Rev. de Psicol. Gral. y Aplic., 1993, *46* (2), 187-198

EVALUACIÓN DE LA EFICACIA DE LOS TRATAMIENTOS PSICOLÓGICOS: UNA PROPUESTA METODOLÓGICA

D. PÁEZ; E. ECHEBURÚA
Universidad del País Vasco

M. BORDA
Universidad de Sevilla

Resumen

Se presenta un diseño cuasiexperimental como alternativa al diseño experimental clásico de evaluación de tratamientos. La comparación entre la medida postratamiento del grupo experimental con la media global de los grupos de control de las investigaciones previas y/o la comparación del tamaño del efecto obtenido con el efecto general de tratamientos previos, obtenido por metaanálisis, es una primera alternativa. La comparación de las puntuaciones postratamiento con una muestra apareada de sujetos normales es la segunda alternativa. La fiabilidad y validez de los instrumentos se pueden contrastar mediante esta muestra general y el grupo de tratamiento. La distribución y las medidas de tendencia central de la muestra aleatoria normal se pueden utilizar como una escala normativa de evaluación —como una «línea de base no causal construida de comparación»—. Esta propuesta metodológica se ejemplifica mediante una investigación sobre el tratamiento de la agorafobia y se muestra que dicha propuesta no sobreestima el tamaño del efecto.

Abstract

A quasi-experimental design is presented as an alternative to experimental designs in order to contrast the effect size of treatment. The comparison between the post mean treatment group and the global mean of control groups of previous research and/or the comparison of the obtained effect with the global effect size, obtained by meta-analysis, is a first alternative. The comparison of post treatment results with a matched sample of normal subjects in the second alternative. The instruments reliability and validity may be contrasted with this general sample and the treatment group. Distribution and central tendency of the random normal sample may be used as a normative scale of evaluation —as a «non causal constructed base line of comparison»—. This methodological proposition is exemplified with a research on agoraphobia and showed that the procedure doesn't overestimate the effect size.

Introducción

Este texto tiene por objetivo reflexionar sobre los problemas en la evaluación de la eficacia diferencial de los tratamientos psicológicos e intervenciones psicosociales y proponer soluciones alternativas extraídas de cuatro líneas de investigación: *a*) el desarrollo de la psicometría y de la epidemiología psiquiátrica descriptiva y explicativa; *b*) el desarrollo de la acumulación y evaluación cuantitativa de resultados mediante el metaanálisis; *c*) el desarrollo de diseños cuasiexperimentales de programas de intervención, y *d*) el desarrollo de la convergencia entre la psicometría y la evaluación conductual (Campbell, 1988; Perachio y Cook, 1988; Rosenthal y Rosnow, 1984; Rothman, 1987; Silva, 1989). A partir de estos avances metodológicos contemporáneos, se presenta en este trabajo una línea de investigación que busca legitimar la evaluación de los tratamientos con medidas antes-después y con la comparación de las medias de una muestra normativa amplia. Se trata asimismo de sacar partido a las medidas realizadas para obtener el máximo de información posible para contrastar las hipótesis.

La aplicación de las propuestas metodológicas presentadas en este trabajo se lleva a cabo, a modo de ejemplo ilustrativo, sobre una investigación referida al tratamiento de la agorafobia (Echeburúa, Corral, García y Borda, 1991; véase anexo). No obstante, y a efectos de simplificar la presentación de dichas propuestas metodológicas, los pacientes tratados se consideran como un único grupo de trata-

miento, pese a haber habido en el estudio original cuatro modalidades terapéuticas (véase Echeburúa y colaboradores, 1991).

Problemas metodológicos en la evaluación de tratamientos

La evaluación de la eficacia diferencial de los tratamientos psicológicos se enfrenta con una serie de problemas, de los que dos son de particular relevancia: *a*) el tamaño de las muestras utilizadas y las dificultades para obtener grandes muestras homogéneas de pacientes, y *b*) el problema deontológico y práctico para conseguir un grupo de control. Es sabido que los sujetos en una lista de espera prolongada pierden motivación, se sienten discriminados y, en último término, abandonan con gran facilidad la investigación (Beck, Andrasik y Arena, 1989).

En lo referente al tamaño de las muestras, hay que recordar que la interpretación correcta de las técnicas estadísticas utilizadas (t-test, Chi cuadrado, análisis de la varianza simple o múltiple, etc.) requiere un tamaño mínimo de las muestras de 20-25 sujetos por grupo y un tamaño ideal de 50-100 sujetos por grupo. Recuérdese que para protegerse contra un error de tipo 1 (afirmar que hay un efecto del tratamiento cuando esto no es real) se adopta un nivel de probabilidades —convencionalmente el de 0,05—. Esto quiere decir que cuando se parte de la hipótesis nula (no hay diferencias entre las medidas o las proporciones) y se observa una desviación de la media general o común a ambos grupos que se dé sólo 5 veces sobre 100 o menos, en ese caso se rechaza la hipótesis nula. Lo que se olvida muchas veces es que otros tres elementos fundamentales son también la fiabilidad de los instrumentos, el tamaño del efecto (la magnitud de la diferencia de medias) y el tamaño de la muestra. Cuanto mayor sea la fiabilidad del instrumento, mayor sea la magnitud del efecto y mayor sea el tamaño de la muestra, mayor será la probabilidad de detectar las diferencias o, en otras palabras, menor será la probabilidad de cometer un error del tipo 2 (afirmar que no hay diferencias cuando en realidad existen).

Con una muestra de 10 pacientes por grupo, la probabilidad de detectar una capacidad de mejoría del 20 por 100 del tratamiento en el grupo experimental en relación con un grupo de control es de sólo el 19 por 100 si se recurre al umbral del 0,05. Aun con 25 pacientes por grupo —con el mismo tamaño del efecto y el mismo umbral anterior—, la probabilidad de detección es menor del 40 por 100. Con 50 pacientes por grupo, la probabilidad es del 64 por 100 y con 100 pacientes por grupo asciende hasta el 88 por 100 (Chassan, 1979). Sin embargo, la posibilidad real de obtener esos tamaños de muestras en la investigación con problemas clínicos graves es bajísima. En este trabajo se plantean alternativas metodológicas para solventar este problema. Además del tamaño de la muestra y del tamaño del efecto, otro elemento importante, como ya se ha señalado, es el referido a la fiabilidad y validez de los instrumentos de medición. Con nuestros datos y nuestras escalas ilustraremos cómo se puede examinar este problema.

Fiabilidad, validez, sensibilidad y especificidad del instrumento utilizado en la evaluación del tratamiento

El instrumento utilizado en esta investigación para determinar el resultado del tratamiento ha sido el *Inventario de Agorafobia,* de Echeburúa y Corral (Echeburúa, Corral, García, Páez, Borda, 1991). Está dividido en dos partes. La primera mide en una escala de 0 a 5 puntos diferentes tipos de respuestas (motoras, psicofisiológicas subjetivas y cognitivas). La segunda parte mide la variabilidad de las respuestas en función de los factores que contribuyen a aumentar y reducir la ansiedad.

La fiabilidad —el grado de constancia de las medidas de un test— se determina por la consistencia interna (evaluada por el método de las dos mitades o por los coeficientes alfa, etha u omega) o por la estabilidad temporal (fiabilidad test-retest). El coeficiente de consistencia interna obtenido mediante la utilización del método de las dos mitades es de 0,75 en una muestra de 357 sujetos «normales» y de 0,83 en una muestra de 29 agorafóbicos en la fase de pretratamiento (significativo en ambos casos). Asimismo, los coeficientes alfa de las tres dimensiones del inventario (conductas motoras, psicofisiológicas subjetivas y cognitivas) superan el 0,90 tanto en la muestra de sujetos «normales» como en la de agorafóbicos.

La validez —el grado en que un test cumple el fin al que se destina— puede determinarse fundamentalmente a través de la validez de constructo, de la validez convergente y de la validez discriminante. La validez de constructo es jerárquicamente la más importante ya que nos permite generalizar de los indicadores al área teórica de referencia. La validez de constructo, considerada como la relación convergente entre los tres tipos de respuestas del *Inventario de Agorafobia* y obtenida a partir de una muestra

TABLA 1

Coeficiente de correlación de Pearson entre las dimensiones del inventario de agorafobia

	N = 357	N = 29
R. motoras – R. cognitivas	r = 0,347 **	r = 0,760 **
R. motoras – R. psicofisiológicas	r = 0,404 **	r = 0,707 **
R. cognitivas – R. psicofisiológicas	r = 0,608 **	r = 0,861 **

* $p < 0,001$.

de 357 sujetos «normales» y de 29 agorafóbicos, es satisfactoria, tal como queda expuesto en la tabla 1. Que los pacientes muestren asociaciones más fuertes que los «normales» concuerda con el hecho de que estos sujetos son «expertos» en el área afectiva en cuestión, por lo que pueden reflejar una mayor estructuración en las respuestas emitidas (al menos, a nivel de ítem de autopercepción).

La validez convergente, obtenida a partir de la Subescala de Agorafobia del *Cuestionario de Miedos* (Marks y Mathews, 1979) y de las tres dimensiones de respuestas del *Inventario de Agorafobia* y realizada con la muestra de 29 sujetos agorafóbicos, es satisfactoria, tanto en general como por dimensiones (0,89 en las respuestas motoras, 0,75 en las respuestas psicofisiológicas y 0,74 en las respuestas cognitivas).

La validez discriminante, obtenida mediante el análisis comparativo de las respuestas dadas en el *Inventario de Agorafobia* por una muestra de 357 sujetos «normales» y de 29 agorafóbicos antes del tratamiento, indica que las diferentes dimensiones del inventario discriminan significativamente entre sujetos normales y agorafóbicos (tabla 2). De hecho, la función discriminante del SPSS-X clasifica correctamente al 85 por 100 de los sujetos, por lo que la eficacia diagnóstica del inventario es alta. Además, este instrumento de medida tiene una especificidad del 85 por 100 y una tasa de falsos positivos del 15 por 100. Asimismo, el *Inventario de Agorafobia* tiene una sensibilidad del 83 por 100 y una tasa de falsos negativos del 17 por 100. Estos resultados en una escala de autopercepción de síntomas reflejan una eficacia diagnóstica alta y pueden considerarse satisfactorios si se tiene en cuenta que, por ejemplo, la entrevista diagnóstica de evaluación normalizada de trastornos depresivos de la OMS —un instrumento mucho más costoso en recursos humanos— tiene una especificidad del 90 por 100 y una sensibilidad media del 86 por 100 (Sartorius, 1983).

Del mismo modo, el análisis comparativo de las respuestas dadas por la muestra de 29 pacientes del estudio antes y después de la terapia señala que las diferentes dimensiones del *Inventario de Agorafobia* discriminan entre las respuestas anteriores y posteriores al tratamiento y sirven, por tanto, para detectar el cambio terapéutico en la agorafobia (Echeburúa y cols., 1991). En resumen, gracias a las medidas repetidas y a la comparación con un grupo de control, se ha podido obtener una primera serie de informaciones relevantes para el contraste de hipótesis.

Estos resultados nos dan una información complementaria en lo referente a la validez y fiabilidad del instrumento. Recuérdese que los conceptos de sensibilidad y de especificidad en epidemiología se refieren a la certeza con que un instrumento detecta una señal verdadera contra un fondo de información incorrecta. La probabilidad condicional de detectar un signo o síntoma cuando la enfermedad está presente es la sensibilidad —o, en términos de la teoría de detección de señales, un «hit»—. La probabilidad condicional de detectar un signo o síntoma cuando la enfermedad está ausente es la razón de falsa alarma. De este modo, se dicotomizan las medidas con el objetivo de aumentar la certeza de detectar la enfermedad —aunque esto se haga a costa de ignorar la variabilidad y, paradójicamente, de reducir la fiabilidad (Mirowsky y Ross, 1989).

Ejemplifiquemos este hecho con el *Inventario de Depresión de Beck*. Supongamos que este instrumento se aplique como una entrevista y se evalúe su fiabilidad mediante la certeza del acuerdo en que clasifican a los sujetos dos entrevistadores. Si se establece un solo punto de corte (20 puntos, por ejemplo) y se divide a la población en depresivos y no-depresivos, se logra una mayor certeza o proba-

TABLA 2

Eficacia diagnóstica del «Inventario de Agorafobia»

Grupos de hecho	Grupos predecidos por función discriminante «Normales»	Agorafóbicos	Total
«Normales»	304	53	357
Agorafóbicos	5	24	29

$$\text{Sensibilidad} = \frac{\text{Verdaderos positivos}}{\text{Total de pacientes}} \times 100 = \frac{24}{29} \times 100 = 83$$

$$\text{Especificidad} = 100 - \frac{\text{Falsos positivos}}{\text{Total de normales}} \times 100 = 100 - \frac{53}{357} \times 100 = 85$$

$$\text{Eficacia diagnóstica} = \frac{\text{Total de bien clasificados}}{\text{Total global}} \times 100 = \frac{328}{386} \times 100 = 85$$

bilidad de acuerdo entre entrevistadores. Sin embargo, de este modo se considera a un sujeto con 15 puntos igual que a un sujeto con 0 puntos. La certeza de acuerdo es mayor, pero al precio de una menor precisión o fiabilidad con la que el instrumento mide la depresión.

Evaluación del tamaño del efecto y metaanálisis

A partir de los años ochenta se ha comenzado a recurrir, a la hora de comparar la eficacia de diversos tratamientos, a técnicas estadísticas más elaboradas que la mera contrastación de las tasas de éxito. El metaanálisis —un avance en los diseños cuasiexperimentales mediante la utilización de análisis múltiples— utiliza el concepto de «tamaño del efecto» y posibilita la reevaluación *post hoc* de las investigaciones sobre valoración de los resultados terapéuticos (Brown, 1987; Kazdin, 1985; Labrador, 1986; Perachio y Cook, 1988; Wilson, 1985). Este procedimiento supone una alternativa a la evaluación cualitativa tradicional, al proporcionar una medida objetiva y cuantitativa para evaluar la bibliografía sobre resultados (Barker, Funk y Houston, 1988; Clum y Bowers, 1990; Matt, 1989).

Los pasos de un metaanálisis son los siguientes (Fitz-Gibbon, 1984): *a*) búsqueda de estudios; *b*) codificación de las características de los estudios; *c*) medida del «tamaño del efecto», que queda indicado por la localización de la media del grupo E como una puntuación *z* en la distribución del grupo de control, y *d*) correlación del «tamaño del efecto» con las variables contextuales para averiguar si hay relaciones entre las magnitudes del efecto y los contextos en los que se han hallado los efectos.

Mediante este procedimiento, el «tamaño del efecto» obtenido en el tratamiento psicoterapéutico se aproxima a una desviación típica. Este «tamaño del efecto» se obtiene generalmente sustrayendo a la media de síntomas postratamiento del grupo de terapia la media del grupo de control (grupo con placebo o sin terapia) y dividiendo la cantidad resultante por la desviación típica del grupo de control más la del grupo experimental.

Según este procedimiento, las terapias cognitivo-conductuales son significativamente más eficaces que las psicodinámicas, con una media de efecto de 0,97 en el primer caso y de 0,74 en el segundo (esta diferencia, según la prueba de *z*, es significativa). Sin embargo, estas diferencias tienden a desaparecer cuando se seleccionan sólo los estudios que recurren a poblaciones clínicas (y no a muestras de sujetos análogos) y cuando se controla la reactividad de la medida de mejoría —las terapias conductuales utilizan medidas de evaluación más directas y que suscitan mayor reactividad—. En trastornos menos graves, la media común del efecto en ambos tipos de terapias es de 1,11 (1,28 en el caso de las fobias) y en trastornos más graves de 0,68 (0,67 en el caso de depresión y de la ansiedad) (Castillo y Poch, 1991). Estos resultados se obtienen a pesar de que los terapeutas que forman parte de las investigaciones controladas tienden a tener una experiencia clínica más bien pequeña y de que el número de sesiones terapéuticas es más bien escaso (una mediana de 12 sesiones por tratamiento). La conclusión general es que las personas que reciben terapia están mejor al final de ella que el 80 por 100 de las que no se someten a tratamiento (Berman, Miller y Massman, 1985; Parloff, London y Wolfe, 1986; Rachman y Wilson, 1983; Rosenthal, 1983; Shapiro y Shapiro, 1982).

En nuestra investigación sobre la agorafobia, el efecto global de las modalidades terapéuticas, referido a la evaluación pre-postratamiento es de 1,28 desviaciones típicas sin corrección y de 1,54 con la corrección para medidas repetidas (tabla 3), lo que sugiere que no se ha sobreestimado el efecto del tratamiento.

Para determinar en qué medida este efecto es superior al efecto medio de los tratamientos psicológicos, se va a utilizar una segunda forma de estimar el tamaño del efecto: el coeficiente *r* entre el grupo de

TABLA 3

Tamaño del efecto en la investigación de la agorafobia

$$\text{Cálculo} = \frac{\text{Media antes} - \text{Media después}}{\text{Desviación típica en el pretratamiento}} = \frac{197{,}5 - 126{,}6}{55{,}4} = 1{,}28$$

Corrección del cálculo para medidas repetidas

$$\frac{\text{Media antes} - \text{Media después}}{\sqrt{SD1^2 + SD2^2 - 2\,(r12\,(SD1 \times SD2))}} = \frac{242{,}0 - 163{,}84}{\sqrt{65{,}6^2 + 75{,}7^2 - 2\,(0{,}7\,(65{,}6 \times 75{,}7))}} = 1{,}54$$

tratamiento y el grupo de control. Este segundo procedimiento consiste en transformar las pruebas de hipótesis en correlaciones de Pearson (0 = grupo de control; 1 = grupo de tratamiento). Las *r* obtenidas por investigación se transforman en *z* y se ponderan por el *N* a partir del cual se han obtenido. De este modo, se puede establecer un tamaño del efecto retransformando las *z* promedio en *r* (Rosenthal y Rosnow, 1984).

La fórmula expuesta en la tabla 4 es la propuesta por Rosenthal (1983) para la comparación de estudios, con la particularidad de que la *z* es en este caso la *z* media de los estudios revisados por Smith y Glass (1977) (un total de 475 estudios y 25.000 sujetos).

Bajo la forma de *r*, el tamaño medio del efecto del tratamiento —antes expuesto como de 0,85 o de una desviación típica— es de 0,32. La *r* común a los cuatro grupos de tratamiento en nuestra investigación es de 0,49. Si se transforman las *r* en *z* y se restan, la diferencia es significativa, por lo que los resultados obtenidos en nuestra investigación son superiores a los obtenidos por la media global de los estudios revisados.

Estos resultados permiten ilustrar la potencialidad del metaanálisis en la evaluación de los tratamientos. No obstante, hay algunas hipótesis alternativas que pueden cuestionar la validez interna de nuestra investigación, basada en la comparación pre-postratamiento. En primer lugar, el paso del tiempo por sí solo puede producir una mejoría de los síntomas. Si bien esto es cierto en la ansiedad y depresión, esta objeción no parece aplicable a la agorafobia, que tiende a ser más bien estable con el transcurso del tiempo (Echeburúa y cols., 1991). Y, en segundo lugar, las medidas repetidas tienden a reflejar una cierta regresión estadística a la media. Este hecho, puesto de relieve en diversas investigaciones (por ejemplo, Jacobson, Wilson y Tupper, 1988), no explica, sin embargo, más que una pequeña parte de la varianza y no altera sustancialmente la significación de los resultados de nuestra investigación.

La inexistencia de un grupo de control en nuestra investigación, necesario para obtener el máximo rendimiento en la aplicación del metaanálisis, se puede compensar de dos modos. Una primera forma es comparar las medias de los grupos de nuestra investigación con la media global del instrumento en los grupos de control de investigaciones previas, lo que presupone la existencia de grupos diagnósticos homogéneos y la utilización de instrumentos de medida comunes. Una segunda, comparar nuestros resultados con una muestra amplia de sujetos «normales». Este segundo método se inspira en el desarrollo de los diseños cuasiexperimentales.

Evaluación de tramientos y diseño de línea base no-causal construida

A nivel cuantitativo, un diseño de evaluación de programas de intervención que se podría postular como poco costoso y factible de construir mediante la aplicación de desarrollos metodológicos cuasiexperimentales es el diseño de línea base no-causal construida, que no es sino una ampliación de la vieja técnica psicométrica de la baremación (Perachio y Cook, 1988). En este diseño, el grupo de control no equivalente es reemplazado por la comparación con una puntuación baremada de una muestra representativa. Si se construyera un baremo de percentiles (a partir, por ejemplo, de una muestra representativa de 800 sujetos) sobre ansiedad, depresión, competencia social, apoyo social, etc., se podría evaluar la eficacia de los tratamientos psicológicos y psiquiátricos en un país determinado mediante la comparación con este baremo. En lugar de comparar la evaluación pretratamiento con la evaluación postra-

TABLA 4

Fórmula de Rosenthal para contraste de heterogeneidad de efectos

$$\frac{\text{Equivalente z de r específico } Z1 \; - \; \text{Equivalente z de r general } Z2}{\sqrt{\frac{1}{N1-3} + \frac{1}{N2-3}}}$$

$$\frac{0{,}536 - 0{,}332}{\sqrt{\frac{1}{30-3} + \frac{1}{25.000-3}}} = \frac{0{,}204}{0{,}13} = 1{,}56 \quad p < 0{,}06$$

tamiento, con las limitaciones de validez interna ya señaladas, se puede comparar el percentil medio en el pre-test con el percentil medio en el post-test. La diferencia entre ambas medias permite estimar la fuerza de la intervención y facilita los procedimientos de acumulación evaluativa mediante el metaanálisis. Asimismo, si se igualan los grupos en, por ejemplo, edad, sexo y nivel de ingresos, se pueden comparar el pre-test y el post-test del grupo de tratamiento con un grupo apareado de la muestra representativa (véase Perachio y Cook, 1988).

Los problemas de este diseño se pueden superar con una sola investigación en una muestra normativa amplia con medidas repetidas. A nuestro entender, este diseño es similar al diseño epidemiológico explicativo de *cohorte*, sólo que este último se halla referido más bien a la acción de un factor de riesgo y no a la de una intervención terapéutica. En un diseño de cohorte se compara prospectivamente una cohorte sometida al factor de riesgo con otra en la que éste está ausente. Conceptualmente, se puede considerar al diseño cuasiexperimental de medidas repetidas con grupos no equivalentes de evaluación de tratamiento, como un estudio de cohorte —salvo que en el primer caso se evalúa la acción de un factor de protección o de curación y que los grupos deben diferir antes (uno debe estar sano y el otro enfermo) y no después.

Con el objetivo de reducir el coste de los diseños de cohorte, Rothman (1987) sugiere medidas que son de aplicación también en los diseños de línea base no-causal construida:

> Una forma de disminuir el gasto es reemplazar una de las cohortes, en concreto, la no-expuesta, con información de la población general. En lugar de recoger información nueva de una población no-expuesta de gran tamaño, se usan, para la comparación, datos existentes acerca de la población general. Este proceder tiene problemas obvios. En primer lugar, es razonable sólo si existe una cierta seguridad de que solamente una pequeña proporción de la población general está expuesta al agente que se estudia. En la medida en que parte de la población general esté expuesta, existirá un error de clasificación incorrecta que introducirá un sesgo en la comparación, en el sentido de subestimar el efecto. Segundo, la información que se obtenga en el estudio deberá ser comparable en calidad a la ya existente para la población general. (Rothman, 1987, pág. 71.)

El diseño de línea base no-causal construida debe estar sujeto también a estas mismas exigencias. En primer lugar, se puede suponer que los tratamientos psicológicos afectan a sólo una pequeña parte de la población, por lo que la presencia del factor de protección en la población general será limitado. En segundo lugar, no siempre se dispone de grupos de tratamiento que respondan a los criterios sociodemográficos de la muestra normativa (por ejemplo, en lo relativo a la edad, el sexo, la ocupación, etc.), cuando el modelo de comparación baremada supone que el grupo de tratamiento y el grupo de control provienen de poblaciones equivalentes en cuanto a historia, tiempo, lugar y características demográficas. Si esto no es así —y una limitación de los tests baremados disponibles es la baremación sobre muestras no siempre representativas—, se cuestiona un supuesto básico: que los intervalos de cambio son estables en el tiempo y que son los mismos para la muestra normativa y para los grupos de tratamiento. Una forma de resolver este problema es utilizar baremos extraídos de muestras del mismo ámbito geográfico al que pertenecen los pacientes y utilizar como grupo de comparación a sujetos «normales» dentro del rango de edad, sexo, profesión, etc., predominante entre los pacientes.

En tercer lugar, el modelo de comparación baremada supone que un cambio en los puntos brutos es igual al número de cambios en el percentil de la población, pero esto no es así ya que los resultados son distintos según sea el nivel de partida. De hecho, si se parte de percentiles medios con menos puntos, se consiguen más cambios (Perachio y Cook, 1988). Y en cuarto lugar, las limitaciones genéricas de los diseños cuantitativos (prestar atención a la evaluación del *resultado* de las terapias psicológicas, pero no así a la evaluación del *proceso* de las mismas, es decir, a cómo funcionan y a qué tipo de factores se atribuye el cambio del comportamiento) se mantienen también en este caso (Clemente, 1989; Hersen y Michelson, 1986).

En nuestra investigación, la distribución por percentiles en el *Inventario de Agorafobia* de una muestra normativa aleatoria de 357 sujetos procedentes del País Vasco figura en la tabla 5.

Una primera forma de proceder con este método es comparar las puntuaciones medias de antes y de después del tratamiento en los grupos terapéuticos con los percentiles de la muestra normativa. La puntuación media de los grupos de pacientes en dicho inventario antes del tratamiento era de 197,5, que supone un percentil 98 de la muestra normativa y

TABLA 5

Distribución por percentiles en el inventario de agorafobia de la muestra normativa (N=357)

Percentil	Puntuación
10	67
20	78
30	90
40	101
50	117
60	129
70	140
80	157
90	172

que significa que los pacientes medios se encontraban peor que el 98 por 100 de los sujetos de la muestra normativa en la dimensión evaluada. Tras el tratamiento, la puntuación media de los grupos terapéuticos es de 126,55, que se corresponde con el percentil 67 de la muestra normativa. Estimado de esta manera, el efecto del tratamiento es de:

$$\frac{98 - 67}{100} = 0{,}31$$

Los resultados de este trabajo, de alcance limitado, son, sin embargo, coincidentes con otros que muestran que un 30 por 100 de los agorafóbicos no se benefician de los tratamientos conductuales y que el 70 por 100 restante mejorado puede quedar con algunos síntomas residuales (Jacobson y cols., 1988; Parloff y cols., 1986).

Si en lugar de tomar los percentiles de la muestra normativa hubiéramos utilizado los de la muestra de pacientes en el pretratamiento, habríamos supuesto que los grupos de pacientes en el postratamiento, con una puntuación media de 126,60 y con un percentil 10 (aproximadamente) en la escala de puntuaciones anterior al tratamiento, estarían mejor que el 90 por 100 de los pacientes no tratados. Recuérdese que la estimación metaanalítica de mejoría era del 80 por 100, por lo que la sobreestimación no resulta tan alta.

Una segunda forma de utilizar la comparación entre los grupos terapéuticos y la muestra normativa es comparar los contrastes de hipótesis antes y después del tratamiento. Si la terapia actúa eficazmente, la comparación postratamiento entre ambos grupos debe ser menos significativa que en el pretratamiento y se puede estimar la probabilidad de que las significaciones difieran. Los resultados con este enfoque figuran en la tabla 6.

Como se puede constatar de la comparación entre ambos grupos, las diferencias son significativas en el pretratamiento, pero sólo tendenciales después del tratamiento. La atribución de un cambio significativo resulta correcta y refuerza la inferencia sobre la eficacia de la terapia.

TABLA 6

Pruebas de T-Test entre la muestra normativa y los grupos terapéuticos

	Muestra normativa	Grupos terapéuticos
Antes de tratamiento	$\bar{X} = 118{,}1$ DT = 39,7	$\bar{X} = 197{,}5$ DT = 55,4
	$t = -7{,}57$ $p < 0{,}000$	
Después del tratamiento		$\bar{X} = 126{,}6$ DT = 62,2
	$t = -0{,}92$ $p < 0{,}37$	

Valores de t para dos colas.

Una tercera forma de utilizar la muestra normativa es la de establecer puntos de corte para obtener un indicador de recuperación significativo que tenga una base estadística. Según Jacobson y Truax (1991), hay cuatro fórmulas para establecer estadísticamente un punto de corte:

1. Coeficiente C1: Cuando las desviaciones típicas del grupo de tratamiento y de la muestra de la población general son similares (circunstancias poco habituales).
2. Coeficiente C2: Cuando las desviaciones típicas entre el grupo disfuncional y la muestra normal son diferentes.
3. Coeficiente C3: Cuando no hay una muestra normal de comparación. Si bien esta fórmula es más exigente que C1 y C2, cuando las distribuciones entre la muestra normal y la disfuncional no se solapan esta fórmula es muy «liberal».
4. Coeficiente C4: Cuando hay una muestra normal. Este último coeficiente se puede utilizar de forma más exigente aplicando la misma fórmula con una desviación típica, lo que, en el ejemplo propuesto, daría un punto de corte de 157,8 (similar a C1 y C2). Éste puede ser un punto de recuperación satisfactorio, según el cual los sujetos después del tratamiento tienen que estar mejor que el 32 por 100 de la muestra normal. Si se aplica este criterio a las puntuaciones del pos-test, el 76 por 100 de los sujetos se han recuperado. En cambio, si se aplica la fórmula más exigente —la C1—, sólo se han recuperado el 34 por 100 (tabla 7). Según Jacobson y Truax (1991), lo preferible es aplicar las fórmulas C2 y C4 a la C1. Ambas fórmulas utilizan estadísticamente la información del diseño de línea de base no-causal construida que se ha expuesto en este trabajo.

Otra crítica clásica que se ha hecho a la investigación sobre tratamientos es que, al basarse en comparaciones de medias, oculta los cambios individuales. Un tratamiento puede aparecer como más eficaz, aunque mejore el estado de algunas personas solamente. Asimismo, las medias pueden permanecer constantes y no revelar que un subgrupo ha mejorado y otro empeorado. Una solución a este problema es examinar las puntuaciones individuales de cambio, clasificar a los sujetos en tres categorías (recuperados, mejorados pero no recuperados y sin mejoría) y aplicar el Chi cuadrado, por ejemplo, como estadístico de contraste.

Se ha criticado, sin embargo, la fiabilidad de las puntuaciones individuales de cambio porque hay una correlación inversa entre las puntuaciones de cambio y la puntuación inicial, y porque las puntuaciones de cambio dependen del nivel inicial (Willet, Ayoub y Robinson, 1991). La alternativa es la puntuación de cambio tipificada, que consiste en apreciar el cambio de cada sujeto a partir del cambio medio observado en los sujetos que han partido del mismo nivel de puntuación. También se ha plantea-

TABLA 7

Fórmulas para calcular C

Fórmula general	Aplicación a nuestro ejemplo
I) El coeficiente $C1 = \frac{Mo - M1}{2}$	$\frac{118,1 + 197,5}{2} = 157,8$
II) $C2 = \frac{DTo \times M1 + DT1 \times Mo}{DTo + DT1}$	$\frac{(39,7 \times 197,5) + (5.547 \times 118,1)}{39,7 + 55,4} = 151,2$
III) $C3 = M1 - (2\ DT1)$	$197,5 - (2 \times 55,4) = 86,7$
IV) $C4 = Mo + (2\ DTo)$	$118,1 + 2 \times 39,7 = 197,5$

do recurrir a la puntuación de cambio utilizando como covariable en el análisis la puntuación inicial (Gerin, 1984; Gerin, Dazord y Sali, 1991).

No obstante, las puntuaciones de cambio no son siempre de baja fiabilidad y, por otro lado, los resultados obtenidos con puntuaciones de cambio estandarizadas y puntuaciones de cambio brutas son similares (Gerin y cols., 1991; Willet y cols., 1991). Por ello, estos últimos autores han sugerido la corrección del error de medición como forma de mejorar las puntuaciones de cambio. A su vez, Jacobson y Truax (1991) han elaborado una puntuación de cambio precisa o fiable según la cual se puede establecer en qué medida ha cambiado cada sujeto según la variabilidad del error. La fórmula figura en la tabla 8.

Una puntuación de cambio precisa superior a 1,96 es poco probable (con una probabilidad de ocurrencia menor de 0,05) en el caso de que no haya un cambio real. Todo cambio superior a 1,96 se puede considerar una recuperación real.

Tal como se ve en la tabla 8, si se corrige el error de medición, el 34 por 100 de los sujetos experimenta un cambio individual significativo. Además, todos los sujetos recuperados según el coeficiente PCP superan el punto de corte C_3. Se confirma de este modo la validez convergente de ambos conceptos. La media de cambio según el PCP es de 1,5, convergente con la estimación de cambio antes-después (1,54) tras corregir el error de medición. De 10 sujetos recuperados, 6 superan el punto de corte más exigente C_1. Con todo ello se pueden realizar tanto estimaciones de cambio más finas como evaluar individualmente a los sujetos. Por ejemplo, un criterio de éxito terapéutico clínicamente significativo sería experimentar un cambio positivo superior a 1 PCP y que la puntuación del pos-test fuera inferior al punto de corte C_1.

Una forma de combinar toda la información del diseño de línea de base no-causal sería la de transformar las puntuaciones brutas de los sujetos en tratamiento en percentiles de la muestra normativa, calcular sobre ello el PCP y utilizar esta puntuación de cambio para evaluar los tratamientos. En este caso, se aplican los mismos cálculos anteriores pero a partir de los deciles. La correlación entre el pre y el postratamiento de la agorafobia es de 0,49 (menor que 0,01). Aunque significativo, se trata de un coeficiente de fiabilidad test-retest bajo. Si se calcula la D diff a partir de los nuevos datos, se obtienen los resultados expuestos en la tabla 9.

El examen de esta tabla condensa el total de información del diseño de línea de base no-causal y permite un examen individual. Si se examina el postest se puede saber en qué decil está el sujeto. Al examinar la diferencia, se puede determinar la cuantía del cambio en relación con una puntuación normativa. Y si se examina el PCP, se sabe la certeza del cambio. De este modo, se puede asegurar que menos de un 44 por 100 de los sujetos tiene un cambio significativo. De estos 13 sujetos, 11 se sitúan debajo del decil 5, es decir, están igual o mejor que la mitad de una muestra normal, con un cambio claro y clínicamente significativo. Un 21 por 100 (n = 6) no muestra cambios, y un 7 por 100 (n = 2) sólo mejora ligeramente. De los 6 sujetos que no mejoran, todos permanecen en el decil 10, es decir, están igual o peor que el 90 por 100 de la muestra normal. Los 2 sujetos que mejoran levemente siguen estando por encima del decil 8 (peor que el 80 por 100 de la muestra normal). El 28 por 100 de sujetos que se recuperan ligeramente (n = 8) se puede dividir en dos grupos: los que empiezan con un decil bajo y el tratamiento les sitúa por debajo del decil 5 (sujetos números 11 y 13) y los que empiezan con una puntuación muy alta y el tratamiento los hace retroceder entre 2 y 3 deciles (los sujetos 9, 10 y 29 pasan de 10 a 7, el sujeto número 5 pasa de 10 a 8 y el número 8 de 7 a 4).

Estos resultados son congruentes con las revisiones metaanalíticas antes citadas sobre el efecto de los tratamientos en la agorafobia (Jacobson y colaboradores, 1988). Aunque el efecto de los tratamientos es significativo, menos del 50 por 100 de

TABLA 8

Puntuaciones individuales de cambio significativo calculadas a partir de la fórmula expuesta debajo

Sujeto	Pretest	Postest	Cortes C_3 C_4 C_2	Difer.	PCP	Recup.
2	245,00	207,00	N N N	38,0	0,88	M
3	195,00	91,00	N S S	104,0	2,42	R
4	159,00	118,00	N S S	41,0	0,95	M
5	305,00	147,00	N S S	158,0	3,68	R
6	195,00	57,00	S S S	138,0	3,21	R
7	134,00	72,00	S S S	62,0	1,44	M
8	137,00	97,00	N S S	40,0	0,93	M
9	210,00	130,00	N S S	80,0	1,86	M
10	180,00	130,00	N S S	50,0	1,16	M
11	115,00	80,00	S S S	35,0	0,81	M
12	205,00	190,00	N N N	15,0	0,34	M
13	95,00	57,00	S S S	38,0	0,88	M
15	194,00	58,00	S S S	136,0	3,17	R
16	182,00	249,00	N N N	−67,0	−1,56	E
17	267,00	227,00	N N N	40,0	0,93	M
18	184,00	129,00	N S S	55,0	1,28	RL
19	239,00	158,00	N S N	81,0	1,88	RL
20	237,00	81,00	S S S	156,0	3,63	R
21	141,00	71,00	S S S	70,0	1,63	RL
22	320,00	264,00	N N N	56,0	1,30	RL
23	169,00	141,00	N S S	28,0	0,65	M
24	243,00	248,00	N N N	−5,0	−0,11	E
25	168,00	63,00	S S S	105,0	2,44	R
26	149,00	64,00	S S S	85,0	1,98	R
27	163,00	92,00	N S S	71,0	1,65	RL
28	180,00	72,00	S S S	108,0	2,51	R
29	262,00	135,00	N S S	127,0	2,96	R
30	179,00	99,00	N S S	80,0	1,86	RL
31	276,00	143,00	N S S	133,0	3,10	R

S = Sí o N = no recuperado según punto de corte C_3 (86,7), C_4 (197,5) o C_2 (151,2).
E = Empeoramiento del sujeto, cambio negativo.
M = Cambio positivo, pero inferior a 1 PCP. Mejoría sin recuperación total.
RL = Cambio positivo importante, superior a uno, pero inferior a 1,96.
R = Cambio positivo importante y muy poco probable que sea debido al azar.

Nota. Los números de sujetos correspondientes a las 29 personas con datos completos de la matríz. El sujeto 1 y el 14 tenían datos desaparecidos, por lo que se excluyeron del análisis.

Fórmula para calcular PCP

Fórmula

$$PCP = \frac{M1 - M2}{D\ diff}$$

$$D\ diff = \sqrt{2\,(DE)^2}$$

$$DE = DT1\ \sqrt{(1 - r12)}$$

Aplicación a nuestro ejemplo

$$\text{Sujeto } 2 = \frac{245 - 207}{42{,}9} = 0{,}88$$

$$\sqrt{2 \times (30{,}3)^2} = 42{,}9$$

$$55{,}4\ \sqrt{(1 - 0{,}7)} = 30{,}3$$

TABLA 9

Puntuaciones de cambio individuales a partir de los deciles

Suj.	Pretest.	Postest.	Differ.	PCP.	Recup.
2	10	10	0	0	N
3	10	4	6	3,75	R
4	9	6	3	1,87	RL
5	10	8	2	1,25	RL
6	10	1	9	5,62	R
7	7	2	5	3,12	R
8	7	4	3	1,87	RL
9	10	7	3	1,87	RL
10	10	7	3	1,87	RL
11	5	3	2	1,25	RL
12	10	10	0	0	N
13	4	1	3	1,87	RL
15	10	1	9	5,62	R
16	10	10	0	0	N
17	10	10	0	0	N
18	10	6	4	2,50	R
19	10	9	1	0,62	M
20	10	3	7	4,37	R
21	8	2	6	3,75	R
22	10	10	0	0	N
23	9	8	1	0,62	M
24	10	10	0	0	N
25	9	1	8	5	R
26	8	1	7	4,37	R
27	9	4	5	3,12	R
28	10	2	8	5	R
29	10	7	3	1,87	RL
30	10	4	6	3,75	R
31	10	8	2	1,25	R
Media Total	9,1	5,48	3,6	2,28	
Desviación Típica	1,6	3,3	2,9	1,8	

En la columna Recuperación los significados son los mismos de la tabla 7. Como hay un efecto techo en las puntuaciones, aquí no se calculan empeoramientos y la N significa «no mejorado».

los sujetos tratados pasan al lado más positivo de la distribución de la escala. Resultados similares se han encontrado en la terapia de pareja: menos de la mitad de las parejas tratadas terminan el tratamiento en la mitad «feliz» de parejas después del tratamiento, incluso aunque el tratamiento haya tenido un efecto significativo (Jacobson y Truax, 1991).

Conclusiones

Los problemas suscitados en la investigación clínica para llevar a cabo la evaluación de la eficacia diferencial de diferentes modalidades terapéuticas y para comparar los resultados de los diferentes estudios son múltiples. Los trabajos controlados se realizan fundamentalmente con sujetos análogos, no se especifican suficientemente los procedimientos utilizados, los tratamientos son limitados en el tiempo y los períodos de seguimiento cortos. Además de los niveles de confianza utilizados, el tamaño de la muestra, la magnitud del efecto del tratamiento y la fiabilidad de los instrumentos de medida son muy variables en los diferentes estudios y dificultan en muchos casos llegar a conclusiones determinantes. No son menores las dificultades planteadas para conseguir, en diseños experimentales de intervención clínica, muestras grandes y homogéneas y grupos de control sin tratamiento (o de lista de espera) que acepten períodos de seguimiento amplios sin experimentar una mortalidad experimental grande.

El metaanálisis posibilita la reevaluación *post hoc* de las investigaciones sobre valoración de los resultados terapéuticos y, de este modo, proporciona una medida objetiva y cuantitativa para evaluar la bibliografía sobre resultados. Los datos expuestos en este trabajo ilustran sobre la potencialidad del metaanálisis en la evaluación de la eficacia diferencial de tratamientos. El problema no está, sin embargo, del todo resuelto porque los resultados de esta técnica varían en función de los criterios empleados para obtener los datos y seleccionar los trabajos, así como del sesgo de los observadores (Erwin, 1984; Haaga, 1986; Searles, 1985; Shapiro y Shapiro, 1982; Smith, Glass y Miller, 1980). Se mezclan con frecuencia estudios metodológicamente adecuados con otros deficientes, los criterios utilizados para seleccionar las muestras no son siempre adecuados, no se describen con precisión las técnicas terapéuticas utilizadas, las medidas de evaluación de resultados o no son explícitas o son excesivamente indeterminadas y, por último, hay una amalgama de variables significativas, tales como el tipo de problema, el grado de deterioro, la duración del tratamiento, etc. (Ávila, 1990; Bayés, 1990; Frank, 1979; Labrador, 1986; Pelechano, 1989).

Ante estas limitaciones expuestas, el alcance de este trabajo deriva de la presentación de un diseño cuasiexperimental —el diseño de línea base nocausal construida— para evaluar la eficacia diferencial de los tratamientos, que, por un lado, permite presentar una alternativa al grupo de control y a las muestras grandes y homogéneas en las investigaciones clínicas y, por otro, facilita los procedimientos de acumulación evaluativa mediante el metaanálisis.

Al margen de las limitaciones ya expuestas, con el recurso a una muestra normativa en el diseño propuesto se consiguen los siguientes objetivos: determinar la fiabilidad y validez de los instrumentos de medida, precisar la eficacia diagnóstica (sensibilidad y especificidad) de los mismos, establecer un baremo de estimación de la mejoría terapéutica y, por último, contrastar la eficacia de la terapia sin una sobreestimación de los resultados mediante el cambio de percentiles y la comparación de las medias prepostratamiento con las de la muestra normativa. La información que se puede extraer de este diseño también permite estimar la fiabilidad y la significación clínica de cada puntuación individual de cambio.

La aplicación en nuestro estudio de esta metodología al ámbito de la agorafobia tiene sólo el objetivo de mostrar la potencialidad de este diseño en la psi-

cología clínica. Esta propuesta metodológica, aplicada en este caso a un estudio concreto, requiere de estudios ulteriores de contrastación y de la aplicación de la misma a diferentes trastornos psicopatológicos para dilucidar su alcance en las investigaciones clínicas futuras. Por último, es evidente que nuestra demostración tiene una cierta circularidad, ya que se han utilizado las mismas muestras para aplicar todos los procedimientos. En estudios futuros, una triangulación de muestras, instrumentos y tratamientos es indispensable. De ahí que la combinación del metaanálisis con el diseño de línea base no-causal construida, tal como se sugiere en este estudio, sea una propuesta metodológica coherente.

Agradecimientos

Agradecemos a P. Apodaka, del ICE de la Universidad del País Vasco, y a M. Clemente, de la Universidad Complutense de Madrid, sus comentarios a una primera versión de este texto.

ANEXO

El objetivo de este anexo es simplemente presentar de una forma esquemática algunos de los datos de la investigación comentada en el texto que se exponen con detalle en otro lugar (Echeburúa y cols., 1991a) y que pueden facilitar al lector la comprensión del marco teórico y clínico de la misma.

Tema de investigación

La autoexposición y las benzodiacepinas en el tratamiento de la agorafobia sin historia de trastorno de pánico: Resultados a largo plazo.

Número de pacientes

Treinta y un pacientes seleccionados según los criterios diagnósticos del DSM-III-R por medio de una entrevista clínica estructurada (SCID) (Spitzer y Williams, 1987). Los pacientes fueron 24 mujeres y 7 varones, con una edad media de 36,5 años y con una antigüedad media del problema de ocho años.

Muestra normativa

Trescientos cincuenta y siete sujetos «normales» elegidos aleatoriamente, pero apareados con los pacientes en cuanto a las variables de edad y sexo.

Diseño experimental

Diseño multigrupo, con medidas múltiples y repetidas de evaluación. Asignación aleatoria a los grupos experimentales. Evaluación y tratamiento de «doble-ciego».

Modalidades terapéuticas

Autoexposición (N = 8).
Autoexposición + Alprazolam (N = 9).
Alprazolam (N = 7).
Autoexposición + Placebo (N = 7).

Evaluación

Instrumentos de evaluación principales:

Inventario de Agorafobia (Echeburúa y cols., 1991b).
Cuestionario de Miedos (Marks y Mathews, 1979).

Momentos de evaluación:

Evaluación pretratamiento.
Evaluación postratamiento.
Evaluación en los seguimientos de 1, 3, 6 y 12 meses.

Tratamiento

Tratamiento psicológico:

Método de la práctica programada (Mathews, Gelder y Johnston, 1981), con una duración de 7 sesiones individuales semanales.

Tratamiento farmacológico:

Alprazolam, con una dosificación de tres dosis diarias de 0,5 mg a horas fijas y con una duración de siete semanas.

Referencias

Arnau, J. (1981). *Psicología experimental.* México: Trillas.
Ávila, A. (1990). La investigación del proceso, alternativa a la integración de los enfoques teóricos y técnicos de la psicoterapia. *Clínica y Salud, 1*, 13-19.
Barker, S. L., Funk, S. C. y Houston, B. K. (1988). Psychological treatment versus nonspecific factors: A meta-analysis of conditions that engender comparable expectations dor improvement. *Clinical Psychology Review, 8*, 579-594.
Bayes, R. (1990). ¿Es posible evaluar la eficacia de las psicoterapias? *Clínica y Salud, 1*, 21-26.
Beck, J. G., Andrasik, F. y Arena, J. G. (1989). Diseño de comparación de grupos. En A. S. Bellack y M. Hersen (Eds.), *Métodos de investigación en psicología clínica.* Bilbao: Desclée de Brouwer.
Berman, J. S., Miller, R. C. y Massman (1985). Cognitive therapy versus systematic desensitization: Is one treatment superior? *Psychological Bulletin, 3*, 451-461.
Brown, J. (1987). A review of meta-analyses conducted on psychotherapy outcome research. *Clinical Psychology Review, 7*, 1-23.

Campbell, D. (1988). *Methodology and Epistemology for Social Science.* Chicago: University of Chicago Press.
Castillo, J. A. y Poch, J. (1991). *La efectividad de la psicoterapia: método y resultados de la investigación.* Barcelona: Hogar del Libro.
Clemente, M. (1989). Metodología de investigación de los problemas psicosociales: la investigación sobre la evaluación de las intervenciones psicosociales. *Revista de Psicología Social, 4*, 85-109.
Clum, G. A. y Bowers, T. G. (1990). Behavior therapy better than placebo treatments: Fact or artifact? *Psychological Bulletin, 107*, 110-113.
Chassan, J. B. (1979). *Research Design in Clinical Psychology and Psychiatry.* New York: Irvington Pub.
Echeburúa, E., Corral, P., García, E. y Borda, M. (1991). La autoexposición y las benzodiazepinas en el tratamiento de la agorafobia sin historia de trastorno de pánico: Resultados a largo plazo. *Análisis y Modificación de Conducta, 17*, 969-991.
Echeburúa, E., Corral, P., García, E., Borda, M. y Páez, D. (1992). Inventario de Agorafobia. *Análisis y Modificación de Conducta, 18*, 101-123.
Erwin, E. (1978). *Behaviour Therapy.* Cambridge. University Press (traducción, Pirámide, 1985).
Frank, J. D. (1979). The present status of outcomes studies. *Journal of Consulting and Clinical Psychology, 47*, 310-316.
Gerin, P. (1984). *L'evaluation des psychotherapies.* París: Presses Universitaires Françaises.
Gerin, P., Dazord, A. y Sali, A. (1991). *Psychotherapies et changements.* París: Presses Universitaires Françaises.
Haaga, D. A. (1986). A Review of the Common Principles Approach to Integration of Psychotherapies. *Cognitive Therapy and Research, 10*, 527-538.
Hersen, M. y Michelson, L. (1986). *Issues in Psychotherapy Research.* New York: Plenum Press.
Jacobson, N. S. y Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*, 12-19.
Jacobson, N. S., Wilson, L. y Tupper, C. (1988). The clinical significance of treatment gains resulting from exposure-based interventions for agoraphobia: A reanalysis of outcome data. *Behavior Therapy, 19*, 539-554.
Kazdin, A. E. (1985). The role of meta-analysis in the evaluation of psychotherapy. Special Issue: Meta-analysis and clinical psychology. *Clinical Psychology Review, 5*, 49-61.
Labrador, F. J. (1986). Controversia sobre una posible convergencia entre los distintos acercamientos terapéuticos o de intervención. *Revista Española de Terapia del Comportamiento, 4*, 259-302.
Marks, I. M. y Mathews, A. M. (1979). Brief standard self-rating for phobic patients. *Behaviour Research and Therapy, 17*, 263-267.
Mathews, A. M., Gelder, M. G. y Johnston, D. W. (1981). *Agoraphobia. Nature and Treatment.* New York: Guilford Press (traducción, Fontanella, 1986).
Matt, G. E. (1989). Decision rules for selecting effect size in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin, 105*, 106-115.
Mirowsky, J. y Ross, C. E. (1989). Psychiatric Diagnosis as Reified Measurement. *Journal of Health and Social Behaviour, 30*, 11-25.
Parloff, M. B., London, P. y Wolfe, B. (1986). Individual Psychotherapy and Behavior Change. *Annual Review of Psychology, 37*, 321-349.
Pelechano, V. (1989). Ejes de referencia y una propuesta temática. En E. Ibáñez y V. Pelechano (Eds.), *Personalidad.* Madrid: Alhambra.
Perachio, L. y Cook, T. (1988). Avances en el diseño cuasi experimental. En I. Dendaluce (Ed.), *Aspectos metodológicos de la investigación educativa.* Madrid: Narcea (II Congreso Mundial Vasco).
Rosenthal, R. (1983). Assesing the Statistical and Social Importance of the Effects of Psychotherapy. *Journal of Consulting and Clinical Psychology, 51*, 4-13.
Rosenthal, R. y Rosnow, R. (1984). *Essentials of Behavioral Research.* New York: McGraw Hill.
Rothman, K. J. (1987). *Epidemiología moderna.* Madrid: Díaz de Santos.
Sartorius, N. (1983). *Les Troubles Dépressifs dans differents contextes culturels.* Ginebra: OMS.
Searles, J. S. (1985). A methodological and empirical critique of psychotherapy outcome meta-analysis. *Behaviour Research and Therapy, 23*, 453-463.
Shapiro, D. A. y Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin, 92*, 581-604.
Silva, F. (1989). *Evaluación conductual y criterios psicométricos.* Madrid: Pirámide.
Smith, M. L. y Glass, G. V. (1977). Metaanalysis of psychotherapy outcome studies. *American Psychologist, 32*, 752-760.
Smith, M. L., Glass, G. V. y Miller, T. I. (1980). *The Benefits of Psychotherapy.* Baltimore, Maryland: John Hopkins University Press.
Spitzer, R. L. y Williams, J. B. (1987). *Structured Clinical Interview for DSM-III-R.* New York State Psychiatric Institute (Biometrics Research Department).
Willet, J. B., Ayoub, C. C. y Robinson, D. (1991). Using growth modeling to examine systematic differences in growth. *Journal of Consulting and Clinical Psychology, 59*, 38-47.
Wilson, G. T. (1985). Limitations of meta-analysis in the evaluation of the effects of psychological therapy. Special Issue: Meta-analysis and clinical psychology. *Clinical Psychology Review, 5*, 35-47.