretest.</u> Uno espera que los puntajes del Pretest no muestren diferencias significativas. En el caso de los grupos experimental y de control asignados al azar (Diseño 1), si no hay diferencia significativa en los puntajes del pretest significa que, como uno esperaría, el azar produjo dos grupos con puntajes en el pretest más o menos equivalentes. La posibilidad de que esta equivalencia ocurra es mayor, por supuesto, si alguna forma de agrupación en bloques o estratificación se ha usado con anterioridad a la asignación al azar (ver en la página 127 una discusión sobre agrupación en bloques).

Si el test-t para la significación de la diferencia entre las medias aritméticas de los dos pretest no es significativa, no es necesario dar una atención posterior al pretest. La interpretación puede descansar en las medias aritméticas de los postest.

Si, por otro lado, las medias aritméticas de los pretest son significativamente diferentes, usted tiene que enfrentarse con este problema. Hay varia posibilidades:

1. Quizás no es demasiado tarde para hacer una nueva asignación a los grupos. Si usted no hizo bloques sobre la base de los puntajes del pretest al formar los grupos

inicialmente, hágalo así esta vez; es más probable que los grupos sean equivalentes cuando se forman de esta manera.

2. Trate este diseño como un diseño de grupo control no equivalente, Diseño 3.

3. Si otra persona, no usted, hizo la asignación al azar, verifique con esa persona para ver si unos pocos "casos espaciales" no fueron, en realidad, asignados al azar y corrija esto si es posible o elimine aquellos casos del análisis.

Comparación de los puntajes del Postest. Si la media aritmética de los puntajes del "postest" del grupo experimental fue más alta que la media aritmética del grupo control y el "test-t" fue estadísticamente significativo, usted puede establecer, para el Diseño 1: "Los resultados mostraron que el programa experimental produjo puntajes en el "postest" que fueron significativamente más altos que aquellos del grupo control". Sin embargo, es quizás mejor evitar los términos "experimental" y "control" y, en cambio es preferible describir los grupos. Aquí, por ejemplo, los grupos forman alumnos que participan en un laboratorio complementario de biología y alumnos del curso regular.

Ejemplo. El ejemplo de tutoría

"Los resultados mostraron que la participación en el programa del laboratorio de biología produjo puntajes en el postest que fueron significativamente más altos que aquellos de los alumnos permanecieron en cursos regulares".

En lugar de un test-t sobre los puntajes del postest, la presentación de *límites de confianza* es a veces más significativa y ayuda a interpretar la significación educacional. Los límites de confianza indican cuán grande o pequeña podría ser la verdadera diferencia entre los grupos E 7 C.

Si la media aritmética de los puntajes del postest del grupo experimental fue más alta que la media aritmética del grupo control, pero no fue significativamente diferente de acuerdo al test-t, usted podría hacer que un analista de información ensayara un test estadístico más poderoso, tal como el análisis de varianza en que el pretest se usa como un factor, siendo los niveles los bloques que usted usó al aleatorizar. (El uso del análisis de varianza se describe en el Capítulo 6). Al emplear este análisis más poderoso, usted podría detectar diferencias significativas en el programa que el test-t dejó de detectar. (El concepto del poder de un test para detectar diferencias se describió en el Capítulo 1).

Por otra parte, usted podría simplemente observar que "La media aritmética de los puntajes del postest de los alumnos en el programa X fue más alta que la media aritmética para los alumnos en el programa C, pero la diferencia pudo haberse debido a variación al azar, es decir, la diferencia no fue suficientemente grande para llamarse estadísticamente significativa".

Si la media aritmética de los puntajes del postest del grupo experimental fue más baja que la del grupo control, el programa experimental no ha producido mejores resultados. Usted necesita verificar si el puntaje del grupo-E fue significativamente más bajo que el puntaje del grupo-C, porque si fue así, esto quiere decir que el programa experimental tuvo un efecto negativo en la medición de resultado.

Suponga que usted encuentra que el grupo experimental tuvo una media aritmética significativamente más alta que el grupo control. Usted aún necesitará decidir sobre la significación educacional de las ganancias derivadas del programa. ¿Cuán valiosa es esta diferencia? A continuación del resumen del análisis del Diseño 1, se examina la significación de los efectos del programa.

Resumen del análisis del Diseño 1

Al informar sobre los resultados de un diseño de un grupo control verdadero, pretest – postest, analice primero el grado en que el diseño se implementó adecuadamente, la pérdida de casos y cualesquier confusión y contaminación posible. Si ninguna de estas amenazas o cualesquier otra que a usted se la pueda ocurrir, parece presentar problemas insalvables para la interpretación de los resultados, entonces presente los resultados. ¿Eran los grupos equivalentes en el pretest? y, si fue así, ¿hubo alguna diferencia estadísticamente significativa entre las medias aritméticas de los puntajes del postest? Si se encuentran diferencias estadísticamente significativas, usted necesitará entonces examinar la significación educacional de los resultados.

LA SIGNIFICACIÓN EDUCACIONAL DE LOS EFECTOS DEL PROGRAMA

La significación estadística solamente le dice si usted está o no está errado al extraer ciertas conclusiones sobre un experimento. Si usted puede decir que en el postest el grupo experimental obtuvo mejores resultados que el grupo control a un nivel de .05 de significación, esto significa que si en realidad no hubo una diferencia verdadera entre los dos grupos después del programa, la diferencia que usted encontró ocurrirá solamente cinco en cien veces. Esta pequeña posibilidad lo hace sentirse seguro al rechazar la idea de que no podría haber diferencia real entre los grupos (la "hipótesis nula"). La significación estadística le asegura que el resultado no se debió a variación al azar, la diferencia que usted encontró fue demasiado grande para deberse al azar.

Sin embargo, la significación estadística no le dice nada sobre cuán *valiosa* fue la diferencia. La significación estadística depende en gran medida del número de alumnos o cursos en sus grupos. Un programa que enseñó a un gran número de alumnos del grupo-E, digamos 300, a responder correctamente dos ítemes más que los que el grupo control pudiera responder, podría mostrar una superioridad estadísticamente significativa. Pero, ¿es educacionalmente significativo aprender dos hechos más en el curso de un año? Y si el programa costó más, ¿valía la pena?

Obviamente, se necesita que haya alguna forma para determinar cuán valioso o significativo, desde el punto de vista educacional, son los "efectos" de un programa. La palabra "efecto" se usará como ligeramente equivalente a la diferencia en medias aritméticas entre los grupos que recibieron o que no recibieron el programa. De este modo, si los resultados en el postest fueran:

media aritmética para el grupo-E = 30

media aritmética para el grupo-C = 26

el "efecto" del programa fue 4 puntos a favor del grupo-E.

Las siguientes son diversas formas para intentar medir la significación educacional de los efectos de un programa. Dicha medición no es fácil. Elija para su propio uso cualesquiera de los métodos descritos a continuación, que le parezca razonable. La mayor parte de la discusión supondrá que los resultados del postest en consideración fueron tests de rendimiento.

1. Juicio de los profesores como un criterio de referencia. Si el test fue un test estandarizado o un test construido en la casa, haga que unos pocos profesores que conocen el nivel del curso determinado y el área de contenido estudien el test en detalle y antes que hayan visto los resultados del test marquen lo siguiente:

a) los ítemes que nadie debería haber errado
b) ítemes que se esperaría que el alumno promedio responda correctamente en el tiempo que se asignó
c) ítemes que se esperaba que sólo los mejores alumnos contestaron correctamente.

Derivado de esta información se puede generar una gama de puntajes que representa los juicios de los profesores sobre los contenidos del test y predicciones sobre un logro razonable. Los resultados obtenidos pueden compararse con estos juicios. Pregúntese ¿qué ítemes determinaron las diferencias entre grupos? ¿los fáciles? ¿o pareció que cualesquiera de los grupos trabajó mejor con el material más difícil?

2. Habilidad como referencia. Compare los efectos del programa en alumnos de diferente habilidad. En el postest ¿cuál fue el rango en puntajes entre los alumnos más lentos en el nivel o curso y los alumnos más inteligentes en el nivel o curso? ¿Amplio o estrecho? ¿Parecía el grupo beneficiarse en forma especial? ¿Cómo es el efecto del programa comparado con los efectos de habilidad?

3. Norma como referencia. Si usted usó un test estandarizado, compare los efectos del programa con normas nacionales, ¿las diferencias de puntaje en la escala, por ejemplo, representaron una diferencia de puntajes en el curso equivalente a medio mes o a seis meses?

4. Ítem análisis. Para cada ítem determine el porcentaje del grupo-E y del grupo-C que respondió correctamente el ítem. Trate de ubicar los ítemes que establecieron las principales diferencias. Quizás la diferencia significativa se debió principalmente a ítemes que trataban sólo sobre un área de contenido y de otra forma no hubiera habido diferencia entre los grupos. Entonces usted tiene que preguntar a la gente cuán importante fue el área de contenido.

Por otra parte, si el programa experimental muestra una mayor aprobación en casi todos los ítemes, pareciera que el programa experimental es el programa de mayor éxito en muchos tipos de contenidos.

DISEÑO 2

EL DISEÑO DE GRUPO CONTROL VERDADERO, POSTEST SOLAMENTE

DISEÑO 2

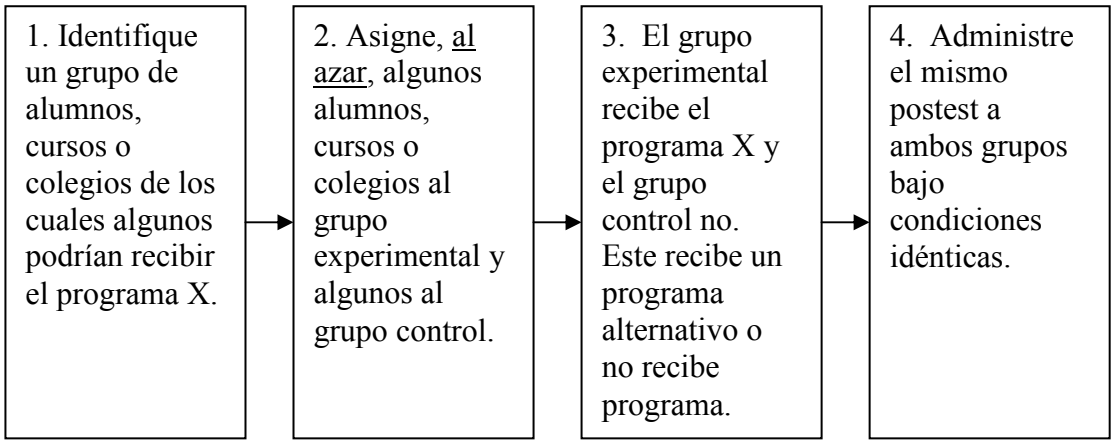
EL DISEÑO DE GRUPO CONTROL VERDADERO, POSTEST SOLAMENTE

DIAGRAMA

	Tiempo		
	1 (pre)		2 (post)
Grupo Experimental	R	X	0
Grupo Control	R		0

Resumen: Se forman aleatoriamente dos grupos, uno recibe el programa X y el otro no.

PASOS ESENCIALES AL IMPLEMENTAR EL DISEÑO 2



Ejemplo 1.

Algunos especialistas en lectura de un distrito querían encontrar un método para introducir lectura en el primer grado, lo que evitaría problemas posteriores. Ellos no estaban seguros si un método fónico intensivo o un método intensivo basado en experiencias sería mejor para los alumnos (y los padres) en este distrito específico. Ellos decidieron ensayar ambos métodos y ver cuál funcionaba mejor. Reunieron una serie de talleres para cada método y aleatoriamente eligieron 10 profesores de primer grado para cada uno de estos talleres. A los profesores se les pidió que usaran extensivamente los métodos y materiales de los talleres en sus cursos y se hicieron visitas regulares para ver si esto se estaba cumpliendo. Al final del año se administró un postest de lectura en todos los cursos.

Ejemplo 2.

Un equipo de profesores de un colegio secundario planificó un programa de relaciones humanas que se componía de películas, discusiones, libros, "role-playing" y explicaciones de los valores y características de diferentes grupos culturales.

Ellos asignaron al azar a los niños al "grupo X" o "grupo asignaron al azar a los niños al "grupo X" o "grupo ", y luego hicieron lo mismo con las niñas. El grupo "X" recibió el programa de relaciones humanas durante 3 semanas, mientras que el grupo "Y" recibió un programa nuevo sobre carreras, un programa que también usó películas, discusiones, libros y "role-playing". Al final de las tres semanas ambos grupos contestaron un cuestionario que contenía preguntas sobre relaciones humanas y sobre carreras. Los resultados del grupo "X" sobre las preguntas de relaciones humanas se compararon con los resultados del grupo "Y", sobre las preguntas de relaciones humanas. Las diferencias se atribuyeron al programa puesto que se podía suponer que los grupos habían sido equivalentes al comenzar, debido a la aleatorización.

PRESENTACIÓN Y ANÁLISIS DE LA INFORMACIÓN PARA EL DISEÑO 2

Hay dos tipos básicos de información para el Diseño 2. El primer tipo de información que se necesita es aquella información que muestra con cuánto éxito se implementó el programa. En segundo lugar, debe informarse sobre los resultados de las medidas de producto para los grupos E y C. A continuación, bajo los títulos: "Informe sobre la implementación" e "Informe sobre resultados" se describen los métodos de informe de estos dos tipos de información.

Informe sobre la implementación del Diseño

Mostrar el éxito con que se implementó el diseño involucrará considerar las siguientes preguntas:

1. Implementación del Programa. ¿Recibió el programa experimental el grupo-E? Esta pregunta la responderá la documentación.

2. Contaminación. ¿Se quedó el grupo-C sin recibir el programa?, es decir, ¿se evitó la contaminación? Esto requiere al menos alguna documentación de lo que pasó en los cursos del grupo-C[*].

3. Confusión. ¿Hubo diferencias consistentes entre lo que sucedió al grupo-E y al grupo-C , sin considerar el programa X, diferencias que con razón podría esperarse que subieran o bajaran los puntajes en las mediciones de resultado? Es decir, ¿hubo confusión o factores ajenos que confunden y que tengan que tomarse en cuenta cuando se interpretan los resultados?

4. Atrición (pérdida de casos). ¿Hubo alguna diferencia entre el número de "casos" que se perdió en el grupo-E y el número de casos perdidos en el grupo-C? ¿Hubo alguna diferencia en los tipos de casos (alumnos, cursos o colegios) que tuvieron que eliminarse del análisis en cada grupo? Por ejemplo, ¿perdieron demasiado del programa los miembros del equipo de "Football" porque hicieron viajes para participar en partidos de "Football"? o, si se encontró que el programa fónico, en el ejemplo 1, no se implementó en tres o cuatro cursos, pero el programa de la experiencia se implementó en todos los cursos, esto necesitará informarse a interpretarse. Una tabla, como la que sigue a continuación, mostrando la pérdida de casos en varias etapas de la implementación podría incluirse en su informe. Una tabla alternativa que muestra pérdida de alumnos por varias razones aparece en la página 50.

[*] Contaminación, confusión y atrición se discutieron en el Capítulo 2 bajo el Título: "Principales Amenazas a la Implementación de los Diseños", páginas 44 y 45.

	Número de profesores a los que se les pidió que atendieran el taller	Número de profesores que completaron el taller	Número de profesores que implementaron los programas satisfactoriamente
Programa fónica			
Programa basado en experiencias			

Interpretación de tablas que muestran pérdida de casos. Examine la pérdida de "casos" (alumnos o profesores o cursos –como quiera que se compongan los grupos) en términos de causas probables. Si un grupo mostró una tasa de mortalidad experimental mayor que el otro, es importante discutir este punto. A menudo se necesitará un sondeo posterior con el fin de descubrir lo que sucedió. Seleccionar una muestra al azar de alumnos, cursos o colegios que se retiraron, ya sea del grupo-E o del grupo-C y entrevistarlos podría explicar lo que causó atrición.

Informe de resultados

Los ejemplos que se dan a continuación muestran una forma de presentación de los puntajes del postest a través de una tabla y de un gráfico. Un gráfico es una forma útil de presentar los resultados cuando la claridad es más importante que la exactitud de un número decimal. El "test-t" en la tabla puede ser uno de los dos tipos que existen. Si la aleatorización* fue simple o en bloques, use el "test-t" para grupos no pareados (también se llama test-t no correlacionado). Si se usó un método de asignación al azar pareado, use un "test-t" para grupos pareados. Si el valor t es significativo, se indica generalmente por medio de un asterisco y la nota al pie de la tabla "p< .05" (o si usted usó un nivel .10, entonces "p< .10").

Ejemplo 1.

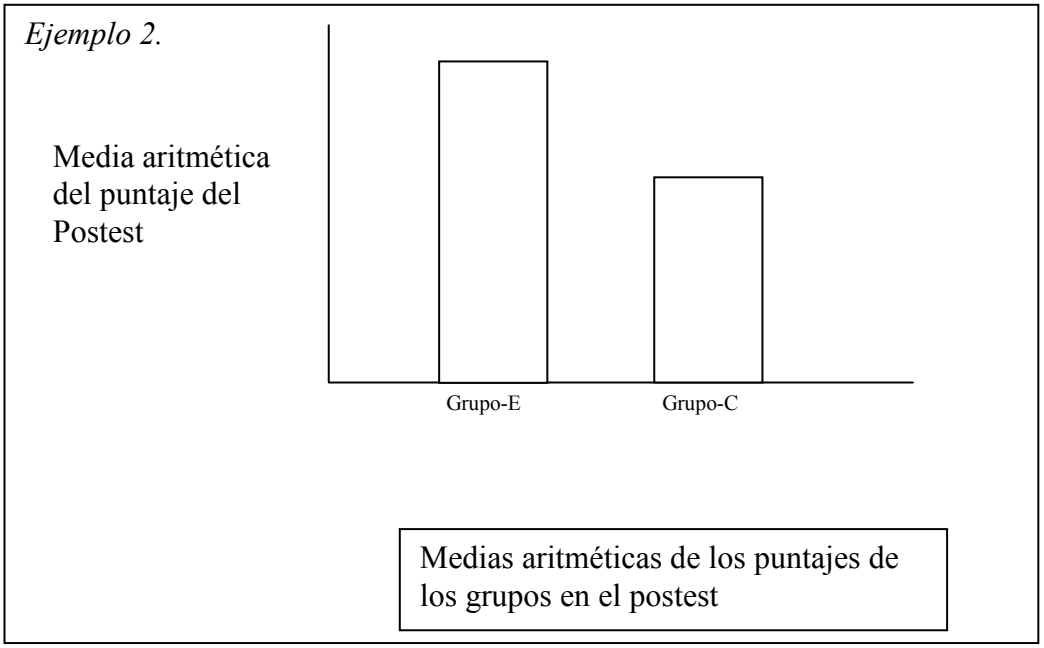
TABLA 3
Media aritmética de los puntajes del Postest

	Postest			
	N	Media	SD	Test-t
	-	--	--	
Grupo-E	-	--	--	--*
Grupo-C				

(* p < --)

* Los métodos para aleatorizar se discuten en las páginas 119-126.

Ejemplo 2.

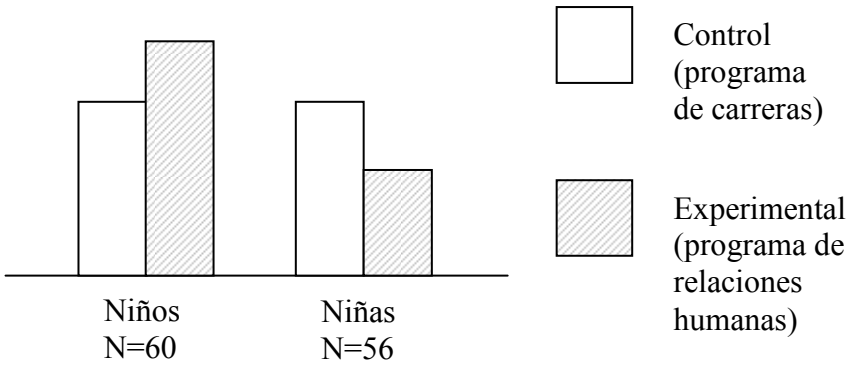


Interpretaciones. Si hay una diferencia significativa entre las medias aritméticas de los puntajes, como se muestra por medio de un "test-t" significativo, usted puede estar muy tranquilo al decir que un programa parece ser más efectivo que otro mientras no haya habido ningún problema importante con pérdida de casos. Usted enfatizará que el procedimiento al azar, que habrá descrito en la sección de su informe sobre la metodología de la evaluación, debe haber asegurado grupos equivalentes al comenzar. Por supuesto, la interpretación debe considerar las condiciones bajo las cuales funcionaron el programa X y el programa control (o alternativo). Si hubiera diferencias en lo que sucedió a los grupos, aparte del programa, éstas debieran mencionarse y discutirse en términos de cuán probable sería que estas diferencias o confusiones fueran a influenciar los resultados.

Usted también necesitará discutir el grado en que los dos programas se mantuvieron separados, es decir, usted necesita considerar el problema de contaminación.

Si las medias aritméticas del postest no fueran significativamente diferentes y usted no cree que esto se debió a contaminación –a la posibilidad de que los programas mismos no se mantuvieron separados—usted podría intentar algún análisis "interno" de la información. Quizás usted sospecha que el programa funcionó mejor con algunos alumnos que con toros. En el ejemplo anterior del programa de actitud (el programa sobre relaciones humanas) pudiera ser que usted quisiera observar la media aritmética de los puntajes de los alumnos de sexo masculino y femenino en forma separada.

Ejemplo 3. Puntajes del Postest sobre actitudes.



Para los niños, el programa experimental parece haber aumentado las actitudes positivas; para las niñas, disminuido. Cuando estos resultados se promedian no muestran diferencia.

Hay una advertencia importante. Cuando usted empieza a dividir los grupos en subgrupos más pequeños, los resultados llegan a ser menos estables debido a los números más pequeños. Los test de significancia verificarán cuán seguro está usted al generalizar partiendo de estos grupos. Además, si usted examina bastante la información –la separa en muchas formas– usted aumenta enormemente sus posibilidades de encontrar algo "significativo" sólo por casualidad. La mejor sugerencia es dividir la información para un análisis interno sólo si usted tiene una buena razón para hipotetizar algunos efectos diferenciales.

Resumen del análisis del Diseño 2

Cuando informe sobre los resultados de un diseño de grupo control verdadero, postest solamente, analice primero el grado en que el diseño se implementó adecuadamente. Debido a que los puntajes del pretest no se presentan puede ser que usted estime que es necesario enfatizar el procedimiento de aleatorización que se usó para formar los grupos. Una descripción detallada del procedimiento puede ayudar a convencer a la audiencia que los grupos se formaron de una manera objetiva. Usted también analizará cualesquier confusión, contaminación y la pérdida de casos que ocurrieron en los grupos E y C.

Si el diseño se implementó adecuadamente, usted presentará los resultados del postest y examinará si la diferencia entre los grupos E y C fue o no fue estadísticamente significativa. Si fue estadísticamente significativa, usted considerará la significación educacional de la diferencia. Los problemas relacionados con la estimación e informe de la significación educacional se discuten en la página 56.

LA SIGNIFICACIÓN EDUCACIONAL DE LOS EFECTOS DEL PROGRAMA

La significación estadística le dice solamente que no es probable que sus resultados hayan ocurrido por casualidad. No le dice si sus resultados son de gran magnitud, o si son importantes. Por favor vaya a la página 56 donde encontrará sugerencias sobre la medición de la significancia de los efectos del programa en términos de su valor educacional.

DISEÑO 3

EL DISEÑO DE GRUPO CONTROL NO-EQUIVALENTE, PRETEST-POSTEST

DISEÑO 3

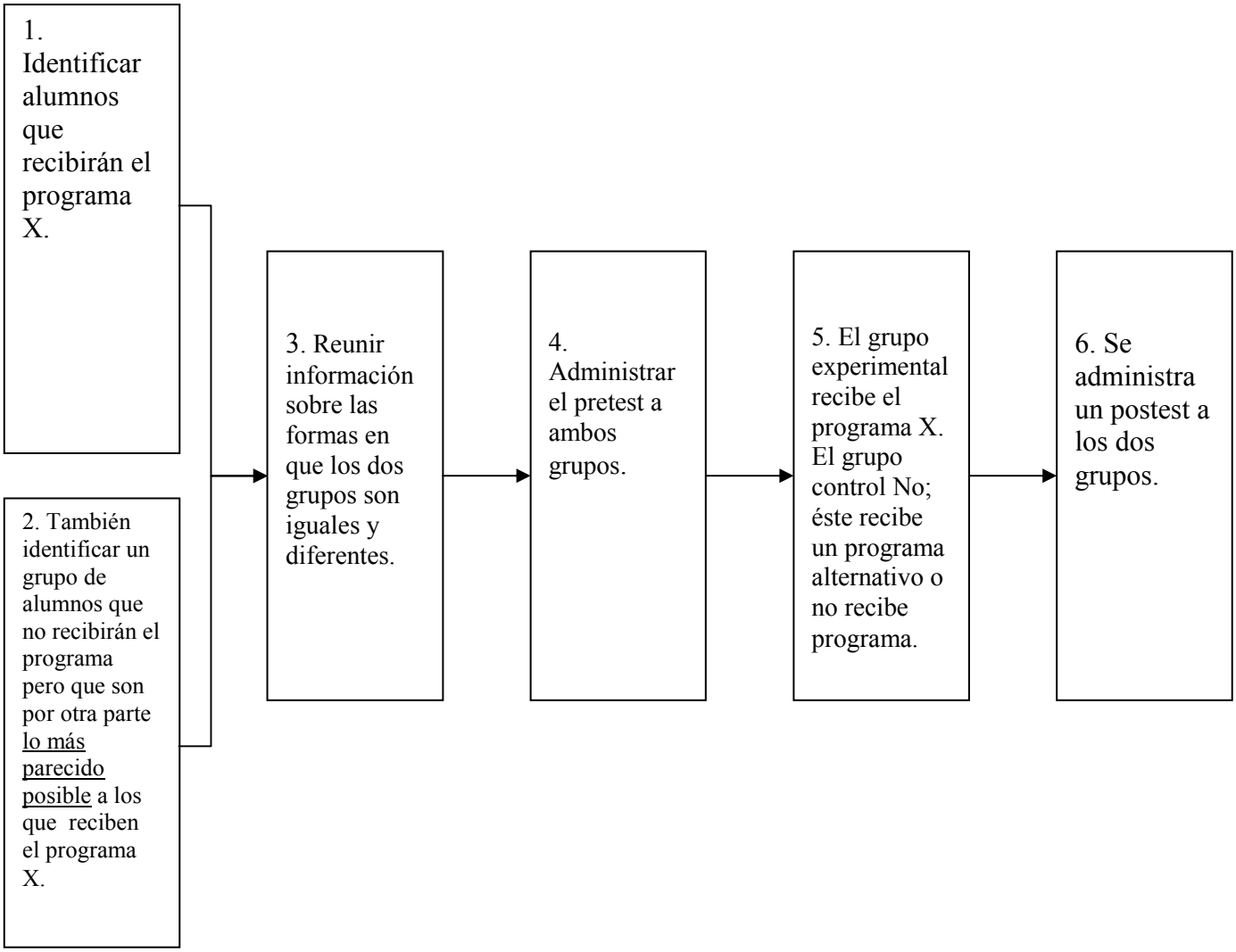
EL DISEÑO DE GRUPO CONTROL NO – EQUIVALENTE, PRETEST-POSTEST

DIAGRAMA

	Tiempo		
	1 (pre)		2 (post)
Grupo Experimental	0	X	0
Grupo Control	0		0

Resumen: Dos grupos que son similares, pero que no se formaron aleatoriamente, se miden tanto antes como después que uno de los grupos reciba el programa.

LOS PASOS ESENCIALES AL IMPLEMENTAR EL DISEÑO 3



EJEMPLOS

Ejemplo 1.

Un distrito desarrolló un programa de recuperación en matemáticas para alumnos del primer año de enseñanza media que tenían bajo rendimiento en destrezas básicas. El programa sería caro debido a los nuevos materiales y a los tutores, y se decidió ensayar el programa en un colegio antes de expandirlo a otros colegios. Puesto que era un programa de nivelación destinado a destrezas, se administró un test de destrezas básicas en septiembre y el 5% más bajo de los alumnos se incorporó al programa de recuperación. Consciente de que tendría que eliminar efectos de regresión como causa de puntajes más altos, el coordinador del distrito previo la necesidad de un grupo de comparación. Él administró el mismo test en dos colegios vecinos y en cada colegio identificó los alumnos que estaban bajo el puntaje límite que se había usado en el colegio original. Al final del año, a los alumnos de los tres colegios se les administró el mismo test como postest. Comparando la media aritmética del grupo experimental con las medias aritméticas de los colegios control, era posible ver si las ganancias en el programa experimental eran mayores que aquellas en los programas de los otros colegios. Puesto que los efectos de regresión y los efectos de los alumnos que están creciendo serían los mismos en todos los grupos, las diferencias pudieran interpretarse como que probablemente se debían a los programas.

Ejemplo 2.

Quince cursos en un distrito habían recibido fondos para programas de educación bilingüe. El evaluador estaba muy interesado en medir el grado en que los programas bilingües mejoraban el rendimiento, estimulaban la amistad entre los grupos étnicos y mejoraban la adaptación de los niños al colegio. Él se dio cuenta que mejoraría sus posibilidades de poder demostrar estos efectos si tenía algún tipo de grupo control.

No pudiendo formar un grupo control verdadero, él buscó un grupo control no-equivalente y pudo obtener la cooperación de nueve cursos en un distrito cercano. Estos cursos también tenían altos porcentajes de alumnos que no hablaban inglés, pero no tenían programas bilingües. El administró un pretest a los 15 cursos del grupo-E y a los nueve cursos del grupo-C, observó muestras de los cursos a intervalos regulares durante el año escolar, y administró un postest a todos los cursos a fin de año. Para su análisis comparó la media aritmética de los 15 puntajes del grupo-E con la media aritmética de los 9 puntajes del grupo control no equivalente, usando un “test-t” para medias aritméticas no correlacionadas.

PRESENTACIÓN Y ANÁLISIS DE LA INFORMACIÓN PARA EL DISEÑO 3

El Diseño 3 es tal como el Diseño 1, excepto que los grupos no se forman aleatoriamente. Sin embargo, el pretest sirve como una forma de averiguar si los grupos son al menos comparables en cualesquier aspecto que este mida.

Puesto que el pretest se usa de esta manera, debería ser una medición lo más estrechamente relacionada posible al postest, de preferencia el mismo test o una forma equivalente o paralela de él.

Hay dos tipos básicos de información que deben darse con el fin de usar adecuadamente el Diseño 3. Primero, debe haber información que muestre con qué éxito se implementó el diseño. Segundo, debe informarse sobre los resultados de las mediciones de producto, tanto del grupo-E como del grupo-C y hacerse las comparaciones adecuadas.

Informe sobre la implementación del Diseño

Mostrar con cuánto éxito se implementó el diseño implicará considerar las preguntas siguientes:

1. Implementación del programa. ¿Recibió el programa el grupo-E? Esta pregunta la responderá la documentación del programa.

2. Contaminación. ¿No recibió el grupo-C programa? Por ejemplo, ¿se evitó la contaminación? Esto requiere, al menos, alguna documentación de la que sucedió en los cursos del grupo-C. [*]

3. Confusión. ¿Hubo diferencias consistentes entre lo que sucedió al grupo-E y al grupo-C, sin considerar el programa X, diferencias que con razón podría esperarse que subieran o bajaran los puntajes en las mediciones de resultados? Es decir, ¿hubo confusión o factores ajenos que confunden y que tengan que tomarse en cuenta cuando se interpretan los resultados?

4. Atrición (pérdida de casos). ¿Hubo alguna diferencia entre el número de "casos" perdidos en el grupo-E y el número de casos perdidos en el grupo-C? ¿Hubo alguna diferencia en los tipos de casos (alumnos, cursos o colegios) que tuvieron que excluirse del análisis en cada grupo? Una tabla como la que se muestra a continuación entrega un resumen adecuado y puede ser incluida en su informe.

[*] Contaminación, confusión y atrición se discutieron en el Capítulo 2 bajo el título "Principales Amenazas a la Implementación del Diseño".

NÚMERO DE ALUMNOS EXCLUIDOS DEL ANÁLISIS
POR DIVERSAS RAZONES

Razón	Número excluido del grupo-E	Número excluido del grupo-C
Ausente del postest		
Ausente del colegio durante el programa		
Trasladado del grupo a petición de los padres		
Abandonó el colegio		
Otras razones		
Número total de alumnos excluidos		

Interpretación de la tabla que muestra pérdida de casos. Si las tasas de deserción en los grupos E y C son muy diferentes, trate de explicarlas. Esto puede requerir que se hagan averiguaciones, tratando de descubrir lo que estaba sucediendo. Si usted sospecha que los alumnos con ciertas características abandonaron el programa, usted podría tubular información adecuada y analizar esta información

Observe que si la pérdida de casos es similar en el grupo experimental y en el grupo control no equivalente, esto es una evidencia más de similitud y por lo tanto de comparabilidad de los dos grupos.

Informe de resultados

La tabla que se presenta a continuación constituye un ejemplo de una forma de presentar resultados. Si la aleatorización* fue simple o "en bloques" se usa un "test-t" para grupos no pareados (a veces llamado "test-t" no correlacionado). Si se usó un método de aleatorización pareado, use un "test-t" para grupos pareados).

* Los métodos de aleatorización se analizan en las páginas 119-126

Ejemplos.

TABLA 4

Resultados del Pretest y Postest de los grupos Experimental y Control.

	N	Pretest			Postest		
		X	SD	test-t	X	SD	test-t
Grupo-E	32	60	10	.90	90	9	5.1*
Grupo-C	35	58	8		80	7	

* = estadísticamente significativo a nivel .05

Si no hay diferencia significativa en los puntajes del pretest, es una gran suerte. Al menos el grupo control no equivalente fue similar al grupo-E en ese puntaje. Por supuesto, los grupos podrían ser diferentes en algunos aspectos que no se miden con el pretest, pero al menos usted partió con grupos que eran similares entre sí en un factor que tendrá un efecto importante en el postest. Usted puede ser especialmente firme en esta afirmación si elige como pretest una medida que probablemente correlaciona poderosamente con el postest, una elección que es, por supuesto, la recomendada.

Si los puntajes del postest son significativamente diferentes y si los puntajes del pretest no habían sido significativamente diferentes, hay una buena evidencia de que los programas diferentes estaban produciendo resultados diferentes. Se podría informar algo como lo que sigue:

"El programa experimental parece haber producido puntajes en el postest que eran significativamente más altos que aquellos del grupo control. Existe, por supuesto la posibilidad de que hubiera diferencias iniciales entre los dos grupos (aunque eran similares en el pretest) y que estas diferencias explicaran algunas de las diferencias en el postest".

Y luego entregar el resto de la información que tienda a fortalecer o debilitar el argumento sobre la comparabilidad inicial de los grupos.

Finalmente, usted necesitará considerar cuán educacionalmente significativos son los resultados. Lea el breve análisis al final de este Capítulo.

Si los puntajes del pretest fueran más o menos equivalentes y los puntajes del postest no fueran significativamente diferentes, usted podría quizás concluir que no hubo efectos derivados de los programas diferentes. Sin embargo, suponga que los resultados parecían promisorios aunque ellos no resultaron ser significativos, entonces puede ser que usted quisiera que un analista ensayara un análisis más poderoso. Esta

persona podría usar análisis de covarianza* o análisis de varianza usando el pretest como un factor de bloqueo.

Si hay una diferencia significativa en los puntajes del pretest, usted tiene un problema. Una forma del abordarlo es anotar el problema, pero proceder de todas maneras. Grafique los resultados como una de las figuras que se van a sugerir más adelante y examine las tendencias. Haga una observación en el sentido de que cualesquier tendencia aparente podría deberse a diferencias iniciales entre los grupos y la evaluación debería repetirse, si es posible, con un grupo control verdadero. Observe que a pesar de la no equivalencia del grupo control, proporciona alguna información razonablemente comparable.

Otra forma de tratar las diferencias significativas del pretest es usar análisis de covarianza los resultados. Sin embargo, esto requiere una cantidad considerable de computación y deberían explicarse algunos supuestos más bien rigurosos. Probablemente vale la pena hacerlo si usted puede trabajar fácilmente en un computador.

Si usted no puede hacer un análisis de covarianza, hay un par de procedimientos de muestra que puede aplicar cuando se encuentra con el problema de puntajes de pretest estadísticamente significativos. Aunque cada procedimiento tiene algunos problemas asociados con su interpretación, como asunto práctico puede ser mejor ensayar estos procedimientos más bien que perder la esperanza de tratar de interpretar lo resultados.

Aquí están los dos procedimientos simples, susceptibles de ser adoptados en la situación en que las medias aritméticas de los puntajes del pretest muestran diferencias significativas. Se recomienda que usted use los dos para ver si los resultados concuerdan entre sí.

Pareo "post-hoc". Obtenga el grupo completo de puntajes de pretest para cada grupo, el grupo-E y el grupo-C (el grupo-C es el grupo control no-equivalente). Para cada caso del grupo-E, trate de encontrar un caso en el grupo-C con el mismo o casi el mismo puntaje de pretest. Haga una lista de estos pares pareados, registrando los puntajes del postest en la lista. Cuando usted tenga la lista más larga posible, habiendo pareado los puntajes del pretest dejando unos pocos puntos de diferencia entre unos y otros, aplique un "test-t" para grupos pareados a los puntajes del postest.

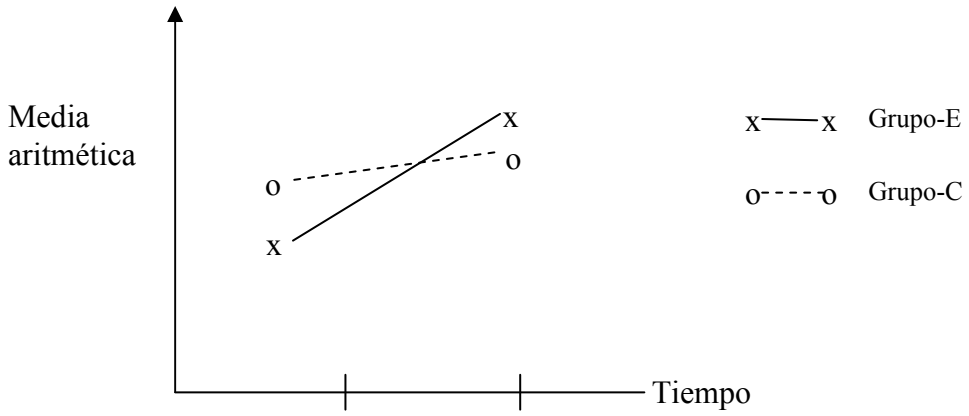
Análisis de puntajes de ganancia. Para considerar las diferencias de los pretest, en una forma ligera pero simple, registre los puntajes ganados por cada estudiante (o "caso"). La ganancia es el puntaje del postest menos el puntaje del pretest, generalmente suponiendo que tanto el pretest como el postest fueron los mismos tests o fueron formas de contenido equivalente. Luego trate los puntajes ganados como un conjunto de puntajes corrientes y aplique el "test-t" para grupos no pareados.

* El análisis de covarianza se usa para ajustar los puntajes del postest a lo que estos habrían sido si los puntajes del pretest hubieran sido equivalentes. Es un ajuste estadístico y no un procedimiento totalmente creíble.

Grafique los resultados

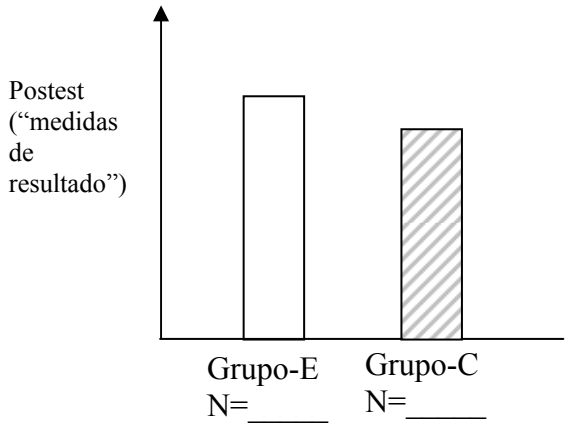
Los gráficos se entienden en forma más rápida que las tablas y se deberían usar casi siempre. Probablemente el mejor gráfico para el Diseño 3 es uno como el que sigue, el que se prepara fácilmente si el pretest fue el mismo test o una forma paralela o equivalente del postest.

MEDIAS ARITMÉTICAS DEL GRUPO-E Y DEL GRUPO-C
EN EL PRETEST Y EN EL POSTEST



Si usted quisiera el medio más claro posible para presentar los resultados, usted podría simplemente mostrar las medias aritméticas del postest como gráfico de barras:

MEDIAS ARITMÉTICAS DE LOS PUNTAJES DEL POSTEST
DE LOS GRUPOS EXPERIMENTAL Y CONTROL



Sin embargo, esto se justifica solamente si usted ya ha mostrado que los puntajes del pretest no eran significativamente diferentes.

Resumen del análisis del Diseño 3

Cuando informe sobre los resultados de un diseño de grupo control no equivalente, discuta primero el grado en que el diseño se implementó adecuadamente: ¿Se controlaron las confusiones y la contaminación ¿cuál fue el grado de atrición?

Debido a que los grupos no se asignaron al azar, usted necesitará presentar evidencia de que los grupos fueron inicialmente iguales en características que podrían haber afectado sus resultados. Una fuente principal de esta evidencia será los puntajes del pretest, pero usted podría presentar otras características de los grupos también, tales como estatus socioeconómico, niveles de habilidad, etc.

Si el diseño se implementó adecuadamente usted presentará los resultados y analizará si la diferencia entre los grupos E y C fue o no fue estadísticamente significativa. Si fue estadísticamente significativa, usted considerará la significación educacional de la diferencia. Los problemas relacionados con la estimación e informe de la significación educacional se discuten en la página 56.

LA SIGNIFICACIÓN EDUCACIONAL DE LOS EFECTOS DEL PROGRAMA

La significación estadística le dice solamente que no es probable que sus resultados hayan ocurrido por casualidad. No le dice si sus resultados son de gran magnitud, o si son importantes. Por favor vaya a la página 56 donde encontrará sugerencias sobre de la significancia de los efectos del programa en términos de su valor educacional.

CAPÍTULO 4

DISEÑOS 4 y 5

PROCEDIMIENTOS, ANÁLISIS, E INTERPRETACIÓN

En este capítulo se describe en detalle el Diseño 4. El Diseño 5 se describe brevemente puesto que el análisis del Diseño 4, en gran medida, sirve para el Diseño 5. Para introducir cada diseño hay un diagrama, un diagrama de flujo de los pasos esenciales y ejemplos.

DISEÑO 4

EL DISEÑO DE SERIES CRONOLÓGICAS

GRUPO-E SOLAMENTE

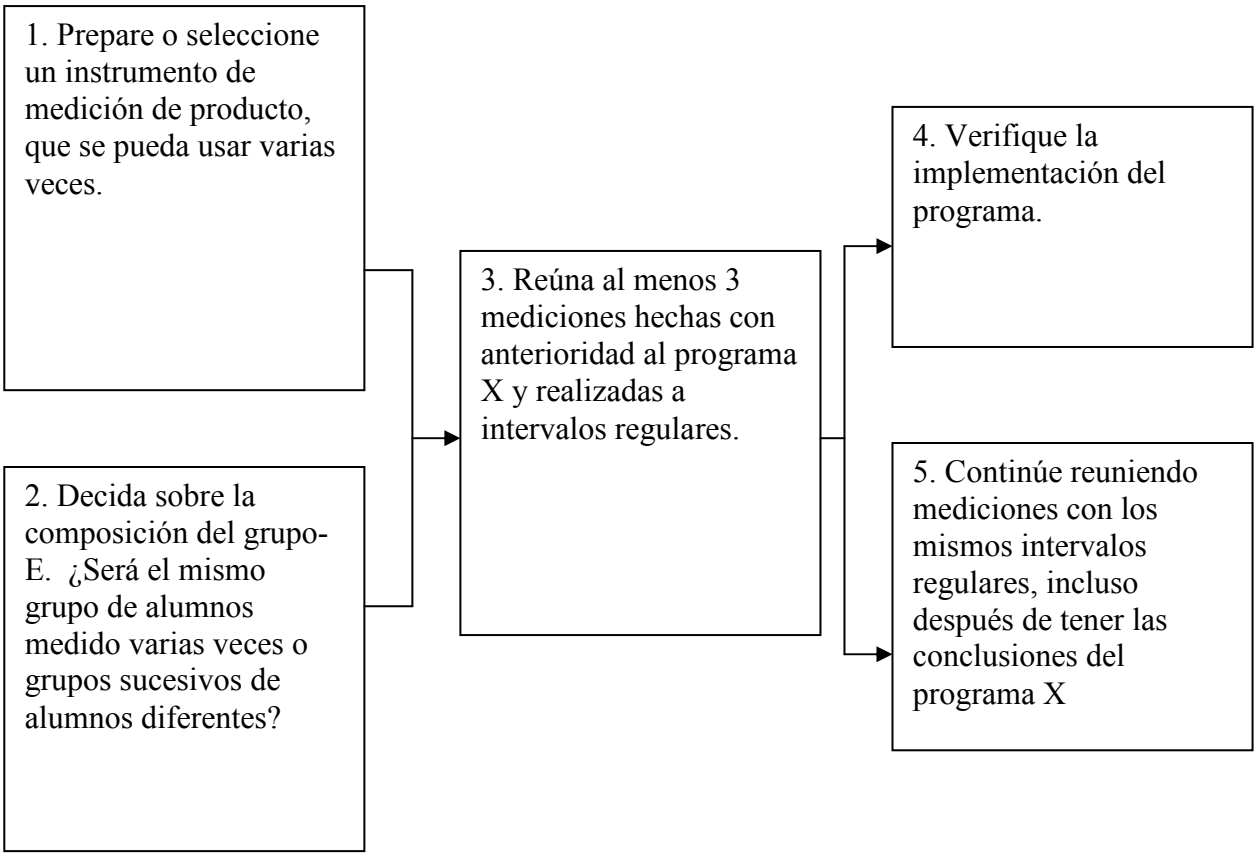
DISEÑO 4

EL DISEÑO DE SERIES CRONOLÓGICAS GRUPO-E SOLAMENTE

DIAGRAMA

	Tiempo								
	1	2	3	4		5	6	7	8
Grupo-E	0	0	0	0	X	0	0	0	0

PASOS ESENCIALES AL IMPLEMENTAR EL DISEÑO 4



EJEMPLOS

NOTA

Hay tres ejemplos para este Diseño. En cada ejemplo, el grupo que se mide es el grupo que recibe el programa, pero con estas diferencias:

En el ejemplo 1, los alumnos medidos cada vez son <u>todos</u> los alumnos del grupo-E.

En el ejemplo 2, los alumnos medidos en cada ocasión so alumnos del grupo-E seleccionados el azar.

} Estas son series cronológicas "longitudinales"*

En el ejemplo 3, se miden grupos sucesivos de alumnos y se consideran representativos de los alumnos que el colegio reciba cada año.

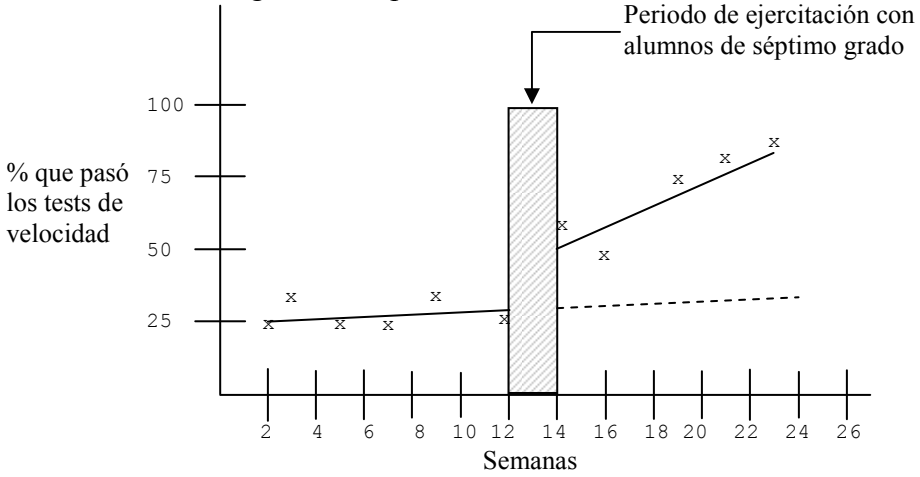
} Esta es un serie cronológica de "grupos sucesivos"

* Glass, Wilson y Gottman en "Design and Analysis of Time Series Experiments" (Laboratorio de Investigación Educacional, Universidad de Colorado, 1975) usan el término "unidad repetitiva" para lo que nosotros llamamos series cronológicas

"longitudinales" y "unidad replicativa" para lo que nosotros llamamos "grupos sucesivos".

Ejemplo 1. (Series cronológicas longitudinales: Grupo-E total Medido cada vez)

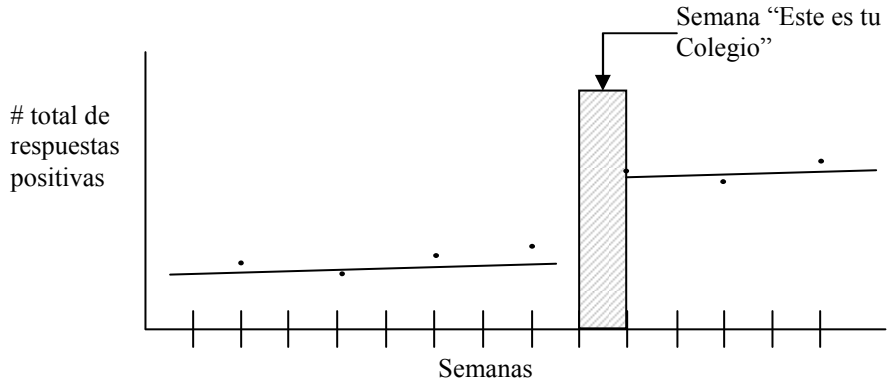
Un objetivo del programa de matemáticas de quinto grado requería que los alumnos fueran capaces de escribir cualesquier tabla de multiplicación en un periodo de tiempo breve y especificado. Cada profesor dio, regularmente, estos test de velocidad a su curso completo, pero muy pocos estaban alcanzando el objetivo. El supervisor de matemáticas sugirió una intervención (programa X). Trajo un grupo de alumnos de séptimo grado a un curso de quinto grado, diariamente y durante dos semanas para que hicieran un trabajo de tutoría. Cada alumno de séptimo grado pasó 15 minutos haciendo trabajar en ejercicios a un alumno, luego 15 minutos haciendo trabajar en ejercicios sobre las tablas de multiplicación a otro alumno. El profesor continuó dando los tests de velocidad, regularmente, después de esta intervención. El supervisor elaboró el gráfico siguiente, basándose en los registros del profesor:



Las líneas de tendencia indicaron que sin el programa un porcentaje mucho menor del curso habría pasado los tests de velocidad. Parecía que el programa no solamente aumentó el porcentaje de los que pasan el nivel de exigencia sino que también aumentó la tasa de mejoramiento. Quizás el programa había mostrado a los alumnos cómo trabajar en esta tarea. Al profesor se le estimuló para que ensayara esta intervención en otros cursos y así pudiera tener evidencias que presentar a la Junta Escolar en apoyo de su petición de transporte para los tutores del séptimo grado.

Ejemplo 2. (Series cronológicas longitudinales: Muestras al azar del Grupo-E que se mide).

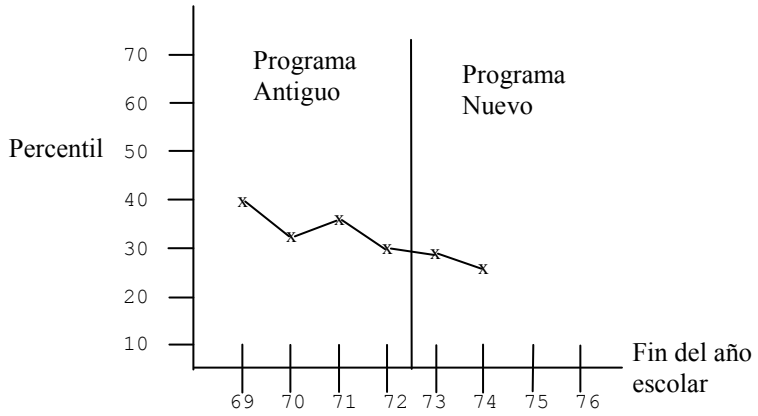
Un director estaba preocupado de aumentar en los niños el orgullo por su colegio y también la participación e interés de los padres por el colegio. Dicho director planificó tener una semana de "Este es tu Colegio", incluyendo exhibiciones de talento, festivales, reuniones sociales, etc. Más que confiar solamente en observaciones subjetivas para indicar el éxito de esta actividad, decidió a comienzos del año escolar controlar regularmente los sentimientos de los alumnos por su colegio. Elaboró un cuestionario que se diseñó para medir la actitud prevaleciente ("esta semana") hacia el colegio. Cada tres semanas, seleccionó al azar 8 alumnos de cada nivel y les administró el cuestionario en la cafetería del colegio. Cada vez sumó todas las respuestas positivas y registró el puntaje en un gráfico. Su gráfico resultó así:



Él quedó muy satisfecho al notar un cambio aparente en nivel. Siguiendo el programa X, el número de respuestas positivas fue generalmente más alto. El cambio en nivel promedio, de antes a después fue mayor que las fluctuaciones generales encontradas en el número de respuestas positivas. Él hizo una presentación a la Junta y surgió un interés considerable por sus actividades "Éste es tu Colegio".

Ejemplo 3. (Series cronológicas de grupos sucesivos)

Un nuevo programa "modular" de matemáticas se había implementado dos años atrás en el Colegio Secundario "Grant". En lugar de ser asignados a un curso de matemática, durante un semestre o año completo, los alumnos tomaron una serie de módulos en destrezas matemáticas durante seis semanas y se asignaron sobre la base de las destrezas que ellos necesitaban dominar. Ellos podían repetir un módulo tantas veces como fuera necesario para dominar la destreza. El programa fue caro debido a los gastos administrativos (tests extras, cambios de curso más frecuentes, etc.), profesores extra y algunos materiales nuevos. Los miembros de la Junta, buscando donde rebajar gastos, plantearon preguntas sobre la efectividad del programa sugiriendo que podría abandonarse. Al evaluador del distrito se le pidió que preparara un informe completo. Entre la información que presentó hubo puntajes de un test estandarizado de matemáticas. Afortunadamente, el mismo test se había administrado durante los últimos siete años, proporcionando información comparable. Él presentó el siguiente gráfico a la Junta:



La Junta no estuvo muy satisfecha, pero estuvo de acuerdo en continuar con el programa por un poco más de tiempo puesto que, se alegó, el programa necesitaba más tiempo para llegar a ser totalmente efectivo. Al menos, no habían ocurrido serias bajas después de la implementación del programa.

ANÁLISIS E INTERPRETACIÓN DE LOS RESULTADOS DE LAS SERIES CRONOLÓGICAS

Informe de la Implementación del Diseño

Antes que usted examine la información, es decir los resultados de las mediciones repetidas de las series cronológicas, es esencial considerar el grado en que las condiciones de la implementación del diseño fueron tales como para permitir extraer conclusiones. Aquí hay algunas preguntas que usted necesitará formularse antes de analizar la información numérica:

Preguntas sobre la Implementación
de los Diseños de Series Cronológicas.

1. *¿Se implementó el programa realmente?, y si fue así ¿cuándo?*

Esta pregunta se refiere, por supuesto, a su documentación del programa. Usted necesitará presentar evidencias que muestren que el programa en realidad se implementó y, para un diseño de series cronológicas, debe confirmarse la fecha exacta o tiempo de inicio y finalización del programa.

2. *¿Hubo algunas influencias ajenas o confusiones?*

¿Tuvieron lugar algunos acontecimientos en el momento o cerca del momento en que comenzó el programa, los que con razón se pudiera decir que influyeron en las mediciones hechas? Suponga, por ejemplo, que usted estaba controlando, con tests mensuales, el progreso en matemáticas de un curso y que usted trajo un programa nuevo al colegio en Enero y encontró que los puntajes del test mensual dieron un salto considerable. Usted estaría inclinado a pensar que los materiales de matemáticas fueron útiles. Pero si el curso recibió un profesor nuevo al mismo tiempo que los materiales nuevos, se podría argumentar que fue el efecto del profesor en que se estaba notando en los puntajes mensuales más altos, más bien que el efecto de los materiales. Probablemente no habría forma de decidir cuál es la explicación correcta. El profesor nuevo fue un elemento de confusión que imposibilitó medir los resultados del programa.

Mientras más grande sea el número de unidades experimentales, las influencias ajenas o confusiones, tendrán menos efecto. Si usted hubiera estado controlando, en el ejemplo recién dado, muchos cursos habría sido poco probable que todos ellos recibieran buenos profesores nuevos, coincidentes con el programa.

No es una tarea fácil descubrir las influencias ajenas y evaluar si amenazan o no seriamente la interpretación de los resultados, y será una tarea diferente en cada situación. El sentido común es su mejor guía y un argumento razonado es su mejor medio para presentar sus ideas en un informe. Un método para examinar una influencia ajena es prestar más atención al factor tiempo. Si parece razonable esperar que un programa produzca un cambio rápido, entonces sus efectos probablemente pueden distinguirse de las influencias ajenas que producen un cambio lento. Por ejemplo, un colegio estableció un periodo más largo que el acostumbrado para almorzar. Poco después hubo quejas de un aumento en el número de alumnos que se

mezclaban en peleas. Algunas personas dijeron que las peleas se debían a la estadía transitoriamente más larga en el colegio, mientras que otras personas pensaron que era el resultado de las horas de almuerzo más largas. Examinar la serie de anotaciones que se refieren a peleas podría mostrar si el número de peleas aumentó repentinamente con la introducción repentina de la hora de almuerzo más larga o aumentó lentamente con el lento cambio en la composición del alumnado.

3. *¿Hubo algún cambio en el método de hacer las mediciones (observaciones) en el momento o cerca del momento de la intervención?*

Si, por ejemplo, el director que introdujo una hora de almuerzo más larga, simultáneamente instruyó a su personal para que enviara a la oficina a cualesquier alumno que encontrara peleando, entonces uno podría esperar un aumento en este tipo de situaciones simplemente por sus instrucciones, independiente del efecto de las horas de almuerzo más largas.
Recuerde que para derivar conclusiones exactas de un diseño de series cronológicas, el método de medición debe ser el mismo a través de la serie de mediciones.

4. *¿Ha cambiado la composición del grupo-E durante el experimento?*

Si usted está usando un diseño de series cronológicas longitudinal, necesita considerar el problema de pérdida de casos. Si la introducción del programa hace que algunos casos abandonen el grupo que se mide, entonces esto sólo podría influenciar las mediciones. Por ejemplo, puede ser que el director de un programa de atletismo que se desarrollaba después de clase llevó los registros de asistencia. Suponga que cuando se introdujo boxeo, el promedio de asistencia se elevó de 60% a 70%; el setenta por ciento de quiénes se inscribieron en atletismo para después de clase asistieron a él. Usted necesitaría preguntar si la composición de los grupos también ha cambiado. Quizás los alumnos que asistían con menos frecuencia cancelaron totalmente su inscripción cuando se introdujo boxeo y usted realmente tiene más del 10% de alumnos nuevos. Si esto ha sucedido, entonces sus mediciones ya no se basan en el mismo grupo. Para medir el impacto del programa sería mejor considerar separadamente la información sobre aquellos que permanecieron desde el comienzo hasta el final, eliminando de las primeras mediciones a aquellos que abandonaron el programa más tarde.

Si usted está usando un diseño de series cronológicas de grupos sucesivos, necesitará considerar cuidadosamente si los grupos mismos están cambiando en o cerca del momento en que usted está esperando ver cambios debido al programa. Cambios rápidos en las características de los alumnos que entran a un nivel pueden resultar debido a la apertura de una industria nueva en el sector, del influjo que resulta de una empresa profesional grande tal como una universidad que se ubica en algún lugar cercano, los cambios que resultan del movimiento hacia los barrios residenciales, etc. Tales cambio demográficos significan que los grupos sucesivos no son representativos de una población estable en la localidad. La población está cambiando, a veces rápidamente, otras veces en forma lenta. Es difícil distinguir entre el efecto de cualesquier programa y los efectos de estos cambios.

Una forma de abordar el problema de los cambios de características de una población es hacer otras mediciones en los grupos además de las mediciones que se espera que muestren el efecto del programa. Por ejemplo, si usted está llevando registros de los puntajes de matemáticas antes y después de la introducción de un programa de

matemáticas para séptimo y octavo grado, usted también podría reunir puntajes verbales durante el mismo periodo de tiempo. Si los puntajes de matemáticas mostraron un salto con la introducción del programa nuevo pero no sucedió lo mismo con los puntajes verbales, esto favorecería su argumento que establece que el programa causó el cambio más bien que la composición del grupo: no fueron alumnos más inteligentes sino un programa mejor.

5. *¿Se introdujo el programa en respuesta a una crisis?*

Las crisis generalmente pasan y las cosas vuelven a lo normal. Por consiguiente, si usted ha introducido un programa nuevo en respuesta a una crisis, fue probablemente sólo una de las muchas respuestas que muchas personas estaban dando a la crisis. Cuando las cosas mejoran, es difícil sostener que fue el programa en que lo hizo, puede haber sido el curso normal de los acontecimientos —las cosas mejoran después de ser muy malas.

Análisis de la Información de Resultados de un Diseño de Series Cronológicas.

Si usted está convencido que el Diseño ha sido bien implementado, y parece ser una buena oportunidad para sacar conclusiones de las mediciones repetidas que usted ha hecho, entonces siga adelante y examine la información. La frase: "examine la información" se usa aquí en lugar de "analice la información" porque el análisis estadístico de la información de series cronológicas es complejo y tendría que procurarse los servicios de un experto en estadística capaz de realizarlo. Sin embargo, se pueden extraer conclusiones tentativas graficando y examinando la información. Un gráfico es evidencia bastante poderosa para la mayoría de las evaluaciones. La información de series cronológicas puede examinarse mejor cuando se presenta en forma gráfica. A continuación se indican los pasos para preparar el gráfico:

1. Por cada conjunto de puntajes reunidos de una vez (o reunidos durante un periodo de tiempo) compute un estadístico que resuma la medida de resultado tal como el total, la media, la mediana, un porcentaje que responde a una forma determinada, etc. Si la medida de resultado se basa en una serie acumulada de registros, entonces puede ser que usted quiera considerar cuidadosamente cómo agrupar la información. Por ejemplo, un registro de ausencias pudiera basarse en totales diarios, pero totales semanales probablemente mostrarían menos fluctuación y producirían un gráfico más claro.

2. Haga un gráfico en el que el eje vertical es la medida de resultados y el eje horizontal es tiempo.

3. Indique las fechas de observaciones sobre el eje horizontal.

4. Grafique el estadístico que resuma la medida de resultados para el tiempo que corresponda.

5. Indique en el gráfico el comienzo y el fin del programa, si el programa ha terminado durante el periodo de observación.

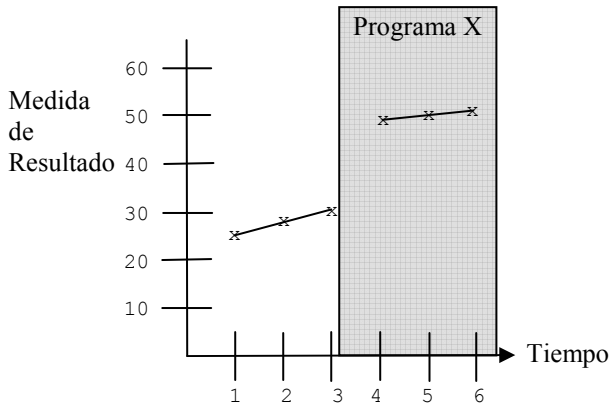
Ejemplo. Una Tabla de Información de Series Cronológicas

Media Aritmética de los Puntajes del Grupo-E en la Medida de Producto

	Tiempo					
	Antes del Programa			Después del Programa		
	1 (Enero)	2 (Febrero)	3 (Marzo)	4 (Abril)	5 (Mayo)	6 (Junio)
Media Aritmética Grupo-E	20	25	30	50	55	60

Ejemplo. Un gráfico correspondiente a la Tabla.

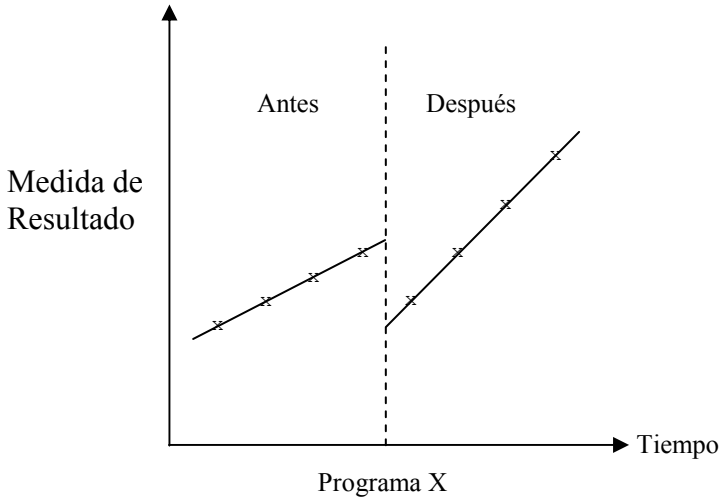
Media Aritmética de los Puntajes del Grupo-E



El Examen de los Gráficos de Series Cronológicas

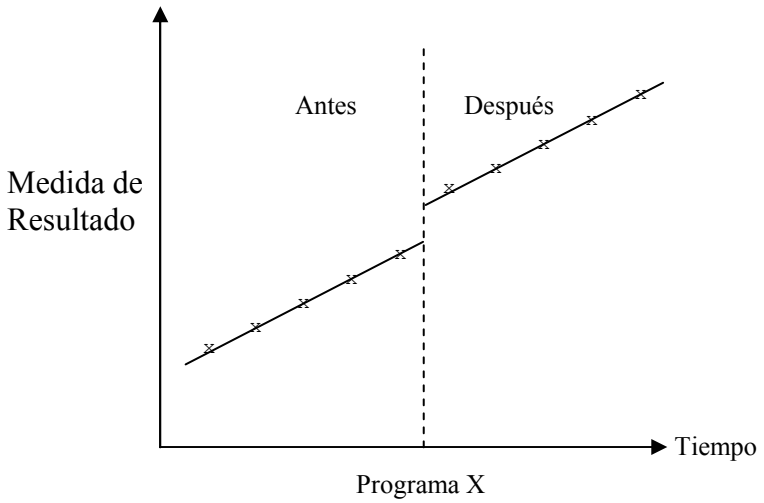
El efecto del programa X se puede determinar viendo si los puntajes obtenidos antes que se implementara el programa X eran diferentes de aquellos obtenidos después que se implementó el programa. Una manera de observar esto es trazar una línea a través de los puntajes "anteriores" y otra a través de los puntajes "posteriores" (este procedimiento se describe en la sección que viene más adelante, llamado "trazando las líneas"). El gráfico puede examinarse con el fin de observar dos tipos de efectos principales: un cambio de tendencia o un cambio de nivel ("salto"). Un cambio en pendiente o tendencias indica un cambio en la tasa de progreso. Un cambio en el nivel (altura) del gráfico podría indicar un aumento repentino en rendimiento. Por supuesto, ambos efectos podrían ocurrir. Examine los ejemplos a modo de ilustración.

Ejemplo. Un cambio de Tendencia



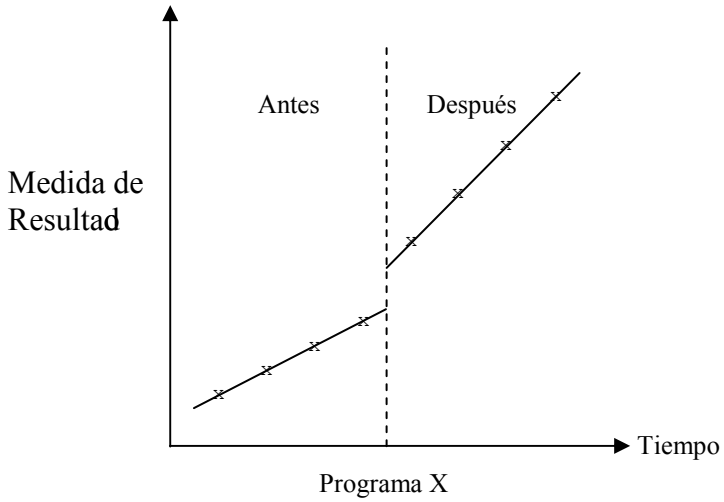
Después del programa X, los puntajes aumentaron más rápidamente −hubo un cambio en la *tendencia* ascendente.

Ejemplo. Un salto (cambio en nivel, pero no cambio en tendencia)



El programa X pareció causar un salto. Después de eso, la tendencia continuó con el mismo ritmo (las gráficas antes y después son paralelas).

Ejemplo. Cambio en salto y tendencia

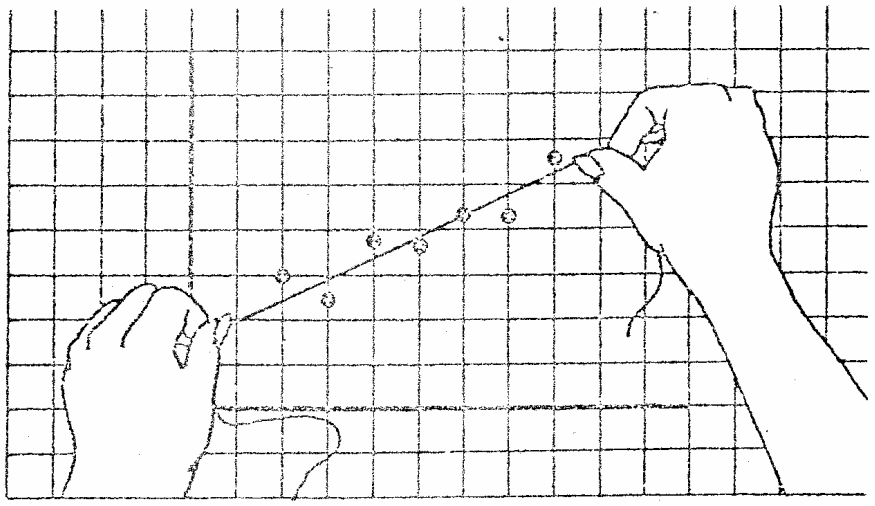


Programa X

Observe que además del salto, la inclinación (tendencia) de los puntajes después del programa no es la misma que antes del programa. Los puntajes se elevan más rápidamente después del programa, por ej., la inclinación de la gráfica es más empinada.

Trazando las Líneas

Con el fin de graficar los resultados de series cronológicas, usted necesitará trazar la mejor línea recta a través de los puntajes obtenidos antes que se implementara el programa X. También necesitará trazar la mejor línea recta a través de los puntajes que se obtuvieron después que se implementó el programa X. Un método simple para trazar "la mejor línea recta" para representar una tendencia en varios puntajes, es usar un pedazo de hilo como se muestra en la ilustración. Ajuste el hilo hasta que parezca que representa lo mejor posible la tendencia en las observaciones.

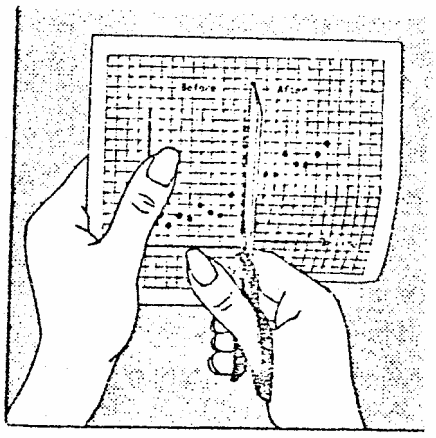


Cuando usted está ajustando la posición del hilo, imagine que cada puntaje (observación, punto en el gráfico) está pegado al hilo con un elástico. Ponga el hilo en la posición que usted piensa equilibraría mejor el hilo entre los puntajes. Haga esto

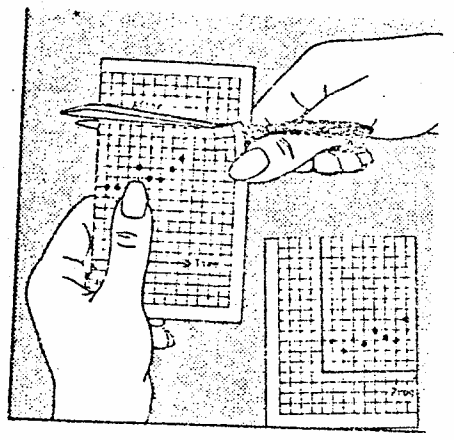
para los puntajes preprograma, para aquellos durante el programa, y aquellos que se obtuvieron después del programa, si es que hubo algunos.

Puesto que adecuar una línea de tendencia usando un hilo no es un procedimiento totalmente objetivo, es aconsejable tomar las siguientes precauciones contra una posible subjetividad, antes que comience con el procedimiento del hilo:

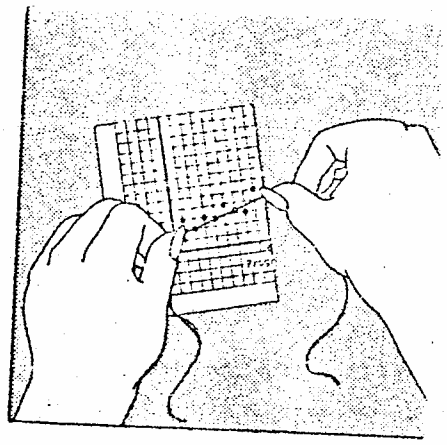
- Haga una copia del gráfico
- Corte la copia en dos a lo largo de la línea "programa X".



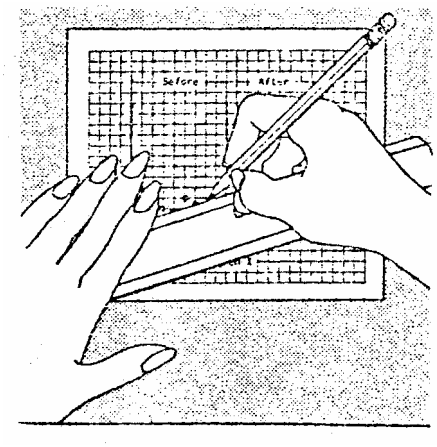
- Recorte el gráfico de tal manera que no se vea cual pedazo es la parte "Antes" y cual es la parte "Después".



- Luego haga que alguien que no esté involucrado, ajuste la mejor línea recta en cada mitad.

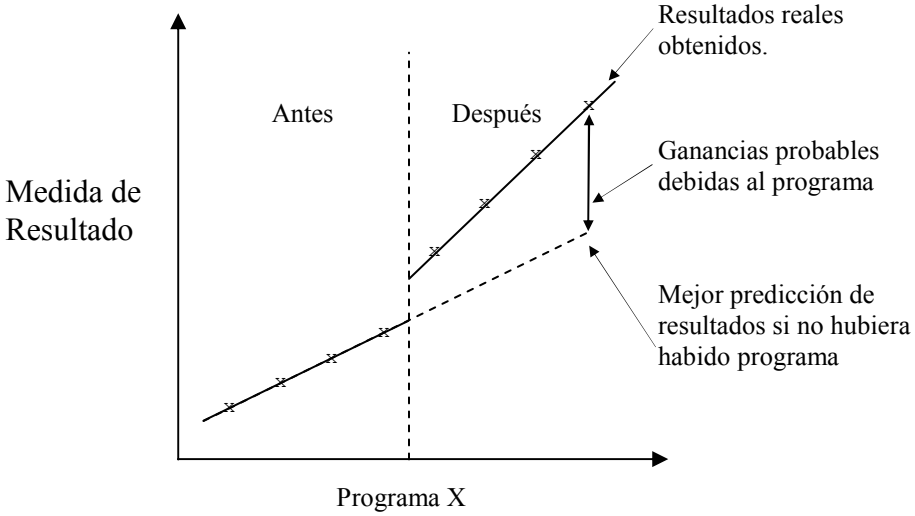


- Copie nuevamente las dos líneas de tendencia en el gráfico original, lo más exactamente posible.
- Ahora usted tiene dos líneas de tendencia, independientemente estimadas, una para antes y otra para después de la implementación del programa X.



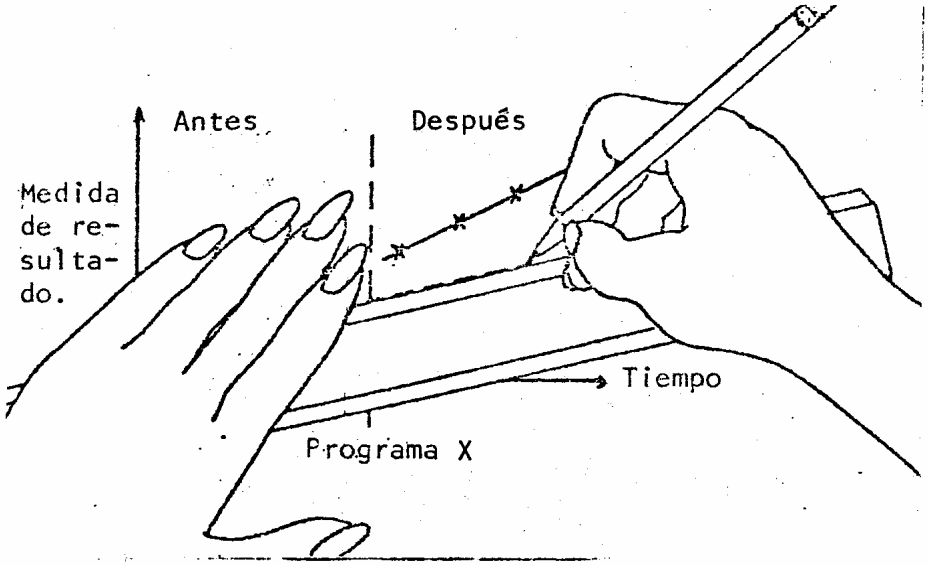
NOTA: En los diagramas de esta sección, las observaciones (puntajes, puntos en el gráfico) están alineadas en forma muy clara. Esto rara vez sucede. Los diagramas se dibujan de esta manera para mayor claridad, a expensas del realismo.

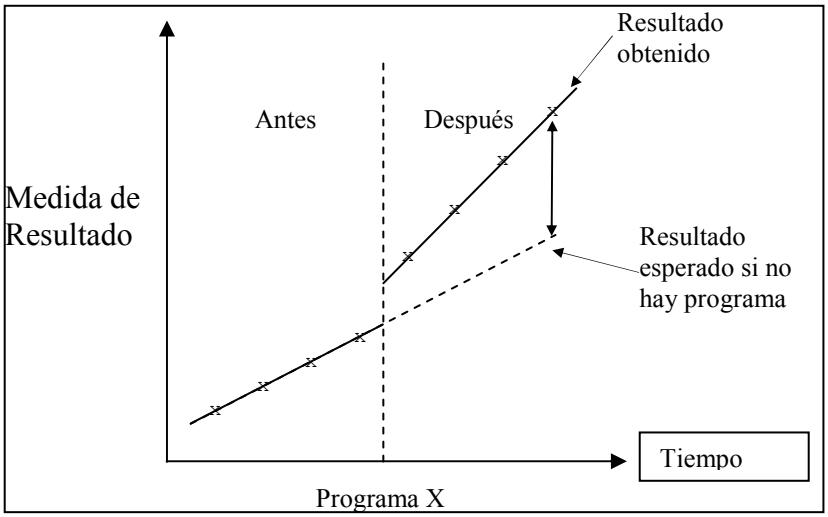
Una manera de examinar los resultados de las series cronológicas es extrapolar las tendencias preprograma para mostrar una estimación de lo que los resultados podrían haber sido sin el programa. El ejemplo hace esta extrapolación extendiendo la línea de tendencia "antes" a la región "después".



Ejemplo.

Para extrapolar, ponga una regla a lo largo de la línea de tendencia"antes" y trace una línea de punto con la regla en la región "después".





Al entregar la información de las series cronológicas, usted puede sugerir que es *el programa* el que ha producido la diferencia entre los resultados *esperados*, ubicados por la línea de tendencia, y los resultados que obtuvo. Sin embargo, usted necesita considerar varios problemas posibles.

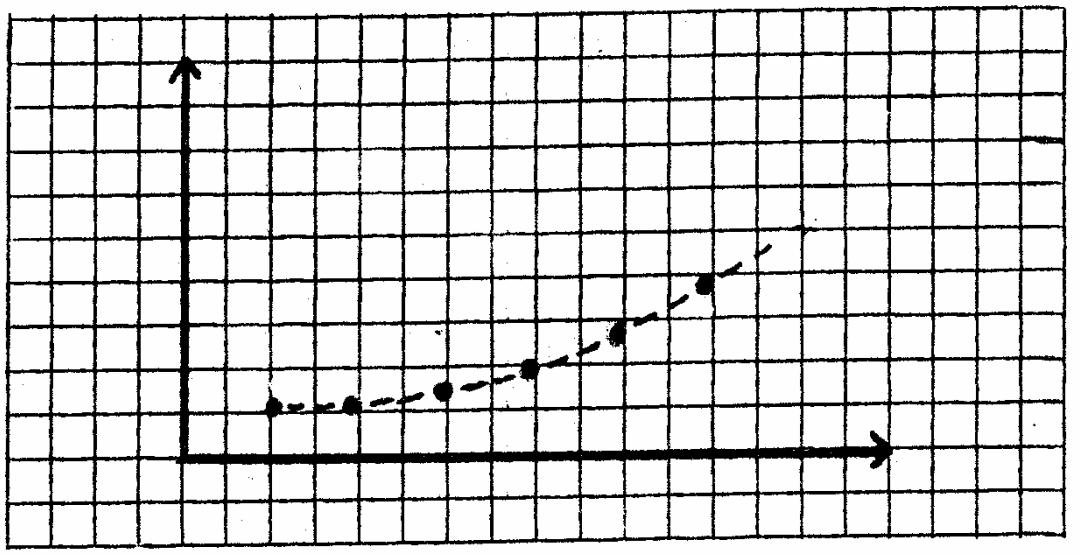
1. *¿Hubo algún otro cambio en el momento o casi al mismo tiempo del programa que pudo haber causado el cambio de antes a después?*

Es prudente sugerir otras causas posibles y discutir cuán probable es que hayan sido efectivas. El problema de tratar con influencias extrañas de discutió anteriormente, y puede recordarse con estas preguntas:

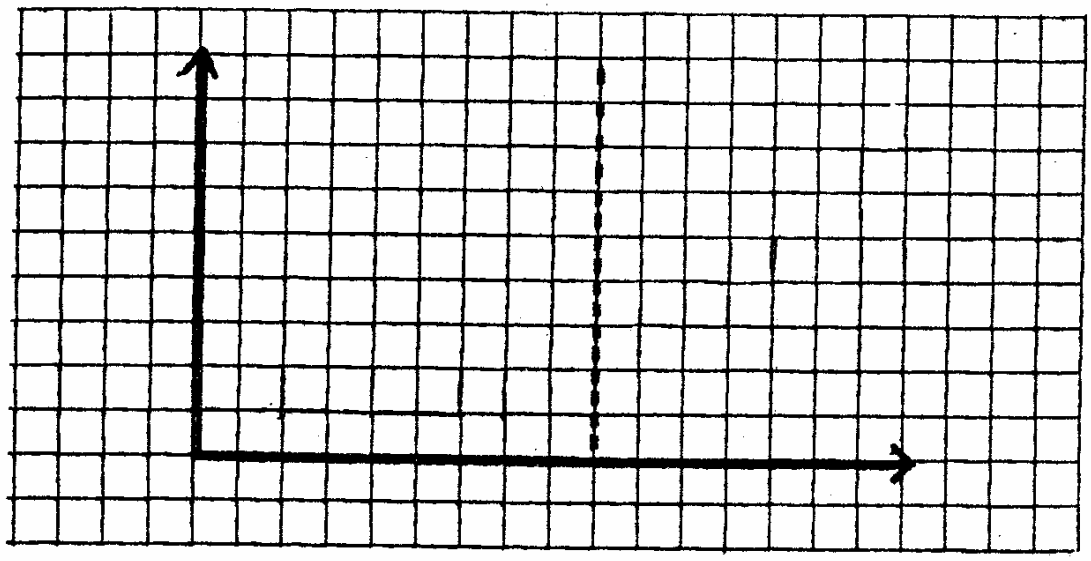
- ¿Quedó igual la composición del grupo?
- ¿Se mantuvo igual el método de medición?
- ¿Hubo otros sucesos aparte del programa que pudieron haber afectado al grupo?
- ¿Hubo una crisis en el momento en que se introdujo el programa?

2. *¿Fue adecuado el uso de líneas rectas?*

Por ejemplo, si el gráfico se hubiera visto como éste , sin un programa,



podrí a ser desorientador traducir eso a dos líneas rectas.

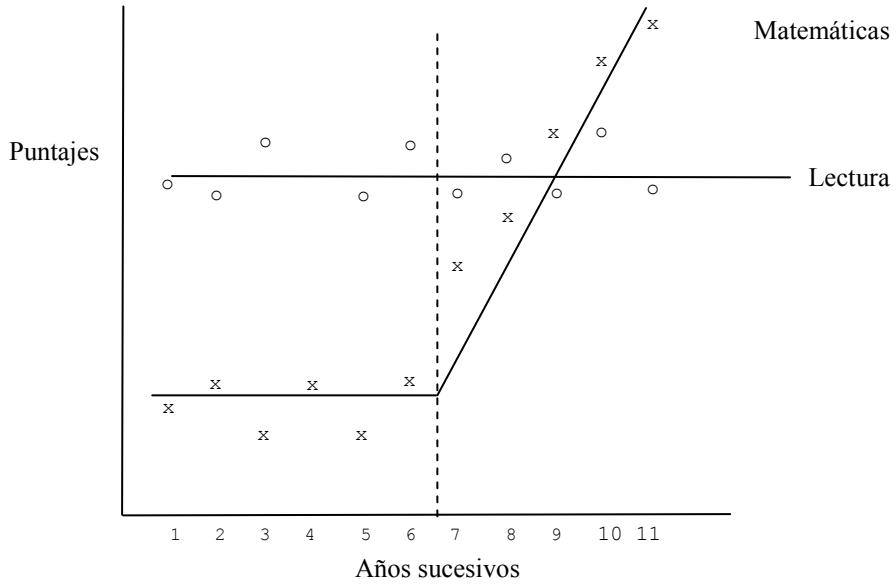


Por supuesto, no es posible estar seguro si el uso de líneas rectas es adecuado o estar seguro de cuáles hubieran sido los resultados sin el programa. Esto es simplemente un punto débil del diseño de series cronológicas de grupo único. Debido a que este diseño no es tan lógicamente poderoso como el diseño experimental verdadero, (Diseño 1) es que se llama un diseño "cuasi experimental".

3. *¿Reflejan quizás los resultados algún patrón cíclico más bien que un cambio debido al programa?*

Por ejemplo, quizás el rendimiento siempre mejora en la segunda mitad del año escolar. Si usted está usando un diseño de grupos sucesivos y usted tiene razón para sospechar una variación cíclica, entonces trate de comparar puntos similares en el ciclo. Por ejemplo, si hay un ciclo anual, compare los puntajes de los tests de mayo con puntajes de tests de otros meses de mayo, no con puntajes de tests de septiembre. A veces el argumento "habría sucedido de todas maneras" puede examinarse graficando otro conjunto de información. Por ejemplo, quizás usted ha estado examinando los puntajes de matemáticas para grupos sucesivos y muestran una

ganancia considerable después de un programa muevo. Alguien argumenta, "Pero los grupos de alumnos ingresados eran precisamente de habilidad mayor. No fue en absoluto el programa". Para investigar esta sugerencia, grafique los puntajes en lectura de los grupos sucesivos:



Si la lectura tampoco mostró un salto, esto fortalece el argumento de que fue el programa, no los alumnos, lo que causó el cambio observado.

La misma estrategia funcionará para un diseño de series cronológicas longitudinales, excepto que aquí usted estará graficando lectura y matemáticas para el mismo grupo durante todo el tiempo. Esto lo resguarda contra el argumento de la variación cíclica durante un solo año escolar.

4. *¿Son los resultados demasiado inestables como para permitir que se extraigan conclusiones?*

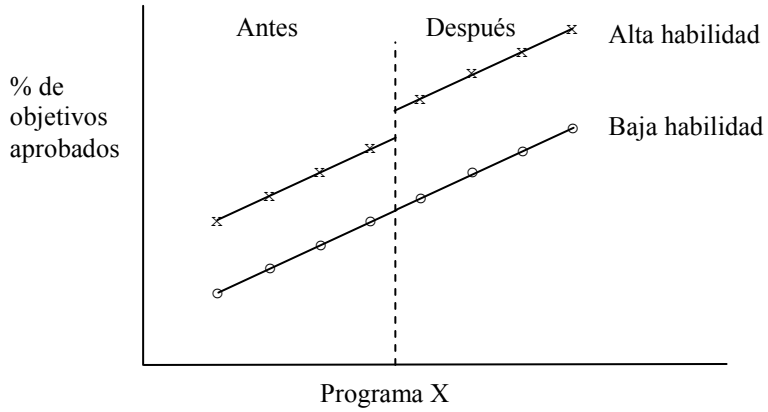
Si los puntajes u observaciones están "en todas partes" y son muy variables, podría no ser posible detectar los resultados del programa, incluso si está teniendo un efecto o no. Los análisis estadísticos de la información de series cronológicas están más allá del alcance de este libro y no se usan mucho todavía. Quizás usted concluirá de la información de series cronológicas que bien vale la pena organizar una prueba más poderosa del programa, empleando el Diseño 1, el diseño de grupo control verdadero, si es posible.

Un mayor Análisis del Diseño de Series Cronológicas

Podría ser que usted quisiera graficar la media aritmética de los puntajes de la medición de resultados, no para el grupo completo, sino para *parte* del grupo. Hacer esto le dará una idea de si el programa afectó a los alumnos diferentes en forma diferente. Quizás si fue particularmente bueno —o dañino— para alumnos de alta o baja habilidad, o para niños o niñas. Quizás si el programa fue bueno o insuficiente para diferentes alumnos por razones desconocidas. Si es así, habrá aumentado la dispersión o varianza de los puntajes que usted obtuvo.

1. Para indicar el patrón de puntajes para alumnos de alta habilidad, usted podría:

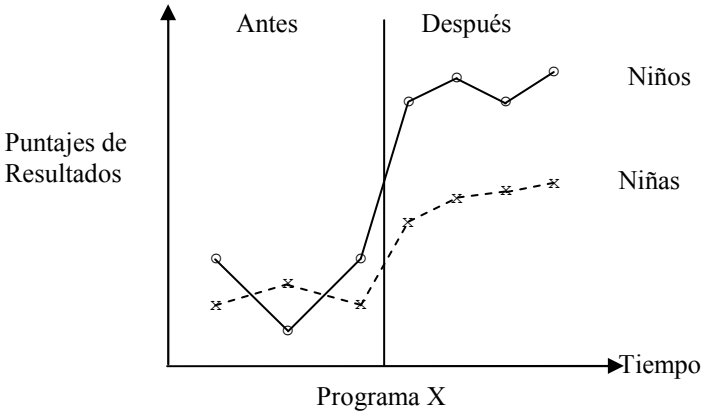
- dividir el grupo en habilidad alta y baja usando alguna medida <u>que no sea la medida de resultado</u> (para evitar regresión)
- haga los gráficos de series cronológicas para los grupos de habilidad alta y baja.



Ejemplo.

Este gráfico indica que el programa X hizo diferencia en los niños de alta habilidad, pero no afectó a los alumnos de baja habilidad.

2. Para indicar los efectos del programa en niños, separadamente, usted podría graficar sus puntajes en forma separada.



Ejemplo.

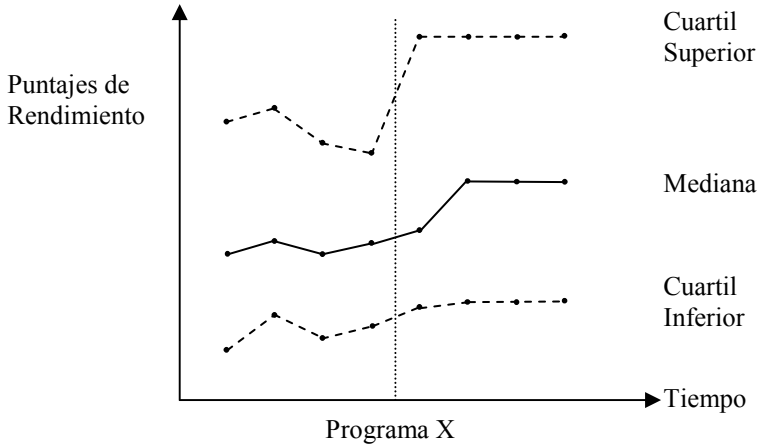
---

3. Para indicar la dispersión o variación de los resultados de la medición de resultados.

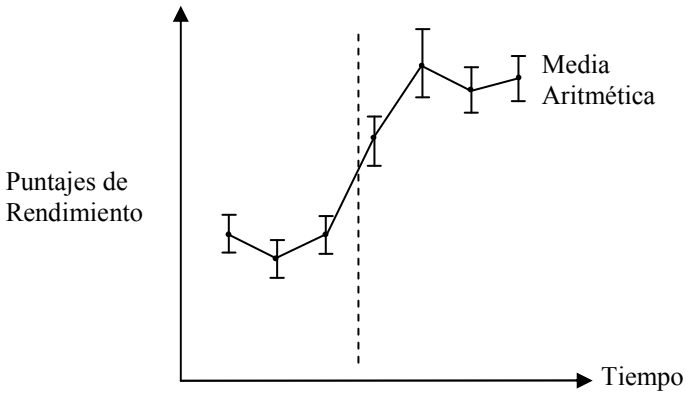
-Si usted grafica puntajes <u>medios</u>, entonces agregue los puntajes de cuartiles superiores e inferiores.

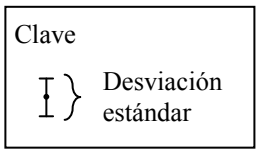
-Si usted grafica puntajes de medias aritméticas, entonces indique la desviación estándar para cada puntaje. (NOTA: No grafique los puntajes más altos ni los más bajos. Son demasiado inestables).

Ejemplos.



Programa X





DISEÑO 5

EL DISEÑO DE SERIES CRONOLÓGICAS
CON UN GRUPO CONTROL NO-EQUIVALENTE

DISEÑO 5

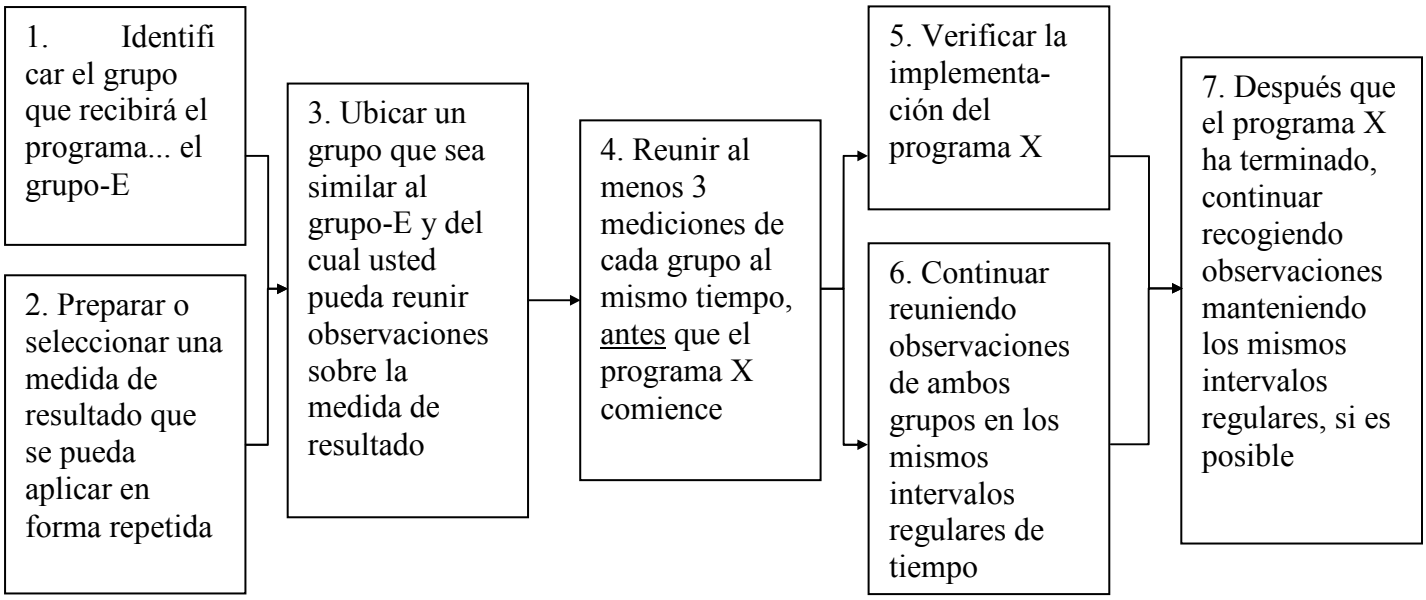
EL DISEÑO DE SERIES CRONOLÓGICAS CON UN GRUPO CONTROL NO–EQUIVALENTE

DIAGRAMA:

	Tiempo							
	1	2	3	4	5	6	7	8
Grupo Experimental	0	0	0	0	X 0	0	0	0
Grupo Control No-Equivalente	0	0	0	0	0	0	0	0

Resumen. Dos grupos semejantes, pero que no se formaron por asignación al azar, se miden a intervalos regulares antes y después que el programa X se implementa. Este Diseño es como el Diseño 4... con el agregado de un grupo de comparación no-equivalente.

PASOS ESENCIALES AL IMPLEMENTAR EL DISEÑO 5

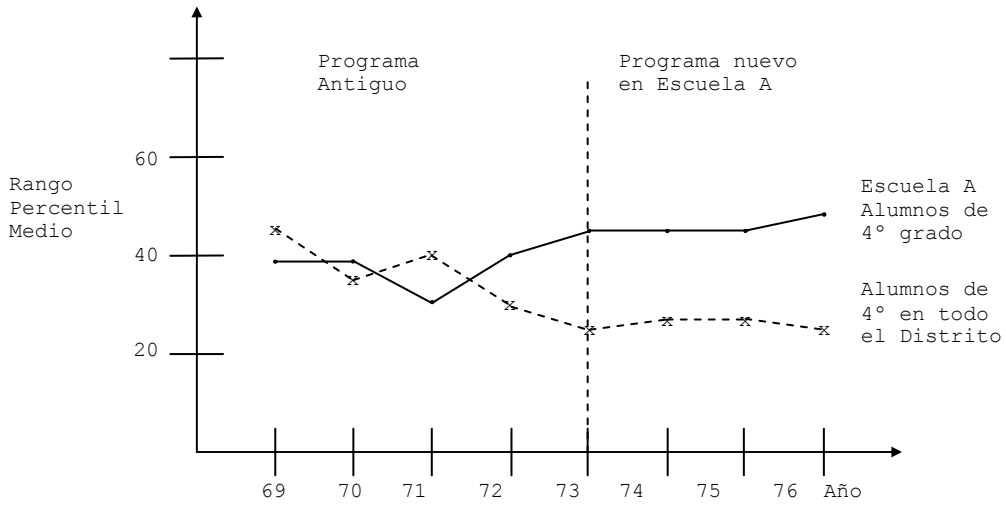


EJEMPLOS

Ejemplo 1

Los alumnos se estaban quejando que la "hora de almuerzo" era demasiado corta y pidieron que se alargara en 15 minutos. Algunos miembros de la dirección objetaron esto basándose en que las peleas estallaban durante las horas de almuerzo y que mientras más largas las horas de almuerzo más a menudo pelearían los alumnos. Otros miembros de la dirección estuvieron en desacuerdo y pidieron una hora de almuerzo más larga como prueba. La directora se dio cuenta que a medida que se acercaran los días calurosos, el número de peleas aumentaría, independientemente de cuán larga fuera la hora de almuerzo. Por consiguiente, decidió que no sería suficiente llevar sólo un registro del número de peleas que se informaban. Ella necesitaría comparar tendencias en su colegio con tendencias en una escuela semejante con el fin de prever el aumento en peleas que parecería inminente.

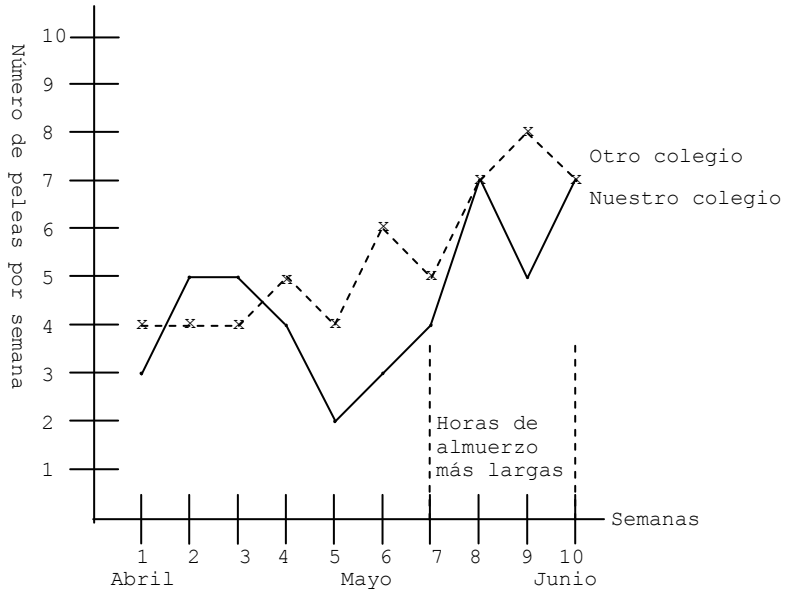
Después de poner en práctica la hora más larga de almuerzo por un periodo de ensayo de cuatro semanas, ella examinó sus propios registros para encontrar el número de peleas que habían ocurrido en las horas de almuerzo durante las 10 semanas anteriores. Ella persuadió al director de una escuela cercana para que buscara en sus registros y le proporcionara el mismo conjunto de datos estadísticos. Ella entonces hizo el gráfico siguiente:



A partir de estos gráficos el cuerpo de profesores y la directora concluyeron que aunque había habido un aumento en el número de peleas, probablemente la razón no fue las horas de almuerzo más largas, puesto que la escuela semejante, vecina, había experimentado un aumento al mismo tiempo, (observe como el gráfico habría sido desorientador si no se hubiera agregado el grupo control no-equivalente).

Ejemplo 2.

Tres profesores de cuarto grado en una escuela de un determinado distrito habían planificado un nuevo programa de lectura. Ellos habían usado el conjunto de materiales "Planificación de programa" del Centro de Estudios de Evaluación, durante una semana de trabajo de verano, y ahora estaban listos para implementar el programa. Al pensar como evaluar dicho programa, ellos decidieron que una medida que deberían examinar era el test de rendimiento estandarizado administrado todos los años en mayo. Los profesores supieron que la misma forma del test se había administrado durante los cuatro últimos años, así es que los resultados de los alumnos de cuarto grado en su escuela se obtuvieron del Director de Investigación y Evaluación del Distrito. El director también les proporcionó un promedio de las notas de los alumnos de cuarto grado del distrito en cada uno de los cuatro años anteriores. Continuando con la recopilación de esta información durante los próximos tres años, se hizo el gráfico siguiente:



ANÁLISIS E INTERPRETACIÓN DEL DISEÑO 5

El Diseño 5 es como el Diseño 4, excepto que agrega una serie extra de información del grupo control no-equivalente. Esta información lo ayuda a excluir explicaciones alternativas sobre cualesquier cambio en las medidas de producto que se encuentra y que coinciden con la introducción del programa experimental.

Debido a que el Diseño 4 y 5 son semejantes, al lector se le pide que lea el Diseño 4 para que tenga una orientación más completa al usar e interpretar el Diseño 5. Por supuesto, si se puede formar un grupo control verdadero, las conclusiones que se van a extraer de un diseño de series cronológicas serían incluso más poderosas. En realidad, un diseño de series cronológicas con un verdadero grupo control sería el diseño más poderoso posible.

CAPÍTULO 5

DISEÑO 6:

PROCEDIMIENTOS, ANÁLISIS E INTERPRETACIONES

El Diseño 6 es el diseño menos adecuado. En evaluación sumativa se debería hacer todo el esfuerzo posible para usar uno de los otros diseños. El Diseño 6 debería considerarse sólo como el último recurso.

DISEÑO 6

EL DISEÑO ANTES Y DESPUÉS

DISEÑO 6

EL DISEÑO ANTES Y DESPUÉS

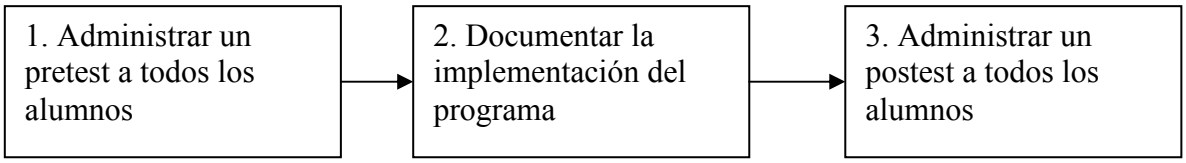
(EL DISEÑO DE GRUPO ÚNICO PRETEST-POSTEST)

DIAGRAMA

	Tiempo		
	1 (pre)		2 (post)
Grupo-E	0	X	0

Resumen: Los únicos alumnos que se miden son aquellos que reciben el programa. A ellos se les administran un pretest y un postest.

PASOS ESENCIALES AL IMPLEMENTAR EL DISEÑO 6



Ejemplo.

Los fondos del Proyecto I se pusieron a disposición de cuatro escuelas en un distrito escolar. A todos los alumnos se les administró un test estandarizado a comienzos a fin de año.

El evaluador informó al Estado sobre los progresos alcanzados durante el año, usando formularios proporcionados por el Estado. El evaluador también estimó que parte importante del trabajo era informar a los padres y al personal docente del colegio.

PRESENTACIÓN Y ANÁLISIS DE LA INFORMACIÓN PARA EL DISEÑO 6

La interpretación de resultados para el diseño "Antes y Después", Diseño 6, presenta serios problemas. Debido a que es difícil saber qué tipos de resultados se podrían haber obtenido sin el programa, es casi imposible decir con seguridad cuán buenos son los resultados obtenidos. Al no tener ningún grupo control o grupo de comparación, uno ha perdido la oportunidad de detectar aquellas ganancias pequeñas pero importantes, que son quizás, todo lo que desde un punto de vista realista se puede esperar en un año o algo así de trabajo con un programa nuevo.

Sin embargo, de todas maneras ésta es una situación común, y este capítulo trata de indicar aquellos pocos aspectos positivos que puedan extraerse de este oscuro diseño.

Una ventaja mínima es el tiempo que se ahorra al no tener que controlar dos grupos. Debido a que el evaluador tiene solamente un grupo que medir, esta persona puede hacer más mediciones, reunir más información, llegar a ser más un investigador de un grupo, que si ellos tuvieran que seguir la pista de las medidas de documentación y de producto para dos grupos.

Al implementar el Diseño 6, el evaluador debería ser capaz de hacer un excelente trabajo de documentación de programa, describiendo en detalle los materiales y actividades del programa y relacionando esto a la razón fundamental sobre cómo se suponía que el programa lograra sus objetivos.

En cuanto a las medidas de resultado, el evaluador podría informar los puntajes obtenidos en los tests estandarizados, en una tabla tal como la que se muestra a continuación.

TABLA 5

PUNTAJES PROMEDIO OBTENIDOS EN EL PRETEST Y POSTEST DE LECTURA Y MATEMÁTICA EN LOS COLEGIOS QUE ESTÁN EN EL PROGRAMA X Y Z

Grupo	n[1]	Pretest	Postest	Test-t
		Lectura		
Colegio A	401	59.4	64.3	3.8[2]
Colegio B	720	50.2	70.5	12.2[2]
Colegio C	364	40.8	60.2	4.5[2]
		Matemática		
Colegio A	461	63.2	70.1	2.4[2]
Colegio B	726	58.4	71.2	3.1[2]
Colegio C	362	32.9	33.4	0.8

Los resultados que se presentan en la Tabla de arriba muestran una situación en que el mismo test se ha administrado como pretest y como postest. El test-t comprueba la significación de la diferencia entre los puntajes del pretest y del postest y es un test-t para grupos pareados. Si los tests fueran tests cognoscitivos (de rendimiento) y el programa durara cualesquier cantidad de tiempo, casi siempre se esperaría una diferencia significativa entre el pre y postest. Los alumnos rinden mejor en los tests a medida que crecen. Sin embargo, usted podría encontrar una situación tal como la que se muestra para el caso de matemáticas en el colegio C en la Tabla 5 una diferencia no significativa entre los puntajes del pre y postest en matemáticas. En este caso alguien ciertamente debería descubrir qué es lo que estaba pasando en el colegio con respecto a matemáticas. ¿Hubo quizás algún error en la prueba? ¿o es que realmente los alumnos no están progresando?

A menudo en la situación del Diseño Antes-y-Después el evaluador tiene que informar sobre los resultados a las agencias Estatales o Federales en formularios que ellos proveen. Estos formularios generalmente requieren los resultados de test estandarizados, referidos a normas. Puede ser que el evaluador también quiera presentar los resultados de pruebas estandarizadas, publicadas, referidas a normas, a un público lego, tal como la unta Escolar o la Asociación de Padres y Apoderados.

Presentación de puntajes de pruebas estandarizadas

Antes que se presente cualesquier puntaje, es mejor hacer unas pocas observaciones sobre la naturaleza de los tests de rendimiento estandarizados. Pudiera ser que usted quisiera describir cómo están "estandarizados" los tests, explicar lo que es un grupo "norma" y luego enfatizar lo siguiente:

1. Debido a la naturaleza de los tests estandarizados, la mitad de los alumnos del grupo norma tienen que estar "bajo el promedio".

[1] número de alumnos presentes tanto en los pretest como en los postest y por lo tanto los alumnos sobre puntajes se calculó el test-t

[2] estadísticamente significativo al nivel .05

Esto es, por lo general, fácilmente comprendido cuando se mencionan los puntajes percentiles, pero se olvida cuando se discuten los puntajes "grado equivalente". Un puntaje "grado equivalente" simplemente es también un puntaje promedio y se calcula de tal manera que la mitad de los alumnos queden bajo el nivel y la otra mitad sobre el nivel del grado. Los puntajes "grado equivalente" parece que fueran puntajes referidos a criterio, pero no son. Los tests estandarizados nunca mostrarán a todos leyendo al nivel o sobre el nivel del grado. Además de ser fácilmente mal interpretados, los puntajes "grado equivalente" tienen ciertos problemas estadísticos asociados consigo, lo que les da una exactitud dudosa. Deberían evitarse en los posible.

2. Los tests estandarizados se mantienen secretos.

Los profesores no pueden saber los conceptos y destrezas que el test pretende extraer y por lo tanto no pueden estar seguros de enseñar lo que el test mide. Esto tiene sus ventajas, pero también significa que los tests pueden ser poco familiares para los alumnos y puede que no midan muchas de las cosas que los alumnos han aprendido. Es decir, los tests estandarizados puede que no sean muy buenos indicadores de cuánto aprendieron los alumnos o de cuán bien se les ha enseñado.

Habiendo advertido sobre lo que significa juzgar un programa únicamente sobre la base de tests estandarizados, usted puede presentar un conjunto de resultados como un indicador general de la efectividad del programa.

Sugerencias para el Evaluador Antes – y– Después

Las actividades hasta aquí descritas para la implementación del Diseño 6, documentación del programa, presentación de la información pre y postest, prueba de significación estadística y presentación de los resultados de tests estandarizados representan el mínimo necesario para darle credibilidad al diseño.

A continuación se dan otras sugerencias para examinar y describir el programa. Además de estas sugerencias observe los comentarios hechos en el capitulo 8 con respecto a la evaluación formativa.

1. *Tenga cuidado en la elección de los períodos de administración de pruebas si usa tests estandarizados.*

¿Cómo pueden interpretarse los puntajes finales si no hay grupos de comparación? Una solución al dilema es usar tests de rendimiento estandarizados, referidos a normas y comparar los resultados del grupo-E con los puntajes del grupo "norma" representativo de la nación. *Esta comparación debería emplearse solamente si usted administra un pretest y un postest al grupo-E en los mismos periodos durante el año escolar en que a los grupos norma se les administró un pretest y un postest.* La información sobre las oportunidades en que a los grupos norma se les administró un test debería entregarse en los manuales que publican los editores. Esta precaución es necesaria porque las normas que el editor da a conocer sobre oportunidades intermedias entre las ocasiones verdaderas de administración de tests son esencialmente adivinanzas basadas en extrapolaciones que pueden ser bastante inexactas.

2. Establezca una evaluación basada en la medición referida a objetivos

Si las metas del programa pueden descomponerse y especificarse como objetivos, entonces usted puede medir el logro de los objetivos con tests referidos a criterios. Haciéndolo así, usted obtendrá las siguientes ventajas:

a) Usted podrá informar sobre las áreas fuertes y débiles del programa estableciendo qué objetivos han sido alcanzados por un gran número de alumnos y qué objetivos no han sido alcanzados por muchos alumnos.

b) Usted puede pedir a los profesores o personal directivo que califiquen los objetivos en términos de su importancia, lo que le ayudará a concentrarse en mediciones cuidadosas de los más importantes.

c) Usted puede pedir a los profesores o personal directivo que establezcan ciertos criterios mediante los cuales ellos juzgarán si los objetivos se han logrado. Cuando los objetivos se han especificado claramente, con ítemes de tests de muestra, se puede pedir a los profesores y otras personas que establezcan criterios o expectativas: ¿Qué objetivos pronostican ellos que serán alcanzados por la mayor parte de los alumnos? Usted puede usar estos criterios para juzgar el éxito o fracaso en cada uno de los objetivos. Hay un par de objeciones serias a este procedimiento, pero es aun mucho mejor que simplemente mirar un sólo puntaje global y tratar de interpretarlo:

-- Una objeción a la evaluación de un programa en términos del logro de sus metas establecidas por el personal directivo, es que el éxito en el logro de los objetivos, de acuerdo al criterio establecido por el personal directivo, podría indicar una enseñanza exitosa o podría simplemente indicar el establecimiento de estándares bajos. Del mismo modo, la falta de éxito en que los alumnos alcancen criterios especificados previamente podría indicar un programa pobre, o podría indicar un personal ambicioso con muchas esperanzas, que fija expectativas altas más bien que expectativas bajas, fácilmente alcanzables.

-- Puede que no sea una buena razón atribuir el logro o no logro de los objetivos al programa del colegio. Quizás el logro se debió al crecimiento natural de los alumnos, o quizás el fracaso en alcanzar los objetivos se debió a un alto promedio de ausentismo entre los alumnos y no a un programa pobre. Sin un grupo control, es difícil atribuir los resultados al programa.

3. Considere el énfasis en los diferentes componentes.

Quizás si todos los lugares (colegios o cursos) en que el programa se implementa están usando los recursos del programa de la misma forma, de tal modo que a usted realmente se le confronta con un programa homogéneo. Quizás en un colegio, por ejemplo, el compromiso de los padres se enfatiza bastante, mientras que en otro lugar una gran cantidad de fondos se usaron para actividades de desarrollo del personal. Pudiera ser que usted quisiera ver si estos énfasis diferentes produjeron resultados diferentes, detectables. Usted podría tratar los lugares que han puesto diferente énfasis, proporcionando grupos control no equivalentes, no para todo el programa sino que para componentes diferentes del programa. Una oportunidad para hacer una comparación podría surgir, incluso, si un lugar asignara más tiempo a lectura que lo que asignó otro lugar. Usted entonces podría querer ver si el tiempo extra gastado en

lectura pareció mejorar los puntajes del postest en lectura. En estos casos, use las instrucciones para el Diseño 3, que comienza en la página 71.

4. Examine el impacto diferencial del programa en alumnos con diferentes niveles de habilidad, o de diferente sexo, o en alumnos con diferentes promedios de asistencia, o en alumnos a quienes los profesores califican como altamente motivados en oposición a alumnos pobremente motivados. Dividiendo y examinando los resultados de varios subgrupos, usted podría encontrar indicaciones de los alumnos para quienes el programa está funcionando mejor. Esto es válido, por supuesto, tanto para medidas de actitud, como también para medidas de habilidad. Esta "división" exigirá uno de dos tratamientos estadísticos: usted puede comparar los resultados promedio de varios grupos usando tests estadísticos o usted puede calcular un coeficiente de correlación para detectar una relación entre las características de los alumnos (por ejemplo, número de días de asistencia) y resultados (por ejemplo, rendimiento o actitud positiva).

5. Cuando evalúa un Diseño Antes -y- Después, desarrolle y ensaye muchos instrumentos que podría ser medidas sensitivas de los objetivos del programa. Tenga preparado a fines de año instrumentos bien producidos, probados en terreno y haga el esfuerzo posible para obtener al menos un grupo control no-equivalente para el año subsiguiente.

CAPÍTULO 6

ALGUNOS DISEÑOS MENOS BÁSICOS: USO DE ANÁLISIS DE VARIANZA

Los capítulos anteriores se han referido a los Diseños que suponen solamente dos grupos de tratamiento. El grupo-E y el grupo-C. Este capítulo introduce los diseños que pueden referirse a tres o más programas y que pueden examinar la influencia sobre los resultados de otros factores que actúan, además de la influencia de los programas. Además de indicar cómo se puede establecer un diseño, el capítulo lo prepara para conversar con quién vaya a analizar la información por medio de análisis de varianza (ANDEVA). Es importante consultar lo antes posible a una persona que esté familiarizada con investigación y estadísticas cuando intenta establecer un diseño complejo.

Suponga que un distrito tiene dos programas nuevos, el Programa A y el Programa B, los dos en matemática, que necesitan ser evaluados como alternativas al programa regular (Programa C). Suponga además que el personal del proyecto sospecha que los resultados podrían variar en el caso de cursos con diferentes características. Estas características de los cursos pueden ser algo como cursos de "estructura - abierta" en oposición a cursos "tradicionales", o cursos "sobre el promedio" en oposición a cursos "tradicionales", o cursos "sobre el promedio" en oposición a cursos "bajo el promedio". El evaluador del distrito necesita responder las siguientes preguntas de evaluación.

1. ¿Tienen los Programas A, B y C efectos significativamente diferentes? (por ejemplo ¿es uno mejor que los otros?)

2. ¿Depende completamente la efectividad de los tres programas de los tipos de cursos en que se implementan? Por ejemplo, ¿es un programa consistentemente mejor en cursos de estructura abierta, mientras que otro es mejor en cursos más tradicionales?

El evaluador tiene que medir los efectos de los dos FACTORES, cada uno de los cuales puede estar influenciando los resultados en matemáticas: Un "factor" es el tipo de programa, el otro "factor" es una característica del programa. La forma de establecer un diseño para descubrir los efectos de varios factores es poner los factores en ambos lados de un recuadro:

FACTOR 1
Tipo de Programa

FACTOR 2
Características del Curso

En seguida, usted escribe los "nivelas" de cada factor. Los "niveles" son las categorías para ese factor. En el ejemplo que se analiza, hay tres "niveles" por tipo de programa, es decir, Programa A, Programa B y Programa C. Hay dos "niveles" por característica de curso, es decir, cursos de estructura abierta y tradicional. Estos niveles se escriben en los lados apropiados del recuadro y se usan para dividirlo en "celdillas" como se muestra en la Figura 1.

Observe que debido a que hay tres niveles de un factor y dos de otro, nosotros tenemos un recuadro de 3 x 2 = 6 celdillas.

		Factor 1		
		TIPO DE PROGRAMA		
		Programa A	Programa B	Programa C
Factor 2 Características del Curso	Estructura abierta	Celdilla #1	Celdilla #2	Celdilla #3
	Estructura tradicional	Celdilla #4	Celdilla #5	Celdilla #6

Figura 1

Ahora la Figura 1 es la base para el diseño que se necesita para responder las preguntas planteadas (¿tienen los programas efectos diferentes? y ¿funcionan los programas en forma diferente en cursos de estructura abierta en oposición a cursos de estructura tradicional?). El paso siguiente es encontrar cursos para poner en cada una de las celdillas.

Idealmente, usted debería asignar al azar un número igual de cursos abiertos y tradicionalmente a cada uno de los tres programas. Suponga que había 12 cursos de estructura abierta y 21 cursos tradicionales en el distrito, entonces a usted le gustaría asignar aquellos 33 cursos de la manera siguiente:

FACTOR 1
Tipo de Programa

Factor 2 Características del Curso		Programa A	Programa B	Programa C	
	Estructura abierta	n = 4	n = 4	n = 4	12
	Estructura tradicional	n = 7	n = 7	n = 7	21
		11	11	11	33

Figura 2

Examinemos parte de la Figura 2. Los cuatro cursos que aparecen en la primera celdilla son cuatro cursos de estructura abierta que reciben el Programa A. En total, 11 cursos implementarán el Programa A: 4 cursos de estructura abierta y 7 cursos tradicionales. Verifique la Figura 3 y asegúrese saber lo que representan los números. Por ejemplo ¿qué es "33" en el vértice derecho inferior? Respuesta: el número total de cursos.

Recuerde que en la Figura 2 los cursos se asignaron al azar a las celdillas. A menudo, sin embargo, usted no puede hacer asignaciones al azar de manera muy fácil. A veces su evaluación simplemente debe funcionar con lo que está ahí, o usted debe invitar profesores para que ensayen nuevos programas. En tales situaciones, usted aún puede llenar las celdillas en el diseño, pero la interpretación será más difícil. Por ejemplo, si el programa tradicional, Programa C, entrega los mejores resultados, ¿es porque realmente es mejor? ¿o se explica por el hecho que los profesores que ya estaban obteniendo buenos resultados con el programa regular (Programa C) no querían ensayar nada nuevo, de tal modo que no ensayaron el programa A o B? Esto significaría que hubo una "subjetividad en la selección" que estaría influenciando los resultados. Los cursos que ensayaban los programas no eran equivalentes. Debido al método voluntario de asignación de programas, las personas que deseaban un cambio eligieron los Programas A 7 B. Aquellos que estaban en el grupo del Programa C eran cursos que tenían profesores satisfechos y probablemente ya muy efectivos.

Aunque llenar las celdillas utilizando un procedimiento que no sea al azar le da algunos problemas, aún le permite usar estadística para volver a verificar resultados al azar. Cuando se aplica el test estadístico (en el caso de la Figura 3, un análisis de varianza), las diferencias entre los puntajes promedios en las celdillas aún mostrarán si los factores no al azar han influenciado sus resultados. El problema, como se ha dicho, será mostrar que estos factores no al azar eran el programa y no algo ajeno como el temperamento del profesor o los horarios de clase de los alumnos.

Ahora suponga que usted fue capaz de llenar las celdillas del diseño al azar con tres cursos cada uno, tal como se muestra en la Figura 3. Los cursos se indican con los nombres de los profesores.

		FACTOR 1		
Factor 2		A	B	C
	Abierta	Jones Humphrey London	Ramírez Martin Reese	Smith Ford Davis
	Tradicional	Ross Brown Higgins	Phillips Johnson Graham	Robinson Fitzgerald Cohen

18

Figura 3

Hay tres cursos en cada celdilla, haciendo un total de 18 cursos. Supongamos que el programa se aplicó durante un año y que los 18 cursos, en verdad "permanecieron en sus celdillas", es decir, los cursos de estructura abierta siguieron siendo de estructura abierta; los cursos que supuestamente estaban recibiendo el Programa A, en realidad recibieron el Programa A, etc. Estas cosas podrían verificarse con sus actividades de documentación.

Al final del año, se da un postest a todos los cursos y se computan los puntajes promedio de cada uno de ellos. De este modo, en la primera celdilla, el curso de estructura abierta de Ms. Jones que recibe el programa A, podría tener un puntaje promedio de curso de 70.

Todo lo que usted necesita hacer ahora es dar los puntajes promedio de los 18 cursos a un analista de datos y pedirle que le entregue un análisis de de varianza (ANDEVA) para el diseño. Usted le dirá al analista "es un diseño de 3 por 2 con 3 observaciones por celdillas". "Los dos factores son programa (3 niveles) y estructura (2 niveles)".

La sección siguiente lo preparará para interpretar los resultados del ANDEVA

CÓMO INTERPRETAR UN ANÁLISIS DE VARIANZA

Observe la Figura 4. Suponga que en la primera celdilla, los puntajes medios del curso ("promedios del curso") fueran:

Jones	70
Humphrey	75
London	65

En ese caso, el puntaje medio para la celdilla es 70: $\frac{70 + 75 + 65}{3} = 70$

Este promedio de los puntajes en la celdilla se llama media aritmética de la celdilla. Suponga que todos los puntajes de cada celdilla se promedian para obtener todas las medias aritméticas de las celdillas y aquellas medias de las celdillas fueran las siguientes:

	A	B	C
Abierta	70	60	50
Tradicional	74	64	54

Figura 4

Ahora, sólo examinando aquellos resultados, piense cómo contestaría sus dos preguntas de evaluación:

1. ¿Tienen los programas efectos diferentes?

2. ¿Depende la efectividad de los programas del tipo de curso en que se implementan?

Mirando la primera fila (la de cursos "abiertos") los números 70, 60, 50, parece claro que el Programa A se asocia con mejores resultados que el programa B, y el B a su vez es mejor que el C. Mirando la segunda fila, cursos tradicionales, usted puede ver el mismo patrón de puntajes descendentes 74, 64, 54. Mirando las columnas hacia abajo, usted encontrará otro patrón consistente: los puntajes de los cursos tradicionales son ligeramente más altos que los puntajes de los cursos de estructura abierta sin excepción.

En la situación que se ilustra en la Figura 5, cada factor tiene un efecto consistente: cada ejemplo del Factor 1 (Programa A, B y C) origina alrededor de 10 puntos de diferencia, y cada uno de los dos ejemplos de Factor 2 (estructura de curso) hace alrededor de 4 puntos de diferencia ¿Pero son estas diferencias significativas? Esto es lo que le dice el análisis de varianza. La tabla ANDEVA que su analista de información puede proporcionarle, podría ser como la siguiente.

Fuente	gl	sc	f
Programa	2		*
Estructura	1		
Programa y estructura	2		
Error	12		
Total	17		

* p< .05

Figura 5. Tabla de Análisis de Varianza para la Información de la Figura 4

Examinemos esta tabla. Bajo el título "Fuente", se ponen en una lista los factores del diseño = tipo de PROGRAMA Y ESTRUCTURA de curso. "fuente" quiere decir "fuente de variación" y le dice qué factor se está probando. En la primera fila de la tabla se prueba la significancia del factor programa.

Sigamos con la primera columna. Hay un 2 bajo "gl". Esto significa que los grados de libertad (gl) para el primer factor son 2. Los grados de libertad son siempre uno menos que el número de niveles para el factor. En este caso había 3 "niveles" de programa, de tal modo que gl es 2. Usted no tiene que preocuparse por el significado de esto, pero es útil verificar dichas cifras para asegurarse que el analista de información no haya cometido serios errores, "SC" quiere decir "Suma de Cuadrados", algo un poco complicado para este análisis. Mire la última entrada en la fila, bajo "F". Este número es el que proporciona el test de significancia. En este caso, se ha determinado ser significativo (al nivel .05). Esto le dice que el factor del programa hizo una diferencia significativa en los resultados.

Verificando la segunda fila, el factor estructura, usted puede ver que el valor F no fue significativo; no se le ha puesto un asterisco. (Si su analista de información no indicó significancia con asteriscos, pídale que lo haga. Es una práctica común usar asteriscos y facilita la lectura de las tablas). Esto le dice que las diferencias entre cursos de estructura abierta y cursos tradicionales pudieron haber ocurrido por muestreo al azar. Las diferencias no son lo bastante grandes como para ser consideradas significativas. Sin embargo, este es un problema lateral; la diferencia entre cursos de estructura abierta y cursos tradicionales no fue una de las preguntas de su evaluación.

Si usted recuerda, estas preguntas de evaluación fueron:

1. ¿Tienen los programas A, B y C efectos significativamente diferente?

2. ¿Depende completamente la efectividad de los tres programas de los tipos de cursos en que se implementan?

La primera fila de la tabla ANDEVA respondió la primera pregunta de evaluación: Sí, los programas tienen efectos significativamente diferentes.

Ahora para la segunda pregunta: ¿Dependen los efectos de los programas del tipo de estructura del curso? Esta pregunta se responde mirando la tercera fila de la tabla

ANDEVA donde programa y estructura es la fuente de variación. El hecho de que el valor F en la tercera fila no sea significativo quiere decir que cualesquiera que sean los efectos que tuvo el programa, estos efectos no variaron con el tipo de estructura de curso. En verdad, esto ya se ve claro simplemente mirando los puntajes promedio en la Figura 4. Muestra el mismo patrón de puntajes para cursos abiertos y tradicionales cuando se lee a través de los programas (74, 64, 54 decrece en forma consistente tal como lo hace 70, 60, 50).

Esta tercera fila se llama término de interacción ("interacción de primer orden" para ser preciso). Es esta fila la que debería examinarse primero porque si el término de interacción es significativo, usted no puede hacer una afirmación general tal como "el programa A es mejor que el programa B o C". Si hay una interacción significativa, usted tendrá que hacer afirmaciones separadas sobre los efectos del programa: una afirmación sobre los efectos en cursos de estructura abierta y otra sobre los efectos en cursos tradicionales.

El análisis siguiente ilustra las diferencias entre situaciones en que hay interacción y situaciones en que no hay, comparando los resultados recién discutidos con un conjunto semejante en que hay una interacción significativa entre los dos factores.

INTERACCIÓN NO SIGNIFICATIVA

TABLA DE MEDIAS DE CELDAS

		Factor 1		
Factor 2		A	B	C
	Abierta	70	60	50
	Cerrada	74	64	54

GRÁFICO DE MEDIAS DE CELDAS

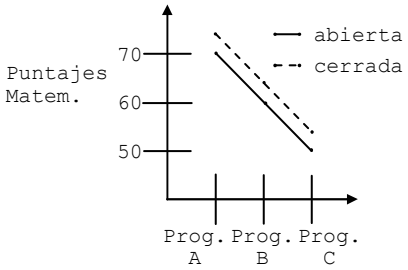


TABLA ANDEVA

Fuente	gl	SC	F
Programa	2		[3]
Estructura	1		
Programa y estructura	2		
Error	12		
Total	17		

INTERPRETACIÓN

1. Hay diferencias significativas entre los tres programas, A, B y C.
2. Los efectos de los programas no están significativamente influenciados por la estructura del curso.

[3] $p < 0.05$

INTERACCIÓN SIGNIFICATIVA

TABLA DE MEDIAS DE CELDAS

		Factor 1		
Factor 2		A	B	C
	Abierta	70	60	50
	Cerrada	74	64	74

GRÁFICO DE MEDIAS DE CELDAS

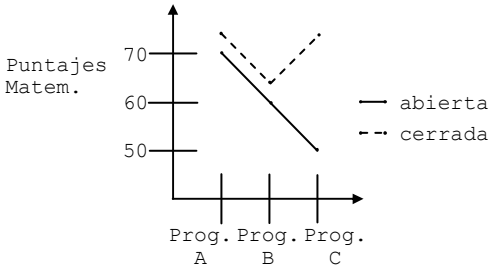


TABLA ANDEVA

Fuente	gl	SC	F
Programa	2		
Estructura	1		[4]
Programa y estructura	2		[4]
Error	12		
Total	17		

INTERPRETACIÓN

1. Hay diferencias significativas entre los tres programas, A, B y C.
2. Los efectos de los programas dependen del tipo de estructura de curso. Parece que el Programa A es el mejor programa para los cursos de estructura abierta, pero cualesquiera de los programas, el A o el C, es mejor para cursos tradicionales.

[4] $p < 0.05$

Para ser estrictos, usted debería seguir buscando una diferencia significativa con ANDEVA comprobando con su analista de datos si la diferencia entre cualesquier par de programas es significativa*. Por ejemplo, si los efectos del programa son significativos y las medidas de resultado son:

Programa A	Programa B	Programa C
70	62	58

bien puede ser que la diferencia entre los programas B y C no sea significativa; solamente el programa A es significativamente diferente del programa C o quizás de ambos, del programa B y C. Por lo general, usted no confiará en un solo conjunto de resultados, de tal modo que tal estrictez no es esencial. Sólo tenga cuidado de no poner mucha fe en pequeñas diferencias y recuerde que el ANDEVA dice que hay una significancia en alguna parte del factor, pero no dice dónde.

Resumen

A continuación se encuentra un resumen paso a paso de un procedimiento para establecer y analizar un diseño de análisis de varianza. Este es un diseño en que no solamente está el factor de programas diferentes, sino también algún otro factor tal como una característica de curso.

1. Escriba los factores en dos lados de un rectángulo

FACTOR II

FACTOR I

2. Llene los niveles de cada factor y luego divida el rectángulo en "celdillas", por ejemplo, un diseño 3 x 4:

FACTOR 2

FACTOR 1		Nivel 1	Nivel 2	Nivel 3	Nivel 4
	1				
	2				
	3				

3. Asigne o ubique "casos" por cada celdilla. Los "casos" pueden ser alumnos, cursos o colegios, depende de la "unidad de análisis".

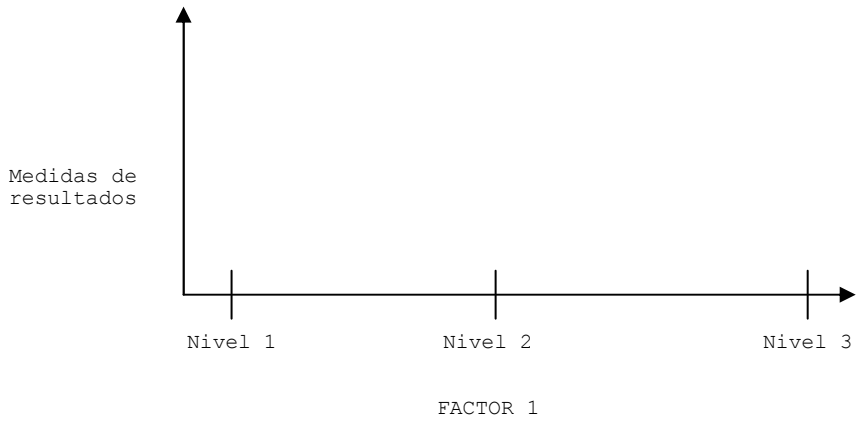
4. Reúna información en forma de puntaje sobre la medida de resultado para todos los casos en cada celdilla.

5. Dé la información al analista de datos, pidiendo un análisis de varianza (ANDEVA) para probar la significancia de los efectos de los factores y la

* *Este análisis se llamo una "comparación post-hoc".

significancia de las interacciones. Asegúrese de decirle que usted quiere tener una tabla de medias de celdas como también una tabla de ANDEVA.

6. Cuando reciba la información, prepare un gráfico con las medidas de resultado en el eje vertical y los niveles de un factor en el eje horizontal.



Grafique los puntajes para el nivel 1 del segundo factor y júntelos. Repítalo para otros niveles del segundo factor.

7. Examine la tabla ANDEVA. Primero compruebe la fila de interacción. ¿Es el valor F significativo? Si es así, hay diferencias significativas entre los resultados del programa, dependiendo de qué nivel del otro factor examina usted. Compruebe el gráfico de medias de celdas e informe los efectos del programa en forma separada para cada nivel del otro factor. Si la interacción F no es significativa, compruebe los efectos generales, efectos que son los mismos independientemente del nivel del otro factor, e informe estos.

8. Si usted tiene dudas sobre qué pares de puntajes son diferentes, pida al analista de datos que haga comparaciones "post-hoc".

CAPÍTULO 7

CÓMO ALEATORIZAR

La aleatorización le será útil en cualesquier momento que quiera asegurar igualdad o representatividad entre grupos de personas. Varios de los diseños de evaluación que se analizan en este libro cuentan con asignación aleatoria de alumnos o cursos a los programas. La asignación aleatoria, o al azar, le permite presumir que los grupos a quienes se les asignarán diferentes programas son inicialmente semejantes*. Saber esto le ayuda a extraer conclusiones lógicas sobre las causas de los resultados que usted mide eventualmente.

A menudo podría usted desear que un pequeño grupo de alumnos, o un test corto, representara a uno más grande. Una representación adecuada puede asegurarse mejor, por medio de una selección al azar de representes del grupo más grande.

Este capítulo da instrucciones paso a paso, tanto para: (1) la asignación aleatoria desde una población total de alumnos o cursos a dos a más grupos, como para (2) la selección aleatoria de una muestra, pequeña, pero muy representativa de un grupo grande.

Asignación aleatoria de alumnos a los programas

Una asignación es "aleatoria" si cada alumno tiene la misma probabilidad de ser elegido. Por definición, la asignación aleatoria es una asignación justa y objetiva de alumnos a diversos grupos. Este es un punto que debe tener presente si usted tiene dificultad en persuadir a las personas para que estén de acuerdo con la asignación aleatoria.

Hay por lo menos cuatro formas para lograr la asignación al azar de alumnos a los programas. Cuál es el mejor método a usar, depende de la situación en que se necesita asignación al azar:

1. Cara o sello. Suponga que los alumnos van a ser asignados al azar a uno de dos programas. Usted podría preparar una lista de todos los alumnos susceptibles de ser elegidos. Tirar una moneda al cara o sello por alumno cada vez (dando por hecho que es una moneda regular) podría ser un método de asignación aleatoria: "Los "cara" van al Programa X, los "sello" al Programa C". Pero si los alumnos van a ser asignados al azar a Tres grupos, el procedimiento cara o sello sería engorroso. Puesto que tirar monedas al cara o sello toma bastante tiempo, es impracticable, excepto en el caso de un pequeño número de alumnos.

2. Tablas de números aleatorios. Los investigadores a menudo usan tablas de números aleatorios como uno de sus métodos. Estas se encuentran al final de muchos libros de estadística. Las tablas de números aleatorios se generan, comúnmente, con

* Del mismo modo, si usted quiere construir dos tests (digamos usar como un pretest y postest) que tengan ítemes ligeramente diferentes pero que midan las mismas destrezas, usted puede asegurar tests similares adecuados asignando ítemes al azar a las dos formas del test tomados de un "pool" grande de ítemes. Estos se llaman técnicamente "formas de contenido equivalente", y aunque todos los ejemplos de este capítulo se refieren a asignación y selección de personas, también son aplicables a ítemes de tests.

computadores y se componen de listas de los numerales 0 a 9, los que son verdaderamente al azar. Son útiles para todas las situaciones aleatorias, pero también han mostrado ser algo engorrosas en la práctica.

3. Baraja de cartas. Usted mismo puede hacer una "baraja práctica de aleatorización" (BPA) numerando 75 cartas de naipe "1" a "75" con un lápiz de marcar negro. (Se sugieren 75 cartas puesto que este número cubrirá la mayoría de las situaciones aleatorias que usted encuentra y aun añade la ventaja de que fácilmente puede barajarse). La BPA hace que los procedimientos de asignación al azar sean comprensibles y fáciles de llevar a cabo. Cada carta puede representar un alumno para que se asigne a un programa… o un ítem que se dé en un test… o cualesquier cosa que se va a asignar o seleccionar al azar. En las páginas *120, 122* y *123* se dan instrucciones para varios procedimientos de asignación al azar que usan la BPA.

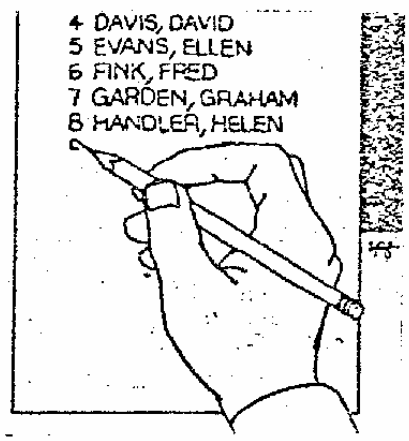
4. Una lista preparada. Las páginas *125* y *126* de este capítulo se titulan "Listas de asignación aleatoria a 2, 3, 4 o 5 grupos". Estas listas se han preparado por medio de aleatorización generada por computador. Para usar dicha página con el objeto de formar grupos, digamos 4, simplemente escriba los nombres de los alumnos en orden alfabético en la lista y lea luego, desde la columna "grupo 4" el número del grupo al cual se asigna cada alumno. Un procedimiento para usar la lista preparada aparece con más detalle en la página *123*.

Instrucción paso a paso para usar la Baraja Práctica de Aleatorización (BPA) para la Asignación Aleatoria Simple.

Los dos procedimientos siguientes constituyen "asignación aleatoria simple", aquellos en que no se toma en cuenta ninguna característica de los alumnos. Ellos asegurarán grupos ligeramente equivalentes, especialmente en el caso en que están involucrados números grandes (más de 30 por grupo). Representan buenos métodos para aleatorizar, lo que unos pocos escépticos cuestionarán. Si usted sabe algo sobre las características de los alumnos o cursos, usted puede asegurar más la equivalencia de los grupos "estableciendo bloques" o "equiparando" estas características antes de hacer las asignaciones al azar. Estos procedimientos se describen en las páginas *126* y *128*, respectivamente.

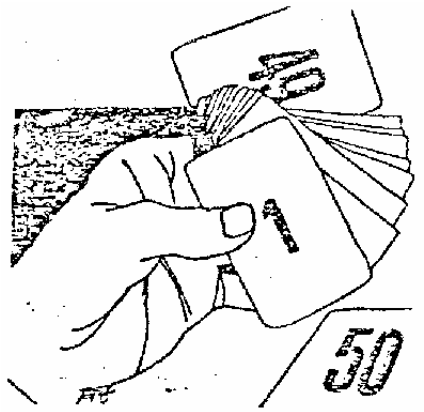
El método BPA de aleatorización descrito a continuación se refiere a asignación de alumnos. También puede aplicarse a asignación de cursos o colegios a programas. La lista descrita en el paso 1 sería entonces una lista de curso (por nombres de profesores o número de sala, etc.) o una lista de colegios.

EJEMPLO: Cuarenta y nueve alumnos se van a asignar al azar a tres programas de lectura, X, Y y Z.

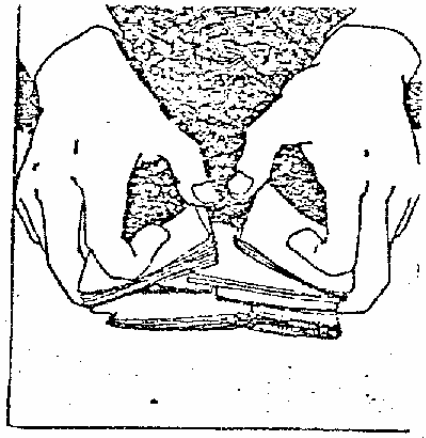


Haga una lista en orden alfabético de los 49 alumnos. Cualesquier orden conveniente bastará, pero generalmente es más útil mantener listas alfabéticas).
Numere cada nombre en la lista del "1" al "49".

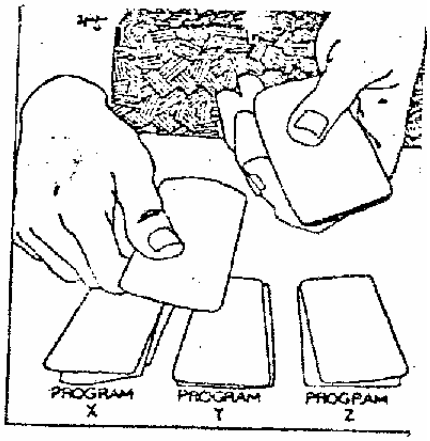
2.



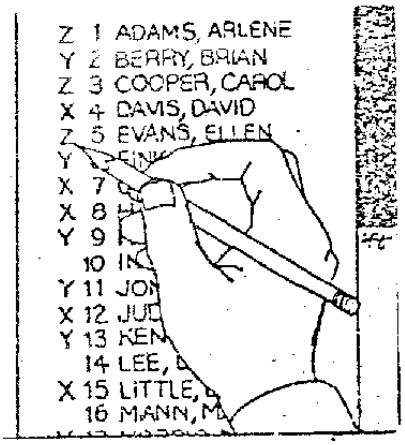
Saque las cartas "1" al "49" de la BPA.



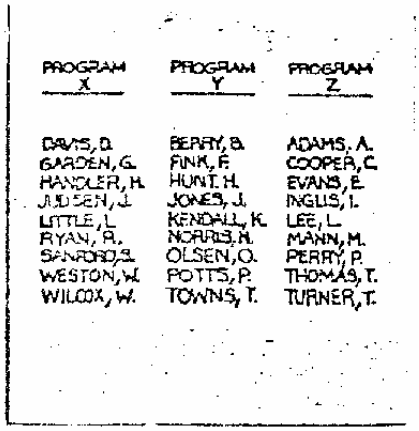
3. Baraje bastante las 49 cartas. Baraje y *corte* varias veces.



4. Ponga las 49 cartas en tres montones, uno para el Programa X, uno para el programa Y y uno para el programa Z (ponga cuadrados del tamaño de las cartas si lo estima necesario). Ponga una carta a X, otra a Y, una a Z, y repita hasta que las 49 cartas se hayan repartido. Un grupo resultará tener un alumno más que los otros. Esto está bien.

5. 

Tome el montón de cartas para el programa X. Mire el número de la primera carta y escriba una "X" junto al nombre del alumno con ese número en la lista. Continúe marcando "X"s para todos los alumnos cuyos números están en el montón X. Tome el montón para el programa Y y marque "Y"s junto a estos alumnos. Haga lo mismo para el montón Z.

6. 

PROGRAM X	PROGRAM Y	PROGRAM Z
DAVIS, D.	BERRY, B.	ADAMS, A.
GARDEN, G.	FINK, F.	COOPER, C.
HANDLER, H.	HUNT, H.	EVANS, E.
JUDSEN, J.	JONES, J.	INGLIS, I.
LITTLE, L.	KENDALL, K.	LEE, L.
RYAN, R.	NORRIS, N.	MANN, M.
SANFORD, S.	OLSEN, O.	PERRY, P.
WESTON, W.	POTTS, P.	THOMAS, T.
WILCOX, W.	TOWNS, T.	TURNER, T.

Copie a máquina su lista original haciendo tres listas separadas con los nombres de los alumnos en cada programa. Si su lista original estaba en orden alfabético, las tres listas pueden copiarse a máquina en orden alfabético. Si la lista original no estaba en orden alfabético, quizás usted quiera pedir al secretario que ponga en orden alfabético los alumnos en la lista X, Y y Z, en forma separada.

Este mismo procedimiento puede usarse cuando quiera que usted asigne alumnos - no importa cuantos grupos pueda haber. Si usted tiene más de 75 alumnos para asignar a los grupos, use esta simple ampliación del procedimiento.

Haga la aleatorización como se describió para los primeros 75 alumnos. Luego usando la lista de alumnos desde el alumno número 76 adelante, comience otra vez en el 1 para numerar a los alumnos. Seleccione tantas cartas como necesite para representar a los alumnos que quedan. (si usted tiene 125 alumnos en total, usted usará tarjetas numeradas del "1" al "50"). Asegúrese, cuando termine, que el número de alumnos que usted ha asignado a cada programa es ligeramente el mismo. Hacer la aleatorización en dos (o más) fases como ésta producirá la asignación al azar.

Asignación Aleatoria realizada en Público utilizando la BPA

A veces es necesario o deseable hacer público el procedimiento de asignación aleatoria. Por ejemplo, si un curso se va a dividir en dos a más grupos, pudiera ser que los profesores quisieran que los alumnos se dieran cuenta que los grupos están formados aleatoriamente. Saber esto evitará cualesquier sentido de favoritismo o un sentimiento de haber sido especialmente elegido para un programa, lo que pudiera influenciar a los alumnos.

Aquí hay un método sugerido:

1. Reúna en una sala a los alumnos que se van a asignar a grupos. Dé a cada uno una tarjeta de registro numerada (numerada de acuerdo al número de alumnos

que Ud. tiene) sobre la cual escribirán su nombre y cualesquier otra información que usted pueda necesitar, tal como el número de la sala donde hacen sus reuniones de consejo de curso, edad, etc.

2. Explique la necesidad de hacer asignaciones aleatorias. Usted pudiera señalar YA SEA que hay un problema en cuanto a que no hay materiales en número suficiente, lo que hace necesario asignar aleatoriamente un pequeño grupo, 0 que usted realmente no tienen idea qué programa es mejor. Cada programa es algo nuevo y solamente los resultados mostrarán lo que es mejor.

3. Seleccione tantas cartas de la BPA como alumnos tenga. Los números deberían corresponder a las tarjetas de registro numeradas que están en poder de los alumnos.

4. Baraje las cartas y repártalas en tantos montones. En realidad, usted pudiera querer que un alumno haga esto, sólo para quitar cualesquier temor de "que el naipe esté arreglado".

5. Recoja el primer montón de cartas y lea los números. Los alumnos que tengan estos números en sus tarjetas de registro deberían entregar sus tarjetas. Este montón de tarjetas de registro se pone entonces en orden alfabético y en ese momento usted tiene un registro de alumnos listo para ese programa.

   Repita este procedimiento para cada programa.

Instrucciones paso a paso para usar la Lista Preparada de Números Aleatorios para una Asignación Aleatoria Simple.

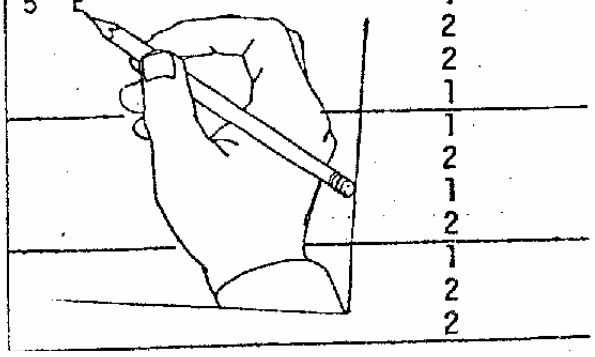
El método de aleatorización que se describe a continuación se refiere a asignación de alumnos. También puede servir para asignar cursos o colegios a programas. La lista que se describe en el paso 1 sería entonces una lista de cursos (por nombre de profesor o número de sala, etc.) o una lista de escuelas.

EJEMPLO: Cuarenta y nueve alumnos se van a asignar al azar a tres programas de lectura, X, Y y Z.

1. Haga una copia de las páginas 125 y 126 de este libro… la lista para asignar al azar a 2, 3, 4 o 5 grupos.

2.

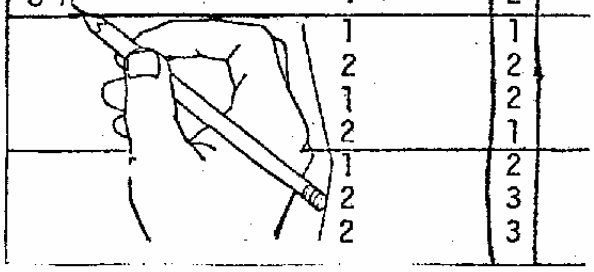
	Nombre del Alumno	2 Grupos
1	Adams, Arlene	2
2	Berry, Brian	2
3	Cooper, Carol	2
4	Davis, David	2
5	E	1
		2
		2
		1
		1
		2
		1
		2
		1
		2
		2



Haga una lista de los nombres de <u>todos</u> los alumnos, en orden alfabético.

3.

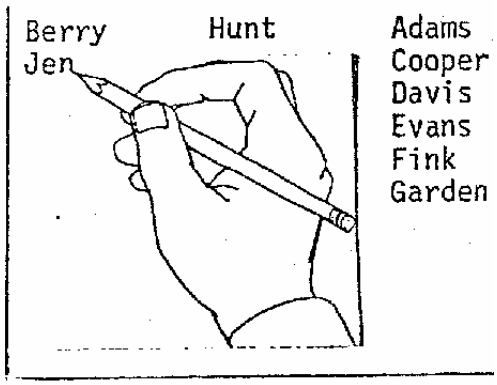
	Nombre del Alum.	2 Grupos	3 Grupos
1	Adams, Arlene	2	3
2	Berry, Brian	2	1
3	Cooper, Carol	2	3
4	Davis, David	2	3
5	Evans, Ellen	1	3
6	Fink, Fred	2	3
7	Garden, Graham	2	3
8	H	1	2
		1	1
		2	2
		1	2
		2	1
		1	2
		2	3
		2	3



Decida cuántos grupos quiere formar usted... 3 en este caso… (correspondiendo al número de programas alternativos que usted desea comparar) y encierre esta columna a lo largo.

4.

Programa X (1)	Programa Y (2)	Programa Z (3)
Berry	Hunt	Adams
Jen		Cooper
		Davis
		Evans
		Fink
		Garden



Mire el número que corresponde a cada nombre de los alumnos asignados a cada grupo. Estos son los grupos al azar.

TABLA 6

LISTA PARA ASIGNAR AL AZAR A 2, 3, 4 o 5 GRUPOS

Nombre del Alumno	2 Grupos	3 Grupos	4 Grupos	5 Grupos
1	2	3	1	3
2	2	1	1	5
3	2	3	3	2
4	2	3	3	4
5	1	3	1	4
6	2	3	4	2
7	2	3	3	3
8	1	2	2	2
9	1	1	3	4
10	2	2	4	3
11	1	2	1	1
12	2	1	1	4
13	1	2	3	3
14	2	3	3	5
15	2	3	4	3
16	1	1	3	3
17	1	1	1	1
18	1	1	2	4
19	2	3	2	5
20	1	1	1	4
21	2	3	2	2
22	1	2	3	2
23	1	3	2	1
24	1	1	1	5
25	1	3	3	4
26	1	2	3	4
27	2	3	4	4
28	2	1	4	5
29	2	1	3	5
30	1	1	4	1
31	2	3	1	5
32	1	1	1	4
33	1	1	2	5
34	1	2	4	4
35	2	2	3	1
36	2	1	4	5
37	1	1	1	3
38	1	2	4	2
39	2	3	2	2
40	1	1	3	4
41	2	2	4	2
42	1	3	2	3
43	2	1	1	2
44	1	1	1	3
45	2	2	2	3
46	2	2	1	3

47	1	2	2	5
48	1	1	3	1
49	2	1	1	2
50	2	1	3	1
51	1	2	4	3
52	1	3	2	2
53	1	3	4	3
54	2	1	4	1
55	1	1	4	5
56	1	3	3	3
57	1	2	2	4
58	1	2	3	2
59	2	3	2	2
60	1	1	1	1
61	1	3	1	1
62	1	1	4	3
63	2	2	4	5
64	2	2	4	1
65	1	2	3	3
66	1	3	3	3

Formación de Bloques

Los procedimientos aleatorios descritos en las secciones anteriores se llaman aleatorización simple. Ellos tratan de formar grupos ligeramente equivalentes a partir de una serie de alumnos o cursos acerca de quienes nada... al menos nada que probablemente afecte los resultados de un programa... se conoce. Cuando se conocen características importantes, éstas deberían usarse para influenciar la asignación aleatoria. Tanto la "formación de bloques", como la "formación de grupos equiparados" permiten esto, asegurando una distribución de grupos equiparados" permiten esto, asegurando una distribución igual de los factores importantes, tales como un cuociente intelectual (CI) alto.

La formación de Bloques o aleatorización por bloques es un método que se emplea para asegurar más el grado de comparabilidad posterior de grupos de comparación. Los procedimientos de "formación de bloques", le permiten detectar aquellos rasgos de los alumnos que pudieran afectar los resultados de los programas y aseguran que dichos rasgos se distribuyan de la misma manera en los grupos que usted construye.

Se llama formación de bloques porque antes que la aleatorización tenga lugar se pone a los alumnos en lista, en "bloque". Por ejemplo, debería haber bloques de alumnos del sexo femenino y masculino o de alumnos de alta o baja habilidad. Los bloques pueden reflejar cualesquier factor que es probable que afecte su desempeño en el programa. Luego, el grupo-E y el grupo-C se forman por selección aleatoria dentro de cada bloque.

La formación de Bloques es el procedimiento recomendado en cualesquier asignación al azar, donde los rasgos relevantes de los alumnos pueden identificarse y premedirse. Es prácticamente esencial donde hay grupos pequeños, menos de 15 por programa. Con un número de alumnos tan limitado, la asignación aleatoria simple puede producir grupos que son complemente diferentes entre sí.

Por ejemplo, suponga que seis alumnos tenían los siguientes puntajes promedio:

Tomás	A*
Ricardo	D
Enrique	D
Minerva	D
Julio	A
Diana	D

Una asignación aleatoria simple a dos grupos podría poner a Julio y Tomás en el mismo grupo, lo que haría que estos dos grupos fueran disparejos: uno con notas A, A y D; el otro con D, D y D. Mientras más probable es que usted encuentre problemas de este tipo.

La solución a este problema específico se logra mediante el uso de PARES EQUIPARADOS, fundamental en los procedimientos de bloques. Si ya hay disponibilidad, o se pueden obtener fácilmente algunas notas o puntajes de los alumnos, se forman pares equiparados (si se van a formar dos grupos). Los miembros de estos pares se asignan entonces aleatoriamente a los programas. En el ejemplo que se da arriba, Tomás y Julio serían un par equiparado de acuerdo a sus notas. Podría usarse un procedimiento de cara o sello para asignar a uno al grupo experimental y al otro al grupo control. Este proceso asegura una "equivalencia" más exacta entre los dos grupos.

Un procedimiento estándar para formar bloques

El procedimiento básico para formar bloques es un poco más complejo que la simple asignación al azar. En esta sección se presentará un esquema paso a paso del procedimiento estándar para la formación de bloques. Para detallar el procedimiento, se describirá la situación frecuente de formación de bloques por sexo.

Usted puede, por supuesto, querer hacer un bloque en torno a otra característica que no sea sexo. Altere el procedimiento de acuerdo a sus necesidades, pero siga los pasos preliminares descritos a continuación:

1. Decida la característica que usted desea usar para formar los bloques... cuociente intelectual (CI), rendimiento en matemática, destreza en lectura, etc. Ésta será la característica que usted está más interesado en distribuir en forma pareja entre los grupos y la que usted siente que podría causar resultados diferentes en programas alternativos, si no se distribuye igualmente entre los grupos. Si, por ejemplo, su o sus programas nuevos se refieren a matemática, puede ser que usted quiera asegurar la distribución en forma pareja de los rendimientos anteriores en matemática.

* *La escala de notas en USA es la siguiente:
A-B-C-D-F, siendo "A" la nota máxima de aprobación y "F" la nota de reprobación (en algunos lugares la nota "D" también indica reprobación).

2. Determine una medida de aquella característica que se ha administrado o que podría administrarse a los alumnos. Esta medida formará la base para la formación de bloques. Nuevamente, si usted desea igualar el rendimiento en matemática, use puntajes de tests de rendimiento en matemática, o el promedio de las notas en matemática del año anterior como base para la formación de bloques. Del mismo modo, si usted quiere que los grupos sean iguales en habilidades generales, use puntajes CI. Si usted no dispone de alguna medida y no puede administrar ninguna, entonces haga que los profesores ordenen a los alumnos de acuerdo a las apreciaciones informales que ellos mismos tienen de sus características.

3. Administre la medida si usted encuentra que aún no tiene los resultados que necesita.

4. Haga una lista de los alumnos ordenándolos según los puntajes que obtienen en la medida de formación de bloques.

5. Decida cuántos "bloques" desea usted definir basados en esta característica. Sexo, por ejemplo, obviamente le da a usted sólo dos categorías, pero CI podría asignar tres o cuatro bloques: BAJO, MEDIANO Y ALTO; o BAJO, LIGERAMENTE BAJO EL PROMEDIO, LIGERAMENTE SOBRE EL PROMEDIO, y ALTO etc. Para medidas CONTÍNUAS, como CI, rendimiento, y nivel socioeconómico, USTED tendrá que decidir el número de categorías para formar bloques. En general, usted debería limitar el número de bloques a no más de tres o cuatro. Elija el número de categorías que mejor parece proporcionar tipos semejantes de alumnos en cada bloque. Si los puntajes parecen bien distribuidos, seleccione tres o cuatro

6. Divida la lista en el número de bloques que decidió formar.

7. Asigne aleatoriamente alumnos de cada bloque a los grupos de tratamiento.

Pares Equiparados

Si se van a asignar alumnos a un grupo-E o a un grupo-C, uno podría llegar al extremo de crear tantos bloques, que solamente dos alumnos quedaran en cada uno. Este extremo en verdad es una buena idea. Es el método de "pares equiparados". De estos pares de "iguales" en la característica importante, los alumnos se asignan al azar, uno al grupo-E y uno al grupo-C.

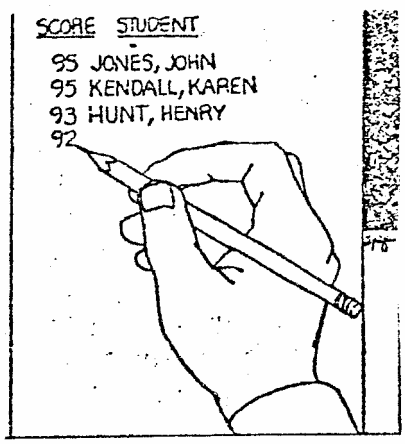
El procedimiento de pares-equiparados es más útil cuando usted evalúa un programa a corto plazo. Un programa corto es probable que muestre solamente una diferencia pequeña en resultados entre el grupo experimental (Programa X) y el grupo control (Programa C). Equiparar hace que los diferentes grupos que reciben cada programa sean inicialmente lo más parecido posible, proporciona un diseño poderoso y se emplea el riguroso pero muy acreditado "test-t de grupos pareados" para probar la significancia de las diferencias en resultados.

Hacer pares equiparados no es aconsejable para evaluar programas que durarán más de 2 meses (o unas pocas semanas más, si usted tiene 50 o más alumnos por programa). La razón para esto es el problema llamado "atrición", es decir, alumnos

que abandonan o tienen largas ausencias durante el curso del programa. Los análisis estadísticos para pares equiparados dependen de ambos miembros del par que está siguiendo el programa. Cuando un miembro de un par abandona o está ausente en forma excesiva, la información del otro miembro del par también debe eliminarse. Por esto usar pares equiparados no es útil cuando se evalúa un programa a largo plazo. Significa correr el riesgo de perder información sobre muchos alumnos puesto que los desertores y ausencias tienden a aumentar durante un trayecto largo. Si cada programa abarca sólo unos pocos alumnos al comienzo, el riesgo es enorme. Además, usted debe recordar que la razón principal para usar pares equiparados es maximizar la probabilidad de obtener significación estadística de un efecto pequeño. Un programa que funciona 6 semanas o más es de esperar que muestre un efecto más bien considerable, lo que eliminaría la necesidad de dicha equiparación cuidadosa.

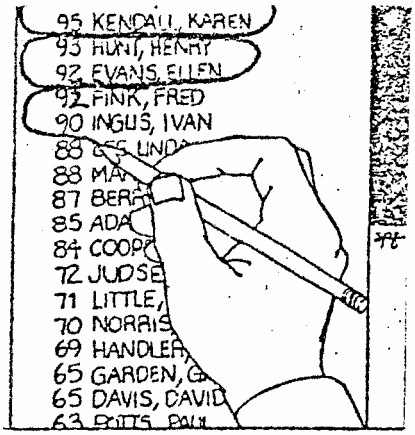
Instrucciones paso a paso Para Formar los Grupos E y C basándose en Pares Equiparados.

1.



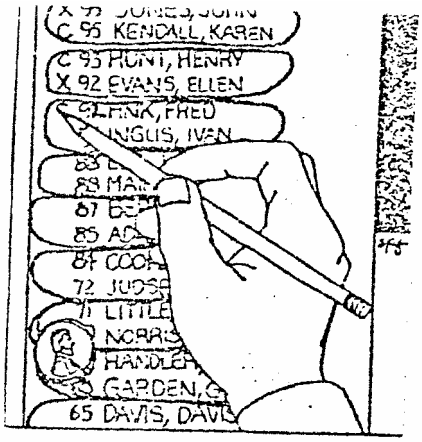
Recoja los puntajes de los alumnos con una medida adecuada y ORDÉNELOS DE MAYOR A MENOR, anotando en una lista el nombre de cada alumno en orden, desde el más alto al más bajo. La medida más apropiada es la que mejor pronostica resultados finales... por ej. un pretest de rendimiento en matemática para un programa en matemática.

2.



Haga pares "equiparados": Encierre en un círculo pares de alumnos que aparecen seguidos en la lista. Puesto que no se pueden conseguir pares de alumnos exactamente equiparados, este procedimiento simplemente produce los pares más cercanos posibles. Si hay un número impar de alumnos, debería elegirse al azar un nombre de la lista y eliminarse del análisis (pero el alumno siempre puede recibir el programa).

3.



Puesto que usted está formando dos grupos usted puede usar cara o sello para hacer asignaciones al azar. Por cada par tire al cara o sello una vez. Decida cuál es sello -el Programa X por ejemplo-, luego tire al cara o sello para el primer alumno anotado en el par. Si sale sello, escriba una "X" junto al nombre del primer alumno. Si es cara, marque una "C" junto a su nombre. Asigne el otro miembro del par al otro programa.

4.

Haga listas separadas, escritas a máquina, de los alumnos en cada programa; pudiera ser que usted quisiera esas listas en orden alfabético.

El método de grupo intermedio para formar un verdadero grupo control en programas especiales.

Cuando se va a seleccionar un grupo de alumnos para, por ejemplo, un programa de recuperación en lectura, los más necesitados, por lo general, se identifican claramente. Sus puntajes en tests de lectura son invariablemente los más bajos o hay un acuerdo total entre sus profesores de que ellos son los peores en lectura. Sin embargo, a medida que usted se acerca al punto de división, las decisiones que usted tiene que tomar son por lo general menos claras. Por ejemplo, Juan obtuvo un puntaje más alto que David en el último test, pero por lo general él obtiene puntajes más bajos. ¿Cuál de los dos niños debiera ponerse en el programa si no hay lugar para los dos?

Esta indecisión es muy justificable. Los tests no constituyen mediciones perfectas. A menudo hay esfuerzo en los puntajes de un test y por lo tanto no se puede confiar en distinciones muy finas. Juan puede haber obtenido un puntaje de 34 mientras que David obtuvo 32, pero David todavía puede ser el mejor en lectura. Sin embargo, usted debe considerar a Miguel quien teniendo un puntaje de 21 se ve claramente que necesita el programa.

De este modo, cuando los planificadores empiezan a seleccionar a aquellos que necesitan el programa, deben elegir de entre un gran grupo de alumnos a aquellos que

claramente lo necesitan, y decidir entre un grupo intermedio, quiénes deberían y quiénes no deberían estar en el programa. (Los psicometristas definirían con precisión el grupo límite usando el error estándar de medición del test; pero, esto no es necesario para los fines de este libro).

Un procedimiento necesario y valioso al evaluar un programa de recuperación es hacer todo el estudio de evaluación centrándose en los alumnos intermedios, formando un grupo Programa X y un verdadero grupo control de alumnos pareados de Este grupo. Suponga que la clase de recuperación en lectura tiene vacantes para 28 alumnos. Los 20 alumnos con mayores problemas en lectura se incorporan al programa. Los otros 16 alumnos forman el grupo intermedio. Una selección aleatoria de 8 de estos alumnos se incorpora también al programa; los otro 8 no van al programa de recuperación. (Vea la figure 6, a continuación, que ilustra cómo una distribución de puntajes en lectura puede usarse para identificar los más necesitados y los grupos intermedios).

De este modo, la asignación al azar de parte de los alumnos del grupo intermedio al programa nuevo (X) y parte a un grupo de comparación o grupo control (C) le permitirá probar la efectividad relativa de su programa nuevo SIN DESCUIDAR EL GRUPO DE NIÑOS A QUIENES USTED SE SIENTE MORALMENTE MÁS OBLIGADO A AYUDAR. Estos alumnos más necesitados recibirán el programa, pero ellos no tomarán parte en el estudio de evaluación.

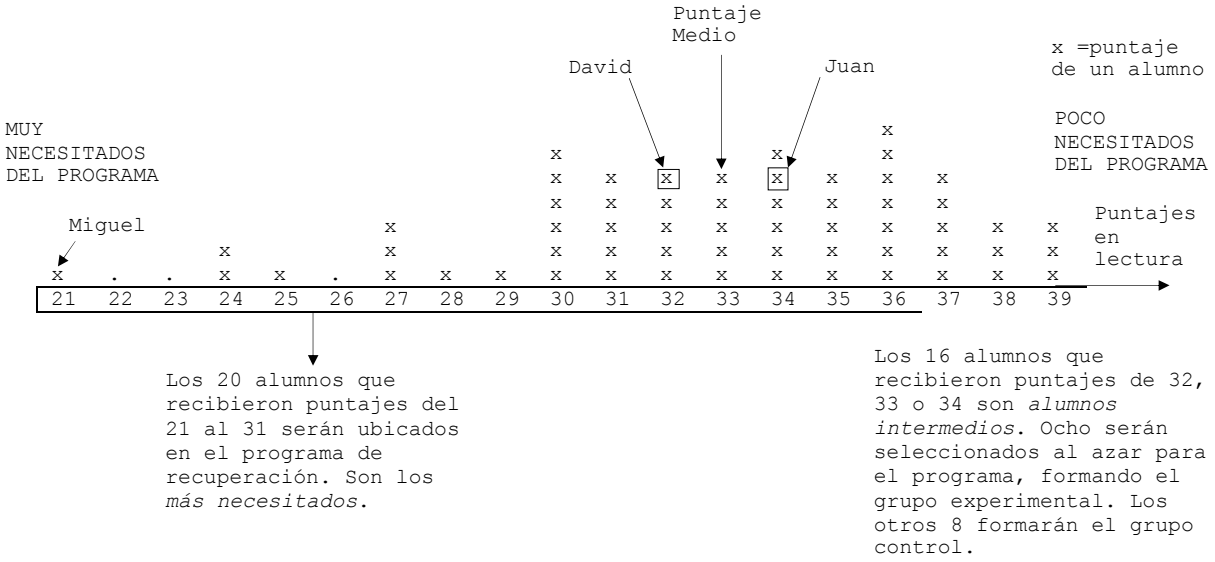


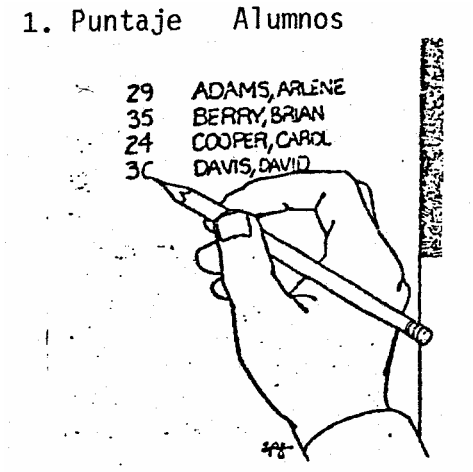
Figura 6. Como se usa un diagrama de puntajes en lectura (i) para asignar a los más necesitados a un programa de recuperación y (ii) para formar un grupo experimental y un grupo control con alumnos intermedios.

Este método de diseño es una forma excelente - e innovativa - para probar el valor de programas especiales. Es realmente preferible al diseño simple Antesy-Después. Elimina los problemas de interpretación de información que resultan de los efectos de regresión. Estos son problemas que abundan en el diseño Antes-y-Después, cuando quiera que se dé un programa a los alumnos con puntajes inicialmente extremos.

Es importante reconocer una característica del método intermedio: Produce resultados que son, hablando en términos estrictos, generalizables sólo a grupos intermedios. Por ejemplo, es evidente que si la evaluación muestra que los alumnos intermedios que recibieron el programa sobrepasaron en puntajes a los alumnos intermedios que no lo recibieron, la generalización que podemos hacer es que el programa es beneficioso *para alumnos intermedios*. El efecto del programa sobre los alumnos más necesitados no puede necesariamente inferirse de los efectos comprobados en los alumnos intermedios, quienes formaron los grupos experimental y control.

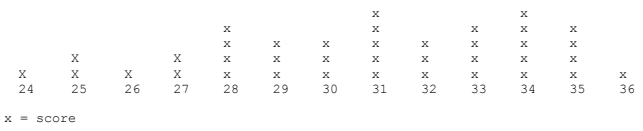
Sin embargo, esta gran limitación del método intermedio a veces puede ser su mayor virtud. Tiene, una aplicación específica en casos en que la decisión sobre un programa no es si dejarlo o continuarlo, sino más bien si expandirlo para incluir más alumnos. ¿Debería el colegio, por ejemplo, crear dos cursos de recuperación en lectura más bien que uno? Si su evaluación ha usado un grupo control intermedio y ha mostrado cuán beneficioso es el programa para los alumnos intermedios, usted está en la situación más firme posible para recomendar una expansión del programa para incluir un mayor número de dichos alumnos. Instrucciones paso a paso para usar el método intermedio aparecen a continuación.

Instrucciones paso a paso Para Formar el Grupo Control Límite.



Reúna los puntajes que los alumnos que posiblemente participarán en el programa nuevo han obtenido en algún test relevante a dicho programa. Estos puntajes podrían ser los resultados de un test objetivo o podrían ser calificaciones hechas por los profesores.

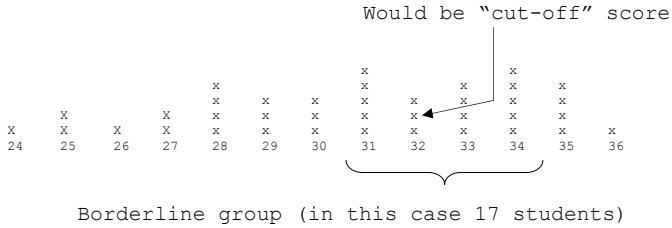
2.



Diagrame estos puntajes y domine los dos extremos "muy necesitados del programa" y "poco necesitados del programa". Un programa de recuperación tendrá los puntajes más bajos marcados "muy necesitados"; un programa avanzado tendría los puntajes más altos marcados "poco necesitados".

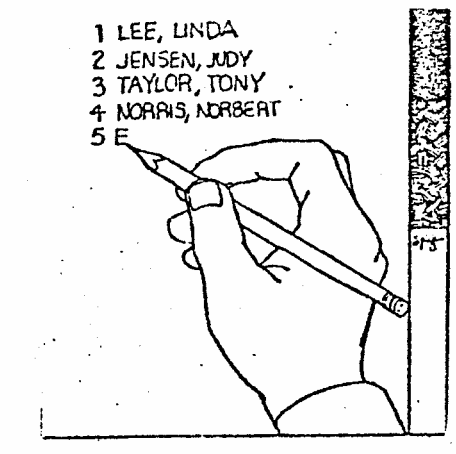
3. Decida cuántos alumnos pueden ubicarse en el programa.

4. Separe este número contando del extremo de los más necesitados" en la distribución y marque el puntaje hasta donde llega (use una flecha). Si usted estuviera simplemente usando esta distribución para asignar alumnos al programa, la flecha estaría marcando el puntaje de separación. Pero usted ahora necesitaría formar un grupo intermedio alrededor del puntaje de separación.

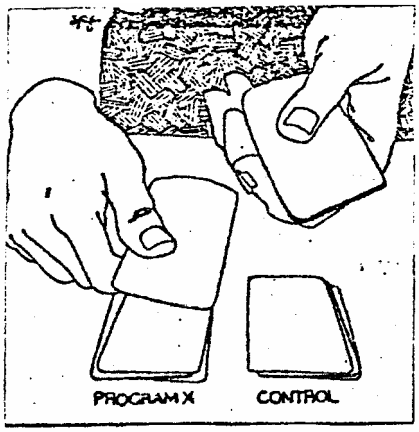


5. Excluya por lo menos a seis alumnos en y a cada lado del puntaje que separa. Los 12 o más alumnos que usted ha contado (seis a cada lado− más cualesquier alumno que tenga el mismo puntaje que tiene la sexta persona en cada lado) forman el grupo intermedio.

6. Haga una lista numerada del grupo de alumnos intermedio.



7.



Use los métodos de cara o sello, de la BPA o la lista preparada de números aleatorios, para asignar al azar a la mitad de los alumnos de la lista al programa y la otra mitad al grupo control.
Alternativamente, ponga los nombres del grupo intermedio en un sombrero o en otro objeto, revuélvalos y saque la mitad de los nombres para el programa.

Selección de una Muestra Aleatoria.

Puede que haya ocasiones en que usted necesite seleccionar una *muestra* aleatoria, de entre un gran número de alumnos. Esta situación será ligeramente diferente de aquella en que usted necesita *asignar* al azar a todos los alumnos a los programas. Por ejemplo, puede ser que usted quiera entrevistar alumnos que están en un programa especial, pero no tiene tiempo suficiente para conversar con todos. Dependiendo del tiempo de que usted dispone para entrevistar y del número de alumnos que están en el programa, usted puede decidir entrevistar a tan sólo el 10%, o a tantos como el 25% de los alumnos involucrados. Otra situación en que usted pudiera necesitar una muestra de alumnos está en relación con el Diseño de Series Cronológicas (Diseños 4 y 5), o puede ser que usted quiera encuestar una pequeña muestra al azar de alumnos en el colegio, sobre bases regulares.

Decidir sobre el tamaño de la muestra será el primer paso necesario. Aquí sirve más un procedimiento práctico. Haga la muestra tan grande como usted pueda en términos de tiempo y dinero. Una muestra grande representa mejor al grupo total; mientras más pequeña, menos puede usted esperar que sus resultados reflejen con exactitud las "verdaderas" destrezas o actitudes del grupo que representa. Por lo general, para realizar análisis estadísticos una muestra del tamaño de 30 se considera adecuada para obtener una medida estable independiente del tamaño del grupo que se representa. Esto es algo diferente de una preocupación por asegurar que la muestra represente adecuadamente el rango de opiniones o habilidades reales en la población mayor.

Bastante se ha escrito y discutido sobre el problema del tamaño de la muestra. Han surgido algunos principios generales y aunque no son unánimemente aprobados, se le entregan aquí: Si usted quiere usar los resultados obtenidos de la muestra para hacer una afirmación general sobre todas las personas que tienen una cierta característica, o que alguna vez han tomado parte en un cierto programa... en otras palabras, si la población a la cual usted desea generalizar es esencialmente infinita... entonces consulte su propia opinión y la de personas competentes. Analice el grado de credibilidad que tendrían para su audiencia muestras de diversos tamaños. Pregúntese:

> "¿Sobre qué tamaño de la muestra debería yo insistir antes de creer en los resultados de la investigación que proponía hacer afirmaciones generales sobre todas las personas con las características X?"

Pregunte a sus colegas y a su propia audiencia de evaluación. Puesto que la provisión creíble es su responsabilidad fundamental al hacer la evaluación, use los juicios de las personas para ayudarlo a decidir sobre un tamaño adecuado de la muestra.

Por lo general, en situaciones de evaluación la población de personas sobre las cuales usted quiere hacer una afirmación es una población que se describe y se cuenta (por ejemplo finita) con facilidad, tal como la población de padres de una determinada escuela de enseñanza básica superior, o los profesores de un determinado distrito escolar. Para estas situaciones, los investigadores han determinado, vía una fórmula compleja, los tamaños de muestra necesarios para una representación creíble de grupos finitos.

La Tabla 7 [*] se entrega aquí como una guía útil para decidir el tamaño necesario de la muestra. Simplemente ponga "N" al tamaño de todo el grupo (población) que va a representar la muestra, y anote la "S" correspondiente. Este es el tamaño de la muestra sugerido. Si usted no puede obtener una muestra de ese tamaño, considere sus resultados como menos creíbles. Si usted puede igualar el tamaño sugerido, puede estar más bien cierto de una representación exacta a través de un muestreo aleatorio simple.

La preocupación fundamental al usar muestras para representar grupos más grandes es, por supuesto, el problema de asegurar una representación exacta El tamaño adecuado de la muestra ayuda a aumentar la representatividad. Lo mismo sucede con la estratificación (o formación de bloques). La estratificación requiere que usted seleccione separadamente de entre grupos de personas a quienes difieren de acuerdo a algunas características importantes que podrían afectar sus resultados... tales como edad, sexo o CI. La muestra completa seleccionada de este modo, debería por lo tanto representar varios subgrupos dentro de un grupo más grande. Usted debería notar que la estratificación le dará bastante información sobre los efectos de diferentes programas en las personas que tienen las diversas características consideradas en la estratificación. Esto probablemente afectará su ambición sobre los análisis estadísticos que usted realiza. Si sus planes de medición fueron determinados por los Diseños 1, 2 o 3, y usted había usado muestreo aleatorio simple, entonces usted se ha dado el poder de obtener información más detallada sobre los efectos del programa. Usted debería usar un análisis de varianza (ANDEVA), discutido en el Capítulo 7.

Si usted estratifica, entonces, la discusión sobre el tamaño de la muestra en la página 172, es válida para cada subgrupo (es decir, cada característica representada por una celdilla en la tabla ANDEVA). Usted fácilmente puede ver que para estratificar se requieren muestras mucho más grandes. En evaluaciones, probablemente sólo será necesario cuando se requiera o cuando usted tenga un presentimiento fuerte y persistente de que una característica afectará en forma diferente los resultados.

Si usted decide estratificar, entonces use las instrucciones para seleccionar una muestra al azar simple, una y otra vez para cada subgrupo. Al realizar el paso 1, haga

[*] *Krijcie, R.V., and Morgan, D.W. Determining sample size for research activities. Educational and Psychological Measurement, 1970, 30, 607-610. Esta tabla se basó en una fórmula publicada por la división de investigación de la Asociación Nacional de Educación.

listas separadas de personas por cada característica para ser muestreada por separado, y repita los pasos 2 al 5 para sacar cada subgrupo de muestra.

TABLA 7

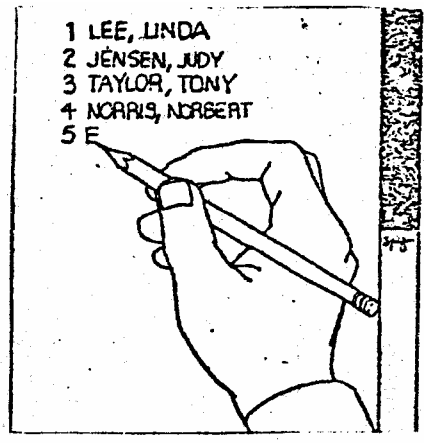
Tabla para Determinar el Tamaño de la Muestra de una Población Dada.

N	s	N	s	N	s
10	10	220	140	1200	291
15	14	230	144	1300	297
20	19	240	148	1400	302
25	24	250	152	1500	306
30	28	260	155	1600	310
35	32	270	159	1700	313
40	36	280	162	1800	317
45	40	290	165	1900	320
50	44	300	169	2000	322
55	48	320	175	2200	327
60	52	340	181	2400	331
65	56	360	186	2600	335
70	59	380	191	2800	338
75	63	400	196	3000	341
80	66	420	201	3500	346
85	70	440	205	4000	351
90	73	460	210	4500	354
95	76	480	214	5000	357
100	80	500	217	6000	361
110	86	550	228	7000	364
120	92	600	234	8000	367
130	97	650	242	9000	368
140	103	700	248	10000	370
150	106	750	254	15000	375
160	113	800	260	20000	377
170	118	850	265	30000	379
180	123	900	269	40000	380
190	127	950	274	50000	381
200	133	1111	278	75000	382
210	136	1100	285	1000000	384

NOTA: N es el tamaño de la Población
s es el tamaño de la Muestra

Uso de la BPA para Seleccionar una Muestra Aleatoria

1

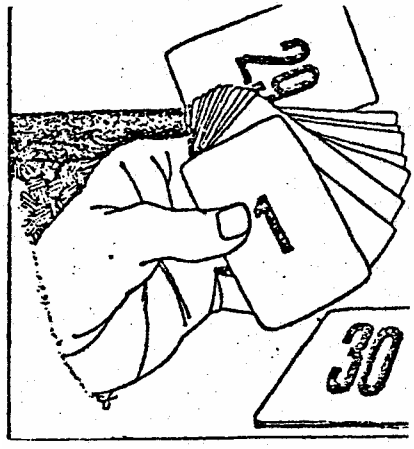


Haga una lista de todos los alumnos que hay disponibles para hacer la selección. Numere los alumnos en la lista.

2.

Use la Tabla 7, página 136 o las sugerencias en la página 134 considerando las restricciones prácticas de su situación para decidir sobre el tamaño de una muestra, s.

3.

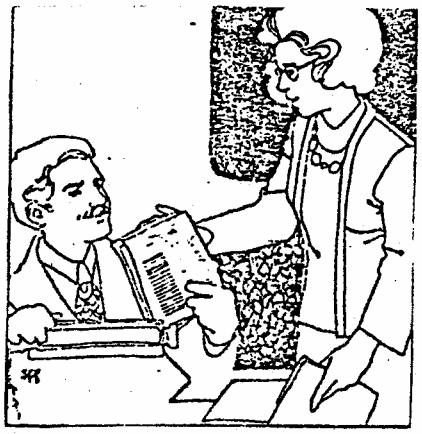


Saque de la BPA las cartas correspondientes al número total de alumnos (N). Barájelas y corte varias veces.

4.

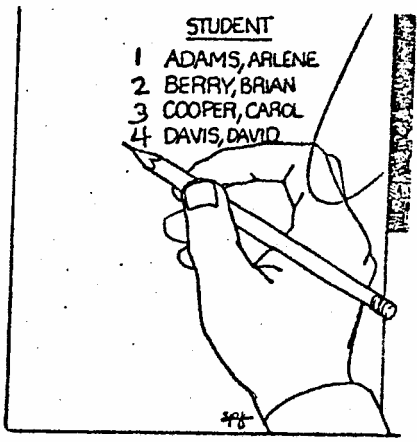
Saque al azar de esta baraja el número de cartas correspondientes al tamaño de la muestra deseada. (s)

5.



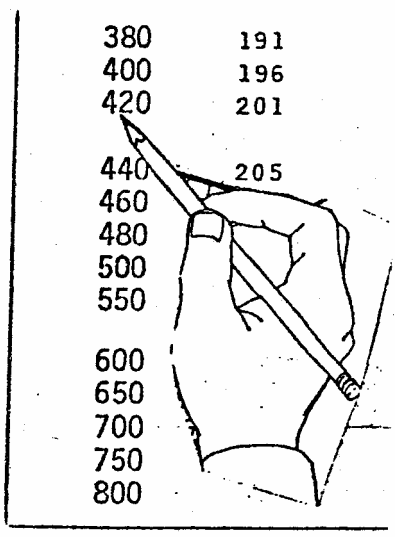
Haga una lista de los alumnos cuyos números se han elegido.

Uso de la lista de alumnos para Seleccionar una Muestra Aleatoria Simple.



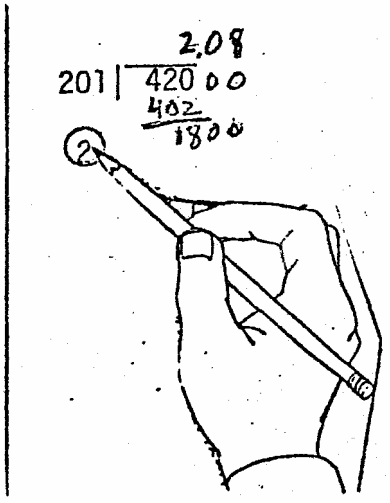
1 Haga u obtenga una lista numerada de todo el grupo disponible. Cualesquier orden que no se base en agrupación por características importantes... por ejemplo, alfabético...bastará.

2.



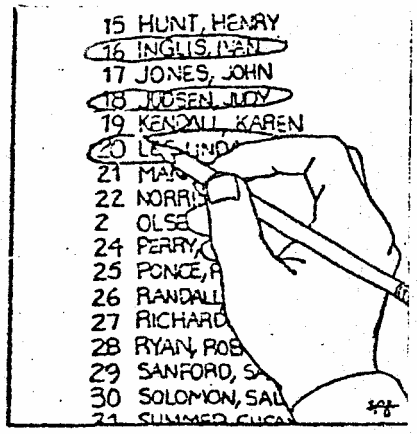
Use la tabla 7, página 136 o las sugerencias en página 134 considerando las restricciones prácticas de su situación, para decidir sobre el tamaño de la muestra, s.

3.



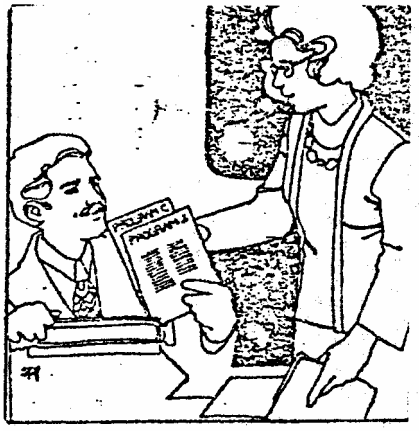
Divida “N”, el tamaño de todo el grupo disponible, por “s”. Aproxime hacia arriba o hacia abajo, es decir, cambie decimales al siguiente número entero.

4.



Use este número entero para contar hacia abajo en la lista y encierre en un círculo los nombres de las personas seleccionadas. Si el número es 2, por ejemplo, entonces encierre en un círculo y seleccione cada dos nombres.

5.



Haga una lista de los alumnos que se han elegido.

CAPÍTULO 8

EL ROL DEL DISEÑO Y UN ESTUDIO DE DOS SITUACIONES COMUNES DE EVALUACIÓN.

EL ROL DE LOS DISEÑO EXPERIMENTALES EN EVALUACIÓN.

Para los lectores que están corrientemente trabajando como evaluadores, la insistencia que se hace en este libro en el sentido que el diseño es importante para la evaluación sumativa puede parecer algo sin base. "Nadie usa diseños experimentales. Nadie usa "grupos control", podría decir usted. Esa afirmación no está lejos de la verdad, desgraciadamente. En realidad, un estudio a nivel nacional de las evaluaciones del Title VII (Educación Bilingüe) del ESEA[*] reveló que solamente un 36% intentó usar un verdadero grupo control para cualesquier aspecto de los programas evaluados[**]. En otro estudio, una investigación hecha sobre 2.000 proyectos que se habían reconocido como exitosos, no se ubicó ninguno con una evaluación que proporcionara evidencia aceptable con respecto al éxito o fracaso del proyecto[*].

Esta situación lamentable, quizás, se origina en el hecho de que el evaluador típicamente no tiene control sobre la asignación de alumnos, ni sobre la implementación del programa. Al evaluador simplemente se le llama y se le pide que "evalúe" un programa que ya está funcionando.

La falta de diseño de investigación se origina en parte por su existencia relativamente nueva como un método para reunir información en ciencias sociales.

El concepto de experimentación (el uso de diseños experimentales) en sistemas sociales complejos, es una idea más bien nueva. El trabajo de Sir Ronald Fisher en estadística fue un paso adelante, esencial, para las ciencias sociales y ello se logró principalmente en los años 30. Eso no es mucho tiempo atrás.

El desarrollo de la idea de experimentación en ciencias sociales no es solamente nuevo, sino que se ha llegado a reconocer como extremadamente importante. Por

---

[*] ESA se refiere a la Ley de Educación Básica y Secundaria aprobada por el Congreso de Estados Unidos en 1965. Estableció ayuda financiera Federal para los programas de Educación especial. Las modificaciones a la ley que especifican los tipos de programas que recibirán fondos se llaman "Titles". "Title" I, por ejemplo, da dinero a los niños limitados.

El "Title" VII financia los programas bilingües. El ESEA se ha cambiado varias veces desde 1965, pero su propósito básico y el de sus "Titles" permanecen igual.

[**] Alkin, M.C. Kosecoff, J. Fitz-Gibbon, C., and Seligman R. Evaluation and decision making: The Title VII experience. CSE Monograph Series in Evaluation, Nº 4 Los Angeles: Center for The Study of Evaluation, 1964.

[*] Foat, C.M.Selecting exemplary compensatory education projects for dissemination via project information packages. Los Altos CA:RMC Research Corporation, May 1974 (technical Report Nº UR 242).

ejemplo, Michael Scriven, un filósofo de la Universidad de California, cree que, "La introducción de los diseños experimentales altamente controlados... y otras herramientas, han hecho posible ahora extraer relaciones interesantes de tipo causal y de correlación a partir de información que anteriormente ere inútil"[**]. Y ese gran hombre del renacimiento, Herbert A. Simon, de la Universidad de Carnegie-Mellon ha establecido que"...el verdadero estudio de la humanidad es la ciencia del diseño..."[***].

La experimentación, entonces, es nueva. Si las personas están interesadas en los resultados, se usará cada vez más. Algunos profesores, directores y miembros de las juntas escolares están pidiendo preguntas de sondeo, preguntas a las cuales sólo buenos diseños experimentales puedan proporcionar respuestas. Sin embargo, hasta que las agencias Estatales y Federales no pidan el uso de buenos diseños, el actual diseño de evaluación antes-y-después continuará usándose.

Un estudio del Rol del Evaluador en Dos Situaciones Comunes de Evaluación

Esta sección considera dos de las situaciones más comunes de evaluación: Las evaluaciones del "Title ESEA" y las evaluaciones de Educación Especial.

Evaluaciones "Title": Por evaluaciones "Title" entendemos evaluaciones de programas fundados bajo el "ESEA", Decretos de Evaluación de Educación Básica y Secundaria del Gobierno de Estados Unidos, tales como:

"Title" I - limitados

III - programas innovativos

VII - educación bilingüe

VIII - prevención de deserción

A menudo, a los evaluadores de estos programas se les pide que presten informaciones de pruebas a los "supervisores" de los programas. Puede ser que ellos también tengan que informar sobre diversos componentes de su programa, e informar sobre la consistencia del programa tal como se implementó, con el esquema del programa que se entregó en la "propuesta" sobre cuyas bases se asignaron los fondos.

Cuando por primera vez se pidieron evaluaciones en la década de los 60, hubo gran preocupación de parte del personal directivo de proyectos, en el sentido de que los informes de evaluación favorables o desfavorables, afectarían directamente las decisiones en cuanto a nuevas asignaciones de fondos. Como han resultado las cosas,

[**] Scriven, M. "The Structure of the social Studies" in Ford, G.W. and Pugns, L. (Eds) The Structure of Knowledge and The Curriculum. Chicago: Rand Mc Nally, and Co, 1964.

[***] Simon, H.A.The Sciences of the artificial. Cambridge: MT press, 1969.

hay poca evidencia de que los informes de evaluación tengan alguna influencia en las decisiones de nuevas asignaciones de fondos.

Uno puede postular varias razones posibles sobre por qué los informes de evaluación no sirven para distinguir entre programas efectivos y programas no efectivos.

1. Los evaluadores son contratados por el personal responsable de proyecto y es poco probable que "muerdan las manos de quienes los alimentan". Por lo tanto, los informes de evaluación son casi siempre positivos.

2. Los evaluadores están en contacto frecuente con el personal responsable del proyecto y es poco probable que sean críticos de sus amigos y colegas. Por lo tanto, otra vez, los informes de evaluación son casi siempre positivos.

3. Los diseños experimentales que proporcionarían una evidencia objetiva de la efectividad de programas experimentales no se solicitan. Por lo tanto, los supervisores de programas generalmente no pueden juzgar el valor de los programas a partir de la información de evaluación. Los diseños antes-y-después no entregan una base adecuada para juzgar efectividad.

4. Los informes de evaluación a menudo se presentan sólo <u>después</u> que los fondos para el año siguiente ya se han asignado, de tal modo, que incluso, si hubiera una influencia debida de los informes, mostraría un considerable retraso.

La inferencia que puede extraerse es que se realiza muy poca evaluación "sumativa" verdadera. Esta situación continuará hasta que se cambien las exigencias hechas por los organismos Federales o Estatales que financian los programas.

Una respuesta adecuada a la situación que se presenta anteriormente es que los evaluadores corrientes de los programas de "Titles" deberían considerarse ellos mismos evaluadores <u>formativos</u> más bien que <u>sumativos</u>. La evaluación formativa es la evaluación que tiene como objetivo retroinformar de forma regular al personal responsable del programa que lo está dirigiendo en los colegios. Su objetivo es mejorar el programa. Trabajando en dicha condición, el evaluador puede hacer muchísimo bien. Verificando constantemente el progreso y la implementación del programa e informando sobre los resultados, el evaluador formativo centra la atención de cada persona en las metas del proyecto.

Para mejorar las posibilidades de que el proyecto alcance sus objetivos y los profesores puedan asignar tiempo en forma apropiada, para concentrarse en los objetivos que no se han alcanzado, y no quedarse innecesariamente en objetivos que ya se han dominado bastaría con hacer pruebas en forma consistente y usar los resultados para retroinformar.

Tan importante como la información que proporciona el evaluador es, quizás, su influencia como persona. El evaluador puede mostrar reconocimiento por el profesor muy trabajador, dar ánimo al profesor esforzado y estimular al profesor cansado.

Para ser honesto, parece que la mejor forma u oportunidad del evaluador del programa del "Title" para impresionar al "supervisor" Federal o Estatal, se encuentra en el hecho de presentar un informe de evaluación masivo: muchas mediciones, informes

en tablas y gráficos, bastante documentación, y registros de las opiniones de muchos grupos. Entregue un informe largo, bien organizado y claro, pero especialmente largo.

Este análisis serio no excluye la necesidad de buenos diseños experimentales en evaluación. Como evaluador formativo, usted puede obtener una formación valiosa usando un experimento para determinar la efectividad de un componente de un programa en oposición a otro. Esto ayudará en decisiones sobre qué componentes...materiales, horarios, énfasis, etc., son efectivos y cuáles son menos efectivos.

Quizás los componentes del programa estén recibiendo énfasis diferentes en los diferentes colegios. Usted puede investigar entonces los efectos de estos énfasis diferentes. Si, por ejemplo, el director del proyecto decide concentrarse en la participación cada vez mayor de los padres en dos lugares, mientras que en otros dos lugares decide concentrarse en profesores en servicio, usted puede comprobar las actitudes de los padres y profesores en los cuatro lugares y puede también comprobar si los énfasis diferentes parecen estar teniendo efectos diferentes. Si las actitudes de los profesores fueran muy parecidas en los cuatro lugares, el profesor "en servicio" no parecería tener mucho impacto en las actitudes. Si la participación de los padres, por otra parte, mostró un salto en los dos lugares donde el director hizo un esfuerzo conjunto, pero no en los otros dos lugares (estos "otros dos lugares" forman un grupo control), entonces ese esfuerzo resulta efectivo.

Tratando de medir los efectos de varios componentes en todos los lugares y de manera concurrente variando la cantidad de cada componente en los diversos lugares, usted puede empezar a reunir indicaciones poderosas de lo que está y lo que no está funcionando. Usted puede guiar el proyecto para que sea efectivo al máximo, en cuanto a distribución interna de tiempo, de esfuerzo y de dinero, un servicio muy valioso y que vale la pena.

La Evaluación de Programas de Educación Especial. Los programas de Educación Especial están diseñados para categorías especiales de niños, tales como aquellos que se ubican en el 2% superior en un test de inteligencia ("superdotados"), aquellos que se ubican bajo un puntaje de 75 en un test de inteligencia ("retardados"), niños que tienen defectos físicos o tienen alguna perturbación emocional, etc.

Las leyes que exigen que todos aquellos niños sean educados, excluyen los diseños de evaluación cuando un programa de educación especial se compara con un programa no especial. Un verdadero grupo control (es decir, asignado al azar) puede por lo tanto formarse solamente si el colegio tiene dos programas disponibles para el mismo grupo de alumnos de educación especial.

Ejemplo.

Un colegio ensayó dos tipos diferentes de programas para sus niños superdotados. Los niños superdotados se asignaron al azar a uno u otro programa durante un periodo de ensayo de 10 semanas. Los costos y beneficios para ambos programas fueron determinados por el director al final del periodo de ensayo. Las reacciones de los alumnos y padres fueron positivas para ambos programas, pero un programa que incluía viajes a terreno

creó bastante resentimiento en alumnos que no estaban en el programa. Puesto que ellos no pudieron justificar los viajes a terreno como era necesario, el director y personal directivo decidió continuar con el otro programa.

Los párrafos siguientes sugieren otras formas posibles para la evaluación de programas de educación especial.

1. *Use el diseño de Grupo control no equivalente.*
(Diseño 3, páginas 67 – 76).

Dicha comparación se podría hacer si otro distrito o colegio que no tenía programas especiales, o programas considerablemente diferentes de los suyos, estuvieran de acuerdo en administrar los mismos tests que usted usa y en compartir los resultados.

Ejemplo
Los profesores de niños mentalmente retardados, susceptibles de ser educados, planificaron un programa en destrezas de lectura que esperaban mejoraría significativamente la lectura de sus niños retardados. Ellos pidieron a una escuela elemental, cercana, que compartiera con ellos los resultados de un test de lectura dado por el distrito en mayo* de cada año, y que permitieran que se diera un test referido a criterio a los alumnos retardados a comienzos y fin del año escolar. El progreso de los dos grupos en lectura podría compararse.

2. *Evalúe los componentes del programa.*

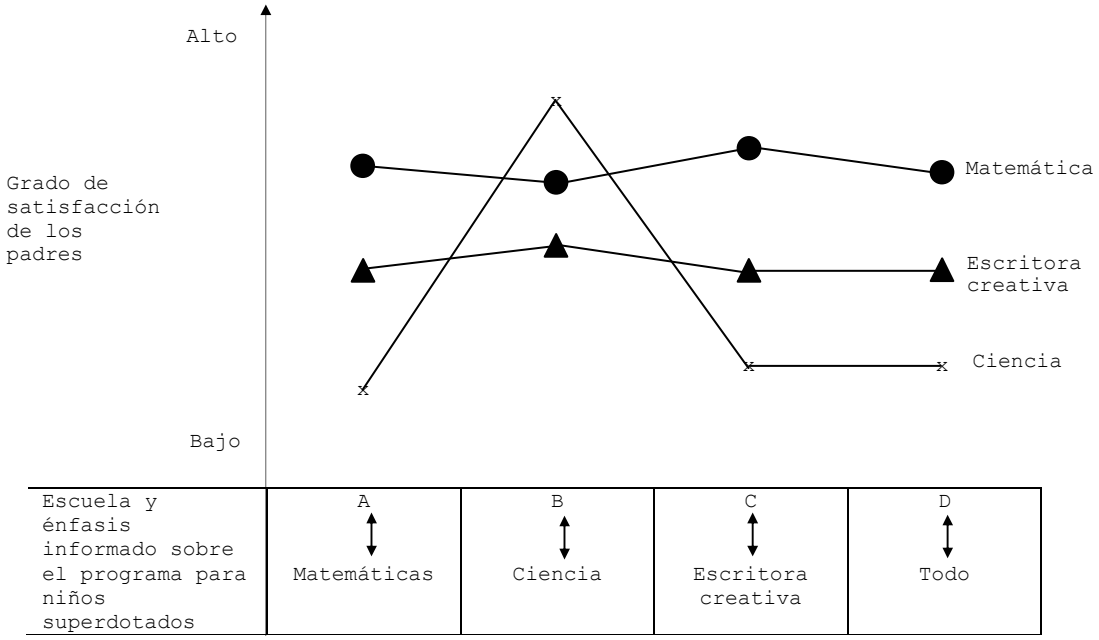
Los estudios comparativos de los efectos de los programas totales no son generalmente la preocupación de los evaluadores de los programas de educación especial en los distritos. Más bien, a ellos se les encarga que evalúen los componentes del programa de educación especial con miras a recomendar cambios que pudieran ser necesarios. En algunos casos, por ejemplo, podría haber materiales alternativos disponibles para los mismos objetivos. Experimentos en pequeña escala podrían realizarse en varios colegios, usando un diseño de grupo control verdadero pretest-postest (Diseño 1) en cada curso, para obtener una información objetiva sobre la efectividad de las diversas alternativas.

3. *Compare diversos programas en términos de satisfacción.*

A veces a un evaluador se le pide que evalúe un número de programas de educación especial que los colegios en forma individual han producido y los

* En Mayo, está finalizando el año escolar en USA.

cuales tienen diferentes metas y objetivos. Quizás en un colegio, el programa para superdotados se concentra en aceleración en matemáticas, en otro colegio en el tiempo que se dedica a ciencias, en otro en escritura creativa, y en un cuarto colegio en todas estas cosas al mismo tiempo. Usted podría medir la satisfacción que sienten los alumnos y padres en todos los colegios con respecto a la instrucción que se ofrece en asignaturas individuales (matemáticas, ciencias, destrezas en escritura). Quizás usted encontraría resultados como estos:



Los padres parecen en general satisfechos, más o menos en la misma proporción, independientemente del énfasis con que informa el colegio, en matemáticas y escritura creativa. Ciencias, sin embargo, parece muy sensible al hecho de si está o no está enfatizada en el programa de superdotados. Cuando está, hay gran satisfacción. El evaluador podría observar que en ausencia de un esfuerzo especial, las ciencias quizás no se enseñen bien a los alumnos superdotados, al menos si la satisfacción de los padres es un indicador válido.

Un punto importante en este ejemplo hipotético es que diversos programas pueden medirse a veces con la misma medida, satisfacción. Es de esperar que cuando se hace un esfuerzo especial en un área, ésta debería aparecer en las unidades correspondientes. Por medio de este tipo de investigación, usted puede ver qué tipo de programas parecen marcar una diferencia. Por supuesto, a menudo es una persona más bien que un programa[x] quien hace la diferencia. Pero esto no se debe dar por hecho tan fácilmente. Si una persona que enseña una asignatura parece tener mucho éxito, haga que otros traten de enseñar la asignatura de la misma forma. Quizás ellos también puedan llegar a ser muy efectivos.

Como evaluador interno (empleado por el distrito), su rol como evaluador a menudo se agrega simplemente a sus roles como facilitador, fuente de estímulo, persona[x] que ayuda al personal responsable a desarrollarse, y persona-recurso. Conducir una evaluación midiendo la satisfacción de varios grupos: padres, alumnos, profesores, comunidad, etc., es una forma simple de lo que un evaluador distinguido, Robert Stake, ha llamado “Evaluación responsable”.

4. *Fije criterios locales.*

Con frecuencia, a los programas de educación especial se les pide que formulen objetivos medibles, y el trabajo del evaluador es medir el logro de los objetivos. Esto a menudo parece producir un juego consistente en establecer objetivos que sean lo bastante altos para ser aceptables, pero bastante bajos para ser logrados, especialmente cuando los objetivos se establecen en términos de logros de tests estandarizados.

A veces, sin embargo, cuando los objetivos se derivan de criterios que tienen valor intrínseco, fácil de reconocer, un método excelente es fijar metas razonables. Por ejemplo, la especificación de alguna destreza básica de supervivencia para alumnos retardados, podría proporcionar objetivos de dominio para un programa de niños mentalmente retardados susceptibles de ser educados (tal como leer señales de tránsito en forma correcta, etc.)

5. *Haga la evaluación basada en teoría.*

Un buen método para la evaluación de programas de educación especial es hacer una "evaluación basada en teoría". ¿En qué teoría de instrucción, teoría de aprendizaje, teoría psicológica, o teoría filosófica se basa el programa? En otras palabras, ¿qué otras actividades considera el personal responsable como importantes para obtener buenos resultados hacia lo que intenta el programa? Una serie de preguntas detalladas hechas al personal responsable pueden ayudarlo a ubicar y hacer explícito el modelo, teoría, o filosofía que el personal responsable está tratando de implementar.

Si usted elige un método basado en teoría, su trabajo será determinar si las actividades que están especificadas por la teoría, están siendo operacionalizadas e implementadas efectivamente. ¿Puede documentarse la existencia de actividades planificadas, es decir, respaldarse por medio de evidencia reunida objetivamente, y no sólo por testimonios? Si usted puede mostrar en su evaluación que los elementos que especifica la teoría como necesarios para el logro de los objetivos están presentes, entonces usted ha mostrado que el programa ha dado un paso efectivo hacia el logro de los objetivos. Si la teoría es correcta, los objetivos deberían alcanzarse eventualmente.

En conclusión, las evaluaciones sumativas de los programas que se van a juzgar por medio de sus resultados, deberían emplear los diseños experimentales. Sin embargo, si usted no pudiera implementar un buen diseño, hay muchas otras formas en que usted puede proporcionar información para ayudar a hacer la instrucción efectiva.

Usted es el mejor juez para determinar cuáles son los procedimientos más apropiados para su situación. No vacile en seguir su propio juicio.