

Metodología de investigación

en psicología general sanitaria

PROYECTO EDITORIAL:
Metodología de las Ciencias del Comportamiento y de la Salud

Directores:
Antonio Pardo Merino
Miguel Ángel Ruiz Díaz



Queda prohibida, salvo excepción prevista en la ley, cualquier forma de reproducción, distribución, comunicación pública y transformación de esta obra sin contar con autorización de los titulares de la propiedad intelectual. La infracción de los

derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (arts. 270 y sigs. Código Penal). El Centro Español de Derechos Reprográficos (www.cedro.org) vela por el respeto de los citados derechos.

Metodología de investigación

en psicología general sanitaria

Juan Botella Ausina • José Manuel Caperos Montalbán



Consulte nuestra página web: **www.sintesis.com**
En ella encontrará el catálogo completo y comentado

Reservados todos los derechos. Está prohibido, bajo las sanciones penales y el resarcimiento civil previstos en las leyes, reproducir, registrar o transmitir esta publicación, íntegra o parcialmente, por cualquier sistema de recuperación y por cualquier medio, sea mecánico, electrónico, magnético, electroóptico, por fotocopia o por cualquier otro, sin la autorización previa por escrito de Editorial Síntesis, S. A.

© Juan Botella Ausina
José Manuel Caperos Montalbán

© EDITORIAL SÍNTESIS, S. A.
Vallehermoso, 34. 28015 Madrid
Teléfono: 91 593 20 98
www.sintesis.com

ISBN: 978-84-9171-399-9
Depósito Legal: M. 24.431-2019

Impreso en España - Printed in Spain

Índice

<i>Presentación</i>	9
1. Introducción a la metodología de investigación	13
1.1. La psicología basada en la evidencia: nota histórica	13
1.2. La práctica clínica como objetivo científico	18
2. Estudios de resultados con grupos	21
2.1. Estudios de eficacia	21
2.1.1. El concepto de validez	22
2.1.2. Diseños experimentales según el grupo de comparación.....	26
2.1.3. Control de variables.....	30
2.1.4. El ensayo controlado aleatorizado	35
2.2. Criterios para una terapia empíricamente válida (TEV)	35
2.3. Estudios de efectividad	37
2.3.1. Eficacia versus efectividad.....	37
2.4. Estudios cuasiexperimentales	41
2.5. Evaluación de la evidencia	41
2.6. Técnicas estadísticas en estudios de resultados con grupos	42
2.6.1. Variables de resultado cuantitativas.....	43
2.6.2. Variables de resultado dicotómicas.....	45
2.7. Valoración de los estudios de resultados con grupos	45
3. Moderación y mediación	47
3.1. Estudio de variables moderadoras	48
3.1.1. Técnicas estadísticas para analizar la moderación	49

3.2. Estudio de variables mediadoras	54
3.2.1. Técnicas estadísticas para analizar la mediación	56
4. Estudios de caso único	59
4.1. Procedimientos de control	60
4.2. Análisis visual de los datos	63
4.3. Análisis estadístico descriptivo	65
4.4. Análisis estadístico inferencial	71
5. Estudios no manipulativos	81
5.1. Manipulación e inferencia causal	82
5.2. Diseños <i>ex post facto</i> prospectivos (de cohortes)	84
5.3. Diseños <i>ex post facto</i> retrospectivos (de casos y controles)	86
5.4. Diseños transversales	88
5.5. Técnicas estadísticas en estudios no manipulativos	89
6. Análisis de datos	91
6.1. Validez de la conclusión estadística	91
6.1.1. Las pruebas de significación y los intervalos de confianza	92
6.1.2. Amenazas a la validez de la conclusión estadística	94
6.1.3. Establecimiento del tamaño de la muestra	95
6.2. Significación estadística y significación práctica	100
6.3. Estudios de equivalencia y de no inferioridad	103
6.4. Análisis de la reducción de muestras y sus efectos	105
6.5. El cambio individual: cambio fiable y cambio clínicamente significativo	108
6.5.1. Fórmulas del cambio fiable y del cambio clínicamente significativo	114
6.6. Índices e indicadores	116
6.6.1. Medidas de frecuencia	116
6.6.2. Medidas de asociación basadas en frecuencias	117
6.6.3. Instrumentos de cribaje y curvas ROC	123
7. Meta-análisis	131
7.1. El marco conceptual e histórico del meta-análisis	131
7.1.1. Antecedentes: revisiones narrativas y meta-análisis	132
7.1.2. Características y objetivos	132

Índice

7.1.3. El proceso de un meta-análisis.....	133
7.2. Transformación a una métrica común	134
7.2.1. Tamaño del efecto y significación estadística.....	134
7.2.2. Índices de tamaño del efecto	136
7.3. Análisis estadístico	141
7.3.1. Estimación combinada.....	141
7.3.2. El <i>forest plot</i>	144
7.3.3. Análisis de la heterogeneidad.....	145
7.4. Evaluación del meta-análisis	146
7.4.1. Validez: el sesgo de publicación.....	147
7.4.2. Críticas y valoración del meta-análisis.....	147
7.4.3. El informe meta-analítico	148
7.5. Recursos: bibliografía y <i>software</i>	149
7.6. Un ejemplo numérico	150
8. Búsqueda de literatura científica	155
8.1. Bases de datos y descriptores	156
8.2. Términos y ecuaciones de búsqueda	158
8.3. Criterios de elegibilidad	160
8.4. Implementación de la búsqueda	163
<i>Referencias bibliográficas</i>	167

2

Estudios de resultados con grupos

Una de las parcelas de mayor importancia en la investigación en psicología clínica la constituyen los estudios sobre el impacto de las intervenciones (o tratamientos) que realizan los psicólogos en su ejercicio profesional. Dichas intervenciones van dirigidas sobre todo a resolver o aliviar los problemas de salud psicológica, pero también pueden ir orientadas a otras metas (como la prevención o la gestión). El objetivo de estas investigaciones, que llamaremos genéricamente *estudios de resultados (outcome research)*, es establecer una relación de causa-efecto entre la intervención y los cambios observados. Si dichos cambios se producen se intentará también establecer su magnitud e incluso, eventualmente, su relevancia clínica. Además, dependiendo de las características contextuales distinguiremos entre los estudios de eficacia (en los que se pone el acento en el rigor y el control) y los estudios de efectividad (en los que el acento se desplaza algo más hacia la representatividad y la validez ecológica). Mencionaremos también los elementos esenciales que definen los ensayos controlados aleatorizados y cómo la pérdida de uno de ellos convierte el estudio en uno cuasiexperimental. Después trataremos la cuestión de la evidencia en sí misma y las estrategias para reunirla, así como las técnicas de análisis de datos más adecuadas para responder a la pregunta sobre el impacto.

2.1. Estudios de eficacia

Los estudios de resultados tienen por objetivo establecer rigurosamente la relación de causa-efecto entre las intervenciones realizadas y los cambios observados. En este capítulo abordaremos los estudios de eficacia con grupos, que son aquellos en los que se intenta implementar el máximo control posible de las variables extrañas. Para ello expondremos el marco conceptual de la validez y luego señalaremos las formas

que adoptan los grupos de control o comparación, discutiendo sus ventajas e inconvenientes, así como las principales variables que se deben controlar. También resumiremos las características del ensayo controlado aleatorizado.

2.1.1. El concepto de validez

Hay un amplio acuerdo en que para establecer una relación causal, la metodología apropiada consiste en hacer lo que se suele denominar un experimento manipulativo. En las ciencias de la salud se les suele llamar *ensayos controlados aleatorizados* (ECA), aunque formalmente hablando constituyen lo que en la literatura general de la psicología se denominan sencillamente estudios experimentales (León y Montero, 2015).

Se entiende que A es causa de B cuando A y B covarían, A antecede a B, y no existen explicaciones alternativas plausibles. El desarrollo y debate teórico en torno a los ECA conduce a reflexionar sobre la identificación de potenciales obstáculos que dificultan cumplir esas condiciones y sobre cómo evitar o eliminar esos obstáculos. La investigación que tiene por objeto establecer relaciones de causa-efecto cuenta con dos elementos clave: el control y la comparación. Se puede decir que estas investigaciones consisten esencialmente en hacer *comparaciones en condiciones controladas*.

Sin embargo, junto al objetivo primero de establecer una relación de causa-efecto entre la intervención y el cambio, hay que añadir otros dos objetivos generales más, a los que tradicionalmente se ha prestado menos atención (Kazdin, 1986; Kazdin, 2001a). El primero de ellos es determinar las condiciones en las que esos cambios son óptimos, estableciendo las variables moderadoras asociadas. De esta forma se pueden hacer predicciones más precisas, así como los perfiles de los pacientes para los que es óptima una intervención. El segundo es la identificación de los mecanismos mediadores que están produciendo el cambio. Aunque la intervención se haya desarrollado, como así es con frecuencia, dentro de un marco teórico que predice un cambio con una determinada base conceptual, muchas veces esto no se pone a prueba. De esta forma, los estudios de resultados han adolecido de un cierto “cajanegrismo”. La metodología de los ECA permitiría estudiar la certeza de que son esos mecanismos los que producen el cambio, aunque es mucho menos frecuente que se estudien directamente esos mecanismos. Dedicaremos el próximo capítulo a estos dos tipos de efectos, la moderación y la mediación.

Para poder establecer las características de los ECA es necesario recuperar el concepto de validez. Lo más común es analizar la calidad de un diseño de investigación en términos de su validez (León y Montero, 2015; Shadish, Cook y Campbell, 2001; Shapiro, 1989). Se entiende por validez de una investigación el grado de confianza que permite mantener respecto a sus conclusiones; es decir, al

grado de solidez o credibilidad de estas. Habitualmente, el constructo general de validez se desglosa en varios tipos de validez, en función de los diferentes aspectos de la investigación a los que se refiera. Aunque se han definido muchos, los cuatro más relevantes para nuestro contexto son la validez interna, la validez externa, la validez de constructo y la validez de la conclusión estadística.

Por *validez interna* se entiende el grado de confianza con que se puede atribuir a la(s) variable(s) independiente(s) los cambios observados en la(s) variable(s) dependiente(s). En este caso, la variable independiente es la propia terapia y las variables dependientes serán las indicadores del cambio terapéutico o variables de resultado (*outcomes*). Conseguir una alta validez interna en una investigación es uno de los objetivos fundamentales del diseño experimental, de forma que en caso de obtener diferencias en la variable dependiente podamos estar razonablemente seguros de que esas diferencias se deben a la aplicación de los diferentes tratamientos y no a otras variables extrañas (Campbell y Stanley, 1963; Kazdin, 1981). Aparte de la obvia cuestión de la precedencia temporal, las principales amenazas a la validez interna en los estudios que nos ocupan son las siguientes:

- a) *Historia*. Son eventos imprevistos que coinciden en el tiempo con la intervención y que pueden producir cambios en las variables de resultado que se confunden con los de la intervención.
- b) *No equivalencia de muestras*. Si dos grupos son diferentes ya antes de comenzar la intervención (no son equivalentes), cualquier diferencia posterior no puede ser interpretada con confianza.
- c) *Regresión a la media*. La vuelta a los valores medios o típicos (línea base) después de haber evaluado en un momento con valores extremos.
- d) *Maduración*. Cambios naturales espontáneos debidos al mero paso del tiempo, especialmente relevantes en las edades en las que el desarrollo es más rápido (por ejemplo, en la infancia).
- e) *Reducción de la muestra*. Pérdida selectiva de participantes durante la investigación. Los que terminan pueden ser una muestra no representativa de los que empezaron.
- f) *Evaluación*. Hacer una evaluación previa a la intervención, que es propia de la investigación pero no tiene un papel tan importante en los contextos clínicos, puede sensibilizar a los participantes de alguna forma que influya en las respuestas post (el recuerdo de las respuestas emitidas y los intentos por mostrarse mejor o peor, la actitud de colaboración con el investigador/terapeuta, también es una amenaza a la validez de constructo).

Sin embargo, la utilidad de una investigación no depende solo de su validez interna, sino también de su validez externa o grado en que sus resultados son generalizables, y en el caso que nos ocupa esto es de vital importancia. Se entiende por

validez externa el grado de confianza en que los resultados obtenidos en un estudio son generalizables a personas, situaciones o contextos de intervención, diferentes de aquellos que se han investigado. En general, el objetivo de los ECA será establecer con confianza afirmaciones del tipo “con esta técnica se incrementan significativamente las habilidades sociales en adultos”. Es posible que en algunos ECA el objetivo se centre en alguna subpoblación muy específica (por ejemplo, estudiantes universitarios), pero en general no será así.

Las conclusiones de un estudio de resultados merecerán ser tenidas en consideración por los psicólogos clínicos solo si en el estudio realmente se han tenido en cuenta las verdaderas claves de la terapia, y aquí aparece una de las dificultades esenciales. En los estudios de resultados se intentan simular las condiciones de trabajo del psicólogo clínico, aunque manteniendo a la vez un cierto grado de control sobre el experimento; el problema reside precisamente en las dificultades que supone controlar el trabajo del psicólogo clínico mediante un puñado de especificaciones de procedimiento. No obstante, y sea cual sea la dificultad en conseguirlo, es claro que entre los objetivos de un buen estudio de resultados está el de no apartarse excesivamente de la situación terapéutica, o al menos de sus claves esenciales, mientras se mantiene un elemental control sobre las variables. Las principales amenazas a la validez externa son las siguientes:

- a) *Baja representatividad de los pacientes o las intervenciones.* Los participantes en los estudios de resultados deben ser similares a los de las poblaciones a las que se pretenden generalizar las conclusiones. Diferencias muy marcadas pueden dificultar esa generalización. Lo mismo ocurre con las propias intervenciones. Para poder ejercer un buen control se suelen implementar condiciones que deforman la intervención en sí misma. Para que las conclusiones sean generalizables lo que se hace durante la intervención debe ser razonablemente parecido a lo que luego se hará en su aplicación en entornos naturales.
- b) *Características del entorno.* El contexto de las investigaciones (facultades de psicología, clínicas universitarias) suele ser diferente de aquel al que se pretende generalizar (todo tipo de centros o consultas particulares, por ejemplo).
- c) *Evaluación.* Ya hemos dicho que la evaluación y su repetición constituyen una amenaza a la validez interna. También lo es a la validez externa, en tanto que el papel de las evaluaciones es mucho mayor en las investigaciones que en los contextos profesionales en los que se aplican las técnicas. La elección de medidas fiables y estandarizadas, con un criterio propio del contexto experimental, puede llevar a reducir los resultados de la intervención a aspectos poco representativos o irrelevantes en la condición natural de la patología.

Cuadro 2.1. Tipos de validez

<i>Tipo de validez</i>	<i>Se relaciona con</i>
Interna	La atribución de las relaciones causales
Externa	La generalizabilidad de los resultados
Constructo	La operativización de las variables
Conclusión estadística	El establecimiento de covariaciones

La *validez de constructo* se refiere al grado de confianza en que la implementación de la intervención y la medida de los resultados representan adecuadamente los constructos de nivel superior implicados. Tiene que ver con la operativización de las variables. Las amenazas a este tipo de validez se relacionan con cualquier contaminación, en el tratamiento y en la evaluación de los resultados, como en los criterios de inclusión y exclusión:

- a) *Contaminación de las intervenciones.* Para poder concluir con confianza respecto al impacto de una técnica debe haber una razonable garantía de que al aplicarla se ha hecho de forma correcta y completa, como se ha definido. Es imposible que al hacerlo no se hayan administrado también otros elementos que van ineludiblemente unidos a la terapia, como los llamados factores inespecíficos, por lo que habrá que ejercer algún control para aislar sus efectos.
- b) *Características de las evaluaciones.* Las medidas e instrumentos empleados deben ser adecuados en el sentido de que midan lo que pretenden medir y no otros aspectos cuya variación podría confundir respecto al efecto de la intervención. La validez, especificidad y relevancia de los test empleados serán la clave para evitar esta amenaza. También intervienen aquí los posibles sesgos del evaluador o la reactividad de los participantes a la propia evaluación (que ya se ha mencionado como amenaza a la validez interna y la externa).
- c) *Deficiente operativización de los criterios de inclusión y exclusión.* Al especificar en un informe de investigación los criterios que deben cumplir los pacientes para ser incluidos en el estudio o las que llevan a su exclusión se pueden cometer errores de especificación. Definir operativamente dichos criterios con ejemplos ayudará a comprender mejor la forma como se ha implementado. Una deficiente implementación hará que se incluyan pacientes que se pretendían excluir y viceversa, lo que debilita la validez de constructo.

Por ejemplo, los manuales de tratamiento son operacionalizaciones de la intervención que pueden incluir elementos contaminantes y dejar fuera otros que son relevantes. La etiqueta “terapia cognitivo-conductual” es demasiado vaga, incluyendo

elementos muy variados (Bas, 1992). Igualmente, en el grupo de control se puede incluir una versión “aguada” de la terapia habitual. La reactividad de los participantes también puede contaminar, por ejemplo, con desmoralización del grupo de control (lo que supone un comportamiento peor que un simple “no tratado”). Lo mismo ocurre con el efecto placebo del grupo tratado, si este componente solo está presente en ese grupo. Las expectativas del investigador también pueden contaminar las medidas operativas de los constructos (el ciego y doble ciego son los mecanismos correctores).

Por último, la *validez de la conclusión estadística* tiene que ver con el grado de confianza con que se puede inferir, a partir de las observaciones recogidas, que las variables estudiadas covarían (habitualmente expresado en este campo como diferencias estadísticamente significativas entre las medias de los grupos). En el capítulo 6 discutiremos las principales amenazas a este tipo de validez.

2.1.2. Diseños experimentales según el grupo de comparación

El primer impulso en la selección de un diseño para evaluar la eficacia de una intervención podría ser seguir un esquema simple, de medición antes y después de la intervención, el llamado por Campbell y Stanley (1963) *diseño simple pre-post*. Sin embargo, son bien conocidas las razones por las que este diseño es inapropiado: las posibles diferencias entre ambas mediciones podrían explicarse de muchas maneras alternativas a los simples efectos específicos de la terapia, como son la regresión a la media (Cummings *et al.*, 2004), la remisión espontánea (Lambert, 1976) o cualquier otra circunstancia que se haya producido en el intervalo entre ambas mediciones (Kazdin, 1981). Este diseño se basa en una comparación comprometida por múltiples amenazas a la validez interna.

La principal alternativa al diseño simple pre-post consiste en la inclusión de algún *grupo de control*, también con medidas antes y después. Esto conlleva un importante número de ventajas (Nock, Janis y Wedig, 2008):

- a) Asegurar la equivalencia entre los grupos antes de la intervención.
- b) Emparejar participantes en variables clave y asignar a los participantes de cada pareja a un grupo distinto de forma aleatoria.
- c) Extraer conclusiones acerca de la cantidad de cambio a lo largo del curso del tratamiento.
- d) Estudiar potenciales moderadoras de los efectos del tratamiento.
- e) Examinar factores que predicen la reducción de las muestras (*attrition*) en el tratamiento.

Una vez convencidos de las ventajas de incluir un grupo de control, la pregunta natural será la de cuál es el grupo de control más apropiado para hacer comparaciones

idóneas en un estudio de resultados. La lógica es impecable: se trata de formar dos grupos entre los cuales la única diferencia sea la aplicación de la técnica terapéutica cuya eficacia se pretende evaluar. De esta forma, cualquier diferencia posterior entre los grupos en las variables o indicadores de interés se podrá atribuir con confianza a la aplicación de la técnica terapéutica en el grupo experimental. Por supuesto, las amenazas asociadas a los sesgos de evaluación se pueden controlar incluyendo evaluaciones ciegas. Es decir, haciendo que las evaluaciones (sobre todo las postratamiento) las realicen personas que desconocen a qué grupo pertenece el individuo que están evaluando. A veces se han utilizado como control a los pacientes que esperan turno para recibir el tratamiento en lista de espera, asumiendo que al llegar al estudio en orden (pseudo)aleatorio (no asociado sistemáticamente a ninguna variable relevante), se puede considerar que las mediciones previas a la terapia son comparables. Sin duda, si se puede es mejor hacer una asignación auténticamente aleatoria a los grupos. Estamos refiriéndonos al *diseño con grupo de control no tratado*, con la variante del *grupo de control lista de espera*. Mientras que en el primer caso no se hace nada con los incluidos en ese grupo, en el segundo se les dice que deben esperar a que haya recursos disponibles, por lo que esperan ser tratados en un futuro próximo. Debemos preguntarnos si todas las diferencias posteriores que puedan observarse entre estos grupos y el grupo tratado se deben a la técnica terapéutica en estudio. Lamentablemente, la respuesta ha de ser negativa. Por un lado, ya hay una diferencia fundamental entre las dos variantes de este diseño. Es sabido que la expectativa de un cambio próximo, así como las propias actividades de evaluación asociadas a un grupo de lista de espera, hacen que en este se produzca cierto cambio.

Pero la contaminación más importante en este diseño es el efecto placebo, que se produce como consecuencia de la creencia de que se está recibiendo un tratamiento eficaz. El efecto placebo se produce en el grupo tratado, pero no en el de control. Al evaluar los efectos de la terapia, estos no se pueden separar del efecto placebo con este diseño. La lógica experimental nos lleva a intentar controlar este factor, tratando de igualarlo en ambos grupos. Para ello habría que inducir el efecto placebo en el grupo de control. Pero ¿qué es un grupo placebo en la terapia psicológica? Esta pregunta tiene por sí misma gran interés. En cualquier caso, recibe el nombre de *diseño con grupo de control placebo*.

Tanto para este diseño como para el anterior se plantean algunos problemas éticos. En concreto, muchas personas no considerarían aceptable que se deje sin tratar a un grupo de personas por la simple razón de que hace falta un grupo de control en la investigación. Los códigos deontológicos lo prohíben. Se han producido grandes escándalos mediáticos y judiciales cuando se ha revelado que algún estudio ha violado estos principios. Pero hay otros factores importantes por su potencial contaminación en el diseño con grupo de control no tratado. Los factores inespecíficos de la terapia (aquellos derivados del hecho mismo de estar en terapia, como son el incremento en la atención recibida y en la atención selectiva del cliente hacia su problema, pero que no

son debidos a la actuación específica de la técnica) han sido ampliamente reconocidos desde hace mucho tiempo. Para aislarlos de la técnica es imprescindible separarlos de los inespecíficos. De lo contrario habría una fuerte amenaza a la validez de constructo.

Los problemas metodológicos y éticos de los diseños anteriores quedan relativamente superados si se utiliza un *diseño de grupos con tratamientos alternativos*, que es el diseño recomendado por muchos autores (por ejemplo, Kazdin, 2001b; Kendall, Comer y Chow, 2013; Luecken y Tanaka, 2013; Nock *et al.*, 2008; Shapiro, 1989). La diferencia entre este diseño y los anteriores es que no se miden los efectos de la técnica en términos absolutos, sino en términos relativos con respecto a otra(s) técnica(s). Los problemas éticos se anulan, ya que el grupo de comparación recibe también un tratamiento de eficacia contrastada. En muchas ocasiones recibe lo que se suele llamar tratamiento habitual o estándar (*treatment as usual*, TAU). En ocasiones este diseño no se emplea con el objetivo de demostrar que un tratamiento es mejor que otro, sino que ofrece una eficacia equivalente. Los estudios de equivalencia tienen por objetivo avalar si el cambio a un nuevo tratamiento que tiene alguna ventaja importante se puede realizar sin que ello suponga una pérdida de eficacia: por ejemplo, un tratamiento más corto, que exige el empleo de menos recursos o que es menos desagradable. En estas circunstancias se justifica cambiar a un nuevo tratamiento si esas ventajas no conllevan una pérdida de eficacia terapéutica. Cuando el objetivo es demostrar que dos tratamientos tienen una eficacia similar, el análisis estadístico debe ser distinto al que se emplea cuando se pretende demostrar una diferencia (capítulo 6).

En el cuadro 2.2 (adaptado de Shapiro, 1989) se resumen los diseños mencionados, identificándolos según el grupo de control utilizado.

Cuadro 2.2. Tipos de grupos de control

Diseño	Características	Controla
Simple <i>pre-post</i>	<ul style="list-style-type: none"> • Un grupo • Se comparan las puntuaciones pre con las post 	
Control no tratado	<ul style="list-style-type: none"> • Los pacientes se asignan al grupo tratado o al de control no tratado 	<ul style="list-style-type: none"> • Remisión espontánea • Maduración • Historia • El “paso del tiempo”
Control placebo	<ul style="list-style-type: none"> • Los pacientes se asignan al grupo tratado o al de control placebo 	<ul style="list-style-type: none"> • Ídem al anterior • Expectativas de tener beneficios
Grupos con tratamientos alternativos	<ul style="list-style-type: none"> • Los pacientes se asignan a uno de los varios tratamientos activos involucrados 	<ul style="list-style-type: none"> • Ídem al anterior • Más convincente y ético

Según lo expuesto hasta aquí, los mejores diseños para un estudio de resultados son los de grupos con tratamientos alternativos que incluyen elementos para protegerse de las principales amenazas a los tipos de validez que hemos señalado. El esquema es el que aparece en la figura 2.1. Primero se recoge evidencia respecto a las variables de interés de forma simultánea en los dos grupos. En segundo lugar se administra la intervención que queremos valorar en el grupo experimental, mientras que en el de control o comparación se hace lo que corresponda (intervenir con el tratamiento habitual, realizar las actuaciones placebo o, simplemente, no intervenir). Por último, se vuelve a recoger evidencia respecto a las variables de interés en ambos grupos. El resultado esperado es el de la figura 2.1: sin cambios o mejoría en el grupo de comparación, pero mejoría significativamente mayor en el grupo experimental. Se contempla que pueda haber también mejoría en el grupo de control porque ha recibido otro tratamiento o porque ha recibido un placebo. Incluso en el diseño con un grupo de control no tratado se puede producir también mejoría por otras razones (como regresión a la media o recuperación espontánea) o puede no haber mejoría.

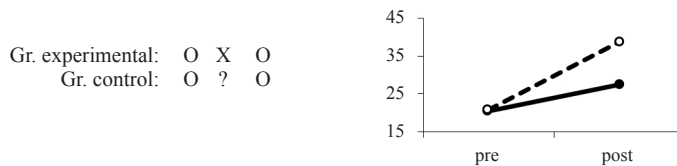


Figura 2.1. Resultado típico en estudios de resultados con dos grupos y dos momentos. El grupo de control podría ser cualquiera de los expuestos en el texto.

Sin embargo, dentro de los límites de estos diseños hay variantes que son más adecuadas para los objetivos específicos de cada estudio. El diseño básico 2×2 (2 grupos \times 2 momentos) admite múltiples extensiones, tanto en grupos como en momentos, que permiten acomodar dichas extensiones. A veces ocurre que una estrategia terapéutica funciona sin que se tenga seguridad de cuál o cuáles son los elementos críticos de esta y qué elementos son redundantes y podrían obviarse. Kazdin (1982) describe dos procedimientos para estudiar esta cuestión, ambos muy costosos en términos de participantes. El primero, que podemos denominar método de *desmantelamiento*, consiste en ir eliminando partes del tratamiento en diferentes grupos de individuos y evaluando la eficacia sin esos elementos. Permite valorar si alguno de los elementos es tan esencial que determina el grueso del cambio o si alguno de ellos resulta superfluo. Tiene el inconveniente de que para estudiar no solo las contribuciones de cada elemento, sino las de sus interacciones con otros elementos del programa, se necesita un número de grupos que crece

exponencialmente con el número de elementos independientes en los que se divide la intervención original. Nock *et al.* (2008) citan como ejemplo el estudio de Jacobson *et al.* (1996) sobre los componentes de la terapia cognitivo-conductual para la depresión: activación conductual (AC), entrenamiento en habilidades para manejar pensamientos automáticos (HPA) y esquemas nucleares (EN). En este estudio había tres grupos. El primero recibió solo AC, el segundo recibió AC+HPA y el tercero recibió el tratamiento completo, AC+HPA+EN.

El segundo procedimiento, llamado método *constructivo*, sigue la estrategia inversa, es decir, hay un grupo con solo el elemento más esencial y se van creando grupos añadiendo otros elementos, estudiando así la contribución relativa de cada uno de ellos al éxito final de la terapia. Se puede tratar de los elementos de un paquete terapéutico previamente configurado o también se puede tratar de complementos o añadidos a una terapia específica. Por ejemplo, en algunos estudios de resultados se emplean grupos en los que a la terapia se le añaden instrucciones poshipnóticas (véase, por ejemplo, el meta-análisis de Kirsch, Montgomery y Sapirstein, 1995). De nuevo nos encontramos con el problema práctico de la cantidad de recursos necesarios para llevar a cabo una investigación de este tipo si el número de elementos que queremos incluir en el estudio son solo tres o cuatro.

Otra variante es el llamado diseño *paramétrico*. Su objetivo es determinar lo que podríamos llamar la “dosis” más adecuada (Howard, Kopta, Krause y Orlinsky, 1986). Los grupos tratados se diferencian por elementos como el número de sesiones, su duración, la duración total de la terapia, la administración de ciertos elementos de forma individual o en grupo, etc. Comparando los resultados de estos grupos se puede establecer el número de sesiones o la duración de estas que resultan más eficientes. Habitualmente esta cuestión no se estudia con diseños aleatorizados en grupos de diferente dosis. Son más frecuentes los estudios *a posteriori* en los que los efectos de la dosis se estudian de forma no diseñada o anticipada, valorando la relación entre el número de sesiones que espontáneamente se dieron y las variables clínicas (por ejemplo, Forde *et al.*, 2005; Salzer, Bickman y Lambert, 1999). Esta pregunta es también típicamente abordada mediante el meta-análisis (véase el capítulo 7; Cape, Whittington, Buszewicz, Wallace y Underwood, 2010).

2.1.3. Control de variables

Dado que la validez de las conclusiones en los estudios de resultados depende del control sobre las variables extrañas, no queremos dejar de destacar algunos de los controles principales que se deben aplicar para poder alcanzar conclusiones válidas en un estudio de resultados. Hemos dividido el apartado en las propias medidas que se llevan a cabo sobre los participantes que han sido incluidos en el

ensayo y que principalmente tienen por objeto mantener la comparabilidad de los grupos durante todo el proceso y en aquellas que tienen que ver con la selección de la muestra que va a participar en el propio ensayo. El primero se refiere a controles relacionados con la validez interna y el segundo con la validez externa.

A) Equiparación de los grupos del estudio

La clave del diseño experimental es que los grupos de estudio sean equivalentes, es decir, que se parezcan en todo, salvo en el tratamiento cuyo efecto queremos evaluar. Esta equivalencia debe cumplirse tanto al principio como al final y durante el estudio. Distintas estrategias permiten garantizar esto.

Equivalencia previa. La mejor solución para conseguir grupos que sean equivalentes antes de introducir las manipulaciones es la asignación al azar. Con muestras moderadamente grandes se garantiza que las muestras sean equivalentes, no solo en las variables de interés sino en cualquier otra que sea relevante, aunque el investigador no sea consciente de ello. Como control adicional es habitual comprobar, tras la asignación al azar, que dicha asignación ha hecho su trabajo y los grupos exhiben niveles comparables en la medición hecha antes del tratamiento (véase una opinión contraria sobre la utilidad de estas comparaciones en Harvey, 2018). Obviamente, este control no debe a su vez introducir una nueva fuente de invalidez asociada al uso repetido de instrumentos.

Bloqueo de variables. Para asegurar la equivalencia de los grupos en ocasiones se emplea, adicionalmente a la aleatorización, algún tipo de bloqueo sobre alguna variable relevante de interés. El bloqueo de variables consiste en la formación de bloques de individuos que presentan valores similares en la variable de interés, para posteriormente aleatorizar los bloques de manera independiente. Esta estratificación garantiza que cada grupo experimental contendrá los mismos individuos de cada bloque. Por ejemplo, si bloqueando hacemos una estratificación por sexo, distribuiremos el mismo número de hombres y mujeres en cada grupo de estudio. Una variable de bloqueo habitual es la puntuación en la medida pre, lo que garantiza que ambos grupos sean equivalentes en cuanto a la intensidad de síntomas. Además, el bloqueo de variables obliga a tener en cuenta dicha variable de bloqueo en el análisis de los datos, lo que permite estudiar su papel como posible moderadora de la eficacia de la intervención.

Valoración de la intensidad e integridad de los tratamientos. A veces ocurre que los tratamientos en una o varias de las condiciones no se aplican completamente o con toda la intensidad prescrita. Por ejemplo, a veces los participantes no hacen todos los ejercicios de relajación prescritos, no hacen un autorregistro completo, no se aplican correctamente ciertas técnicas de control del pensamiento o, sencillamente, no acuden a todas las sesiones. Aunque en grupos grandes con

tratamientos básicamente iguales, estas deficiencias de la implementación se distribuirían al azar, el hecho de que en cada grupo se intervenga de forma diferente hace más probable que no se distribuyan aleatoriamente, sino que en alguno de los grupos haya una mayor incidencia de estos problemas. Por ello es imprescindible controlar cuidadosamente que en todos los individuos de todos los grupos se ha aplicado el tratamiento como se diseñó, como se describe en el informe de investigación. Con frecuencia, en los informes de investigación se especifica el mínimo de sesiones para considerar que se ha aplicado la intervención adecuadamente. Por debajo de ese número, el participante queda excluido del estudio.

Homogeneidad de las aplicaciones. Este aspecto se relaciona con el grado de definición de las técnicas que están siendo estudiadas. Si en un estudio participan varios terapeutas, podría darse que cada uno aplicase la técnica introduciendo pequeñas variantes de preferencias personales que podrían en definitiva ser la clave de diferencias importantes en la eficacia. Por ello es conveniente utilizar en estos estudios manuales de tratamiento (protocolos detallados), de los que cada vez disponemos en mayor cantidad y variedad. Es una condición imprescindible para que el lector del informe sepa exactamente qué es lo que se está comparando.

Composición y selección de las muestras de participantes. Ya hemos mencionado la asignación al azar de los participantes a los grupos como medio para garantizar la comparabilidad previa. No se debe confundir con la selección probabilística de los participantes. Mientras el primer procedimiento mejora la validez interna, el segundo mejora la validez externa. En este punto se trata la cuestión de que la muestra sea representativa de la población a la que se pretende generalizar los resultados. Esta representatividad no solo ha de darse al comienzo del estudio, sino que se debe mantener durante su curso completo.

Evaluación de las pérdidas. Otra fuente de amenazas a la validez reside en la reducción de las muestras, generada por la pérdida de participantes (*attrition*) que se produce durante el desarrollo del estudio. Aparte del obvio efecto que la reducción del tamaño de la muestra tiene sobre la potencia de los contrastes estadísticos (validez de la conclusión estadística), el principal problema es que la pérdida no sea aleatoria. Es aleatoria cuando todos los participantes de la muestra inicial tienen la misma probabilidad de permanecer en el estudio. Es no aleatoria cuando está asociada a alguna variable relevante, ya sea de resultados, demográfica o de características de los participantes. De especial relevancia son los casos en los que la pérdida está asociada precisamente al problema que se esté afrontando. De ser así se puede perder la equivalencia original (validez interna) o la representatividad (validez externa).

Aparte de la amenaza a la validez interna y externa que pueda suponer la reducción no aleatoria, se debe analizar la pérdida también como un resultado en sí mismo. Si la asignación a un grupo determinado implica un recorrido en el que se acaba perdiendo un alto número de participantes (tanto los que abandonan como

los que no alcanzan los criterios mínimos para considerar que se ha completado el tratamiento), este resultado debe consignarse en el informe. Por ejemplo, al calcular la tasa de éxitos por grupo se puede calcular como porcentaje sobre los que comenzaron o sobre los que completaron el mínimo tratamiento aceptable. La tasa de individuos que no completan en cada grupo se debe calcular y comparar. Se pueden alcanzar conclusiones en términos de una mayor (o menor) propensión a abandonar o no completar alguno de los tratamientos. Los diagramas de flujo son una herramienta eficaz para comunicar la información a este respecto (véase como ejemplo la figura 2.2). Retomaremos esta cuestión en el capítulo 6.

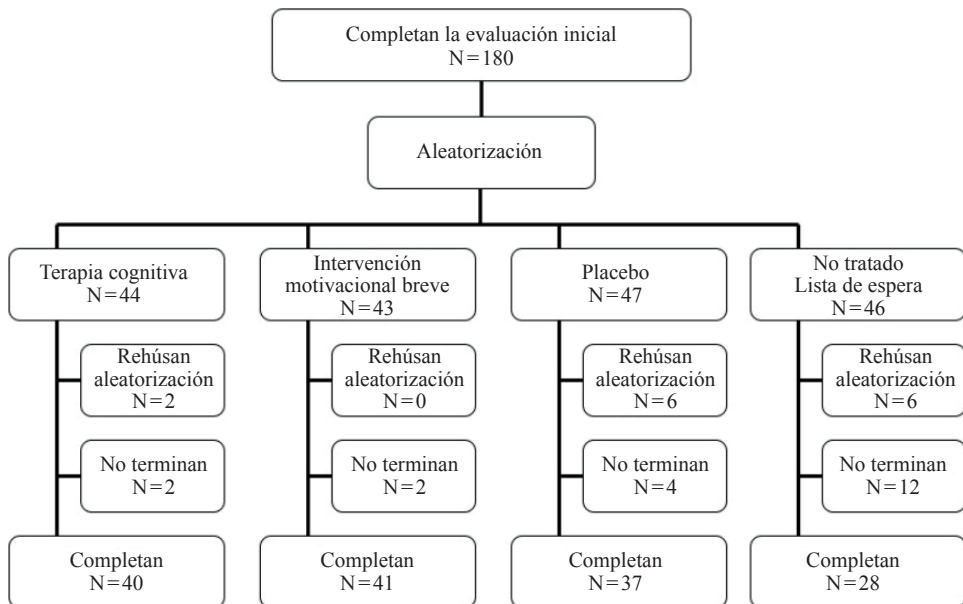


Figura 2.2. Ejemplo de diagrama de flujo del progreso de los participantes en el proceso de un estudio de resultados.

B) Selección de los participantes

En todo estudio se debe dar una definición de la población real a la que se pretende referir las conclusiones. La muestra no es más que una parte de la población, cuya característica de mayor interés para nosotros es que sea representativa de esa población. Para alcanzar esa representatividad lo mejor es hacer una *selección aleatoria*, que se define por la equiprobabilidad de los individuos para formar parte de la muestra, de forma que los no seleccionados son similares a los

seleccionados. Esto se debe hacer empleando un *marco de muestreo* apropiado, aunque en estudios de resultados no es siempre posible.

Dadas esas dificultades, es frecuente que se empleen procedimientos no aleatorios de selección, cuyo impacto potencial sobre la generalizabilidad debe ser analizado. Entre estas están las muestras de conveniencia, definidas por su disponibilidad. Un ejemplo claro son las muestras análogas (por ejemplo, estudiantes de psicología). Otro procedimiento no aleatorio en el que se intenta corregir sus efectos es el *muestreo por cuotas*. También se puede emplear el método boca a oreja o bola de nieve, con el que se consiguen pacientes similares a una pequeña muestra inicial. Los voluntarios que responden a reclamos tienen sus propias peculiaridades.

Respecto a los criterios de inclusión y exclusión, se deben dejar bien establecidos, pero hay que tener en cuenta que estos criterios pueden estar asociados a variables demográficas que pueden afectar mucho a la interpretación de los resultados. Además, una excesiva heterogeneidad muestral dificulta la detección de los efectos en el análisis estadístico. Por otro lado, si los criterios son demasiado restrictivos se puede perder generalizabilidad (validez externa) por una excesiva homogeneidad de la muestra. Festinger y DeMatteo (2008) proponen hacerse cuatro preguntas a la hora de definir estos criterios: ¿cómo afectan a la validez interna y externa?, ¿son necesarios para la seguridad de los pacientes?, ¿son éticos?, ¿cómo van a ser medidos (fiabilidad y validez)? En términos generales, si son excesivamente laxos se pierde rigor, pero si son excesivamente restrictivos se pierde generalizabilidad.

Otra cuestión se refiere al hecho de que las muestras de los estudios de resultados pueden estar constituidas por análogos o por clínicos. Los estudios análogos son aquellos en los que en lugar de trabajar con poblaciones clínicas se trabaja con personas que, aunque son funcionales, exhiben valores extremos en la variable de interés y que tradicionalmente han sido estudiantes universitarios. De este modo, se puede estudiar la eficacia diferencial de dos técnicas de intervención para la depresión seleccionando para ello a los estudiantes de un curso que den mayores puntuaciones en el BDI (*Beck Depression Inventory*), aunque no lleguen a mostrar valores típicos de la población clínica. Estos estudios pueden aportar informaciones muy valiosas, dada la dificultad de acceder a muestras clínicas grandes y de controlarlas, permitiendo la detección de manipulaciones a las que son sensibles las técnicas en estudio. Sin embargo, a la pregunta de si con ellos se reproduce el entorno terapéutico, hay que responder indicando que hay varias diferencias que tradicionalmente se han señalado entre ambos tipos de estudio (gravedad de los trastornos, terapias con poca experiencia, tratamientos más breves y estandarizados, los individuos suelen cobrar en lugar de pagar o la motivación terapéutica).

Todas estas dificultades sugieren una respuesta concreta a la pregunta anterior acerca de la validez externa de los estudios análogos o grado de generalizabilidad

de sus resultados a las poblaciones clínicas reales. Nuestra conclusión es que los estudios análogos son útiles para identificar técnicas que, por ser sensibles a ellas los individuos con problemas de intensidad menor, son candidatas a ser también aplicables de forma eficaz con poblaciones clínicas. Al final, para poder concluir confiablemente acerca de esa eficacia es imprescindible evaluarla con individuos de esta población.

Tanto si son muestras clínicas como análogas, en los informes se deben indicar las características de los pacientes, tanto demográficas como de severidad y antigüedad de los síntomas, los criterios diagnósticos, etc.

2.1.4. El ensayo controlado aleatorizado

De todo lo anterior se deduce que la forma idónea de un estudio de resultados incluye algunas características nucleares. La presencia de dichas características define lo que cada vez se nombra más frecuentemente en psicología como un *ensayo controlado aleatorizado* (ECA). Es este un término importado de la medicina para un tipo de estudios que en psicología se ha llamado tradicionalmente “experimentos”. En cualquier caso, incluimos aquí este nombre porque cada vez se utiliza con mayor frecuencia y con él se facilita la comunicación con los profesionales de otras disciplinas.

Lo que caracteriza principalmente un ECA es:

- a) La inclusión de algún grupo de comparación o control.
- b) La asignación al azar de los individuos a los grupos.
- c) El ejercicio de un buen control sobre las variables extrañas potenciales.

Su forma ideal es aquella en la que se incluyen tratamientos alternativos como forma idónea de controlar el efecto placebo y otros factores contextuales de la terapia. Además, se realizan evaluaciones de las variables de interés tanto antes como después de la intervención.

2.2. Criterios para una terapia empíricamente válida (TEV)

Los criterios propuestos y aceptados por la APA para considerar que una terapia cuenta con una validación empírica fueron los derivados del trabajo de la comisión a la que encomendó la tarea de desarrollarlos. Hoy en día no se utilizan como referencia. Nosotros los incluimos de todas formas porque nos proporcionan la oportunidad de constatar la presencia de los elementos metodológicos que hemos venido exponiendo a lo largo de este capítulo.