

REVISIONES CUANTITATIVAS DE LA LITERATURA: EL META-ANÁLISIS

Carmelo VÁZQUEZ*

Universidad Complutense de Madrid

El propósito de este trabajo es presentar las características principales del *meta-análisis* (MA). No me detendré en consideraciones *técnicas* específicas de este procedimiento (por ej.: empleo de técnicas multivariadas, selección de estadísticos no sesgados, etc.) sino en aspectos estratégicos de tipo más general y en la discusión de si el meta-análisis pudiera ser un instrumento eficaz para los científicos sociales. Aquellos lectores interesados en los principios metodológicos del MA disponen de varios textos clave (fundamentalmente los de Glass, McGaw & Smith, 1981; Hunter, Schmidt & Jackson, 1982; Light & Pillemer, 1984; Cooper, 1984; Hedges & Olkin, 1985; y, en castellano, Gómez, 1987) que exponen de modo más detallado los procedimientos técnicos para llevar a cabo un estudio de dicha naturaleza. El objetivo de esta presentación, no es por tanto ofrecer una descripción minuciosa y técnica del MA sino plantear sus principales características y discutir el papel que puede tener en la construcción de una ciencia.

El MA ha sido objeto de ácidas controversias y aún hoy algunos autores cuestionan la validez de este procedimiento (véase Wortman, 1983, p. 241 y ss.; Searles, 1985). Así, un polemista tan acreditado como Eysenck ha llegado a calificar como "mega-estupidez" (*megasilliness*) a las técnicas meta-analíticas (Eysenck, 1978). En este trabajo, lejos de cualquier intención apologética, pretendo exponer cuáles son las aportaciones y limitaciones del MA y su papel en el quehacer *científico*. Un científico social, no puede ignorar el avance que el MA puede significar a este respecto, especialmente con la proliferación cada vez mayor de estudios meta-analíticos en las revistas especializadas.

* Carmelo Vázquez, Departamento de Personalidad, Evaluación y Psicología Clínica. Facultad de Psicología. Universidad Complutense. 28023 Madrid. España.

AUNQUE el MA tiene antecedentes históricos que se remontan a los años 30 y 40 con algunos estudios de Thorndike, Ghiselli, y otros (véanse Hunter et al., 1982 y Bangert-Drowns, 1986), el desarrollo de esta técnica es muy reciente. En efecto, "meta-análisis" es el nombre que Gene Glass otorgó a un procedimiento peculiar de *revisar* la literatura empírica en un área determinada (Glass, 1977). Se trata, y de ahí el nombre, de un "análisis de análisis" (Smith, Glass & Miller, 1980, p. 80) o, más sencillamente, de una *revisión cuantitativa* de los resultados hallados en un campo concreto de investigación.

Existen dos elementos caracterizadores del meta-análisis. En primer lugar, es un *procedimiento cuantitativo* para evaluar tales resultados (Cooper & Arkin, 1981); y en segundo lugar, lo que se pretende es proporcionar un *índice global* de la consistencia de esos hallazgos a través de las investigaciones consideradas en el estudio (Leviton & Cook, 1980). Este índice, como explicaré más adelante, proviene de *agregar* los índices parciales de cada uno de los estudios introducidos en la revisión.

EL META-ANÁLISIS Y OTRAS TÉCNICAS DE REVISIÓN TRADICIONALES

La pregunta más obvia que puede hacerse es: ¿en qué se diferencia el MA de las revisiones tradicionales (también llamadas "narrativas", o "cualitativas") de la literatura? Todos tenemos una profunda familiaridad con las típicas revisiones que un autor efectúa sobre un tema determinado, en las que se analizan los resultados habidos a lo largo de los años en un tema concreto. Este tipo de estudios es muy común y, de hecho, hay publicaciones de enorme prestigio (como el *Psychological Bulletin*, el *Annual Review of Psychology*, o el *New England Journal of Medicine*) que se nutren en buena medida de tales trabajos de revisión. Como se puede adivinar, el debate entre ambos tipos de revisión, se enmarca en el debate más amplio entre procedimientos cuantitativos y cualitativos en la metodología de la investigación. Un debate que, por otra parte, se presenta habitualmente en falsos términos dicotómicos o mixtificadores.¹ Aunque algo más adelante expondremos con

¹ Donald Campbell, uno de los metodólogos más importantes en Psicología, afirmaba a este respecto: "A fin de cuentas, el hombre es, en su quehacer cotidiano, un conocedor competente, y al conocimiento cualitativo del sentido común no le reemplaza el conocimiento cuantitativo. Más bien, el conocimiento cuantitativo tiene que confiar y erigirse sobre lo cualitativo, incluyendo la percepción ordinaria. Nosotros, los metodólogos, debemos alcanzar una epistemo-

más detalle algunos de los principales problemas de las revisiones tradicionales, en la Tabla 1 adelantamos resumidamente algunas de las diferencias básicas entre éstas y el meta-análisis.

Como han señalado Cook & Leviton (1980), el papel de una buena revisión consiste en establecer "*hechos*". Es decir, resultados que, con un margen razonable de error, sean incontestables y replicables. Sin embargo, las revisiones tradicionales como todos sabemos, y Hunter et al. (1982) han ridiculizado abiertamente en su espléndido trabajo, suelen acabar con las consabidas frases del tipo: "...los resultados parecen indicar... pero se necesitan más investigaciones en el futuro para clarificar la *confusión* existente en este área". Como veremos más adelante, muchos de estos lugares comunes se deben a que los procedimientos tradicionales de revisión normalmente son bastante *inefectivos* para obtener conclusiones generales, o bien las prácticas metodológicas habituales de los revisadores son manifiestamente pobres (Cook & Leviton, 1980; Leviton & Cook, 1980).

TABLA 1

Comparación entre las características de las revisiones tradicionales de la literatura y el meta-análisis

<i>Revisiones narrativas</i>	<i>Meta-análisis</i>
1. Criterios subjetivos en la selección de estudios.	1. Criterios explícitos y verificables de selección.
2. Interpretación subjetiva	2. Interpretación de un estadístico.
3. Número limitado de estudios.	3. Número ilimitado de estudios.
4. Vulnerable a sobrecarga cognitiva.	4. Apenas vulnerable a sobrecarga cognitiva.

¿Cuáles son los *problemas* de las tradicionales revisiones narrativas (que constituyen, por cierto, la totalidad de las revisiones que aparecen publicadas en nuestro país)? Normalmente, los teóricos del meta-análisis distinguen al menos tres tipos diferentes de dificultades:

1. *Sesgos en la selección de trabajos*

El revisor normalmente no es exhaustivo en el área que desea acotar y selecciona los trabajos a incluir en función del acceso que tenga a revistas, de sus preferencias teóricas, o de su familiaridad con aspectos parciales del tema que desea estudiar. Además, es frecuente que en la revisión se excluyan estudios que, *a juicio* de quien efectúa la revisión, no satisfacen unos criterios metodológicos adecuados.

logía aplicada que integre ambos" (p. 191). Una visión comprehensiva semejante también es expuesta por Cook & Reichardt (1982).

Algo que tampoco se puede desdeñar es que es muy probable que exista un sesgo de resultados positivos en las publicaciones a las que normalmente tenemos acceso. Es decir, los autores de trabajos originales de investigación, siguiendo la nefasta práctica científica de plantear experimentos para rechazar la hipótesis nula, son reacios a enviar para su publicación resultados que no han alcanzado significación y, a su vez, los editores de revistas también rechazar la hipótesis nula que trabajos en los que los resultados permiten rechazar la hipótesis nula que trabajos en los que ésta no resulta rechazada. Una caricatura extrema de esta posibilidad es la hipótesis de que los estudios publicados podrían representar simplemente el 5 % de falsos positivos en una población de estudios en la que realmente la hipótesis nula sería cierta (Smith, 1980; Rosenthal, 1979).

Otro potencial sesgo de las revisiones cualitativas es que normalmente se incluyen sólo estudios publicados, mientras que no se consideran trabajos no publicados (tesis, tesinas, informes de laboratorio, etc.). Esto es importante pues parece demostrado que los resultados de los trabajos publicados suelen ser mejores que los de los trabajos no publicados (Glass, Smith, & Barton, 1979; Smith, 1980; White, 1982; Straw, 1983; Green & Hall, 1984). Incluso, como en el meta-análisis de Smith (1980) sobre sesgos sexuales en psicoterapia y consejo psicológico, la diferencia entre estudios publicados y no publicados reside no sólo en la magnitud de los efectos encontrados sino incluso en la dirección del efecto!

En todo caso, la selección de los trabajos a estudiar juega un papel central en todo MA. Dadas las pretensiones de objetividad e imparcialidad de este procedimiento, el objetivo de un MA ideal sería efectuar una inclusión *exhaustiva* de los estudios, publicados o no, en el área estudiada; para esta búsqueda de estudios originales, se puede recurrir al Psychological Abstracts, revisiones computarizadas, escribir a los principales autores del área, etc. (véase Cooper, 1982, 1984; Light & Pillemer, 1984).

Smith (1980) incluso llega a afirmar que "ningún meta-análisis debería considerarse completo si se omitiese algún subconjunto de la población [de estudios analizados]" (p. 22). Aunque normalmente esto no es posible, los estudios meta-analíticos se distinguen por la gran cantidad de trabajos que suelen introducir (publicados o no publicados, como tesis, tesinas, presentaciones a congresos, etc.) y por la elaboración de determinados índices para *atenuar* los efectos de no incluir el universo de estudios en ese área (ej.: Rosenthal, 1980b).² Lo que sí es importante señalar es que en cualquier

² Este es el denominado "problema del archivador" (*file drawer problem*), es decir, la consideración de que en los escritorios de los científicos se almacena un número indeterminado

estudio meta-analítico se señalan *explícitamente* las fuentes de información utilizadas y los *criterios* de inclusión/exclusión de estudios. No obstante, es necesario decir que existe una polémica muy activa, de gran alcance metodológico, respecto a los criterios de selección de estudios en un MA. Mientras que autores como Glass sugieren una estrategia “omnívora”, otros autores señalan que, con el fin de extraer resultados más *útiles*, deberían introducirse estrictos filtros de calidad en la selección de artículos incluíbles en un MA (ej.: Kazdin, 1985; Kendall & Maruyama, 1985).

2. *Criterios subjetivos de interpretación*

El método tradicional para extraer conclusiones en las revisiones tradicionales es el llamado “recuento de votos” (*box count* o *vote counting*) (Hedges & Olkin, 1980); es decir, el autor, después de seleccionar los estudios y describirlos más o menos detalladamente, se limita a contar cuántos son significativos y cuántos no. La efectividad del tratamiento (ej.: tipos diferentes de terapias psicológicas) o, más sencillamente, el grado de relación entre cualesquiera variables (ej.: sexo e inteligencia), se infiere sin más de la diferencia derivada de este simple conteo. Sin dudar las innegables virtudes democráticas de tal sistema, hemos de reconocer sus evidentes limitaciones como táctica científica.

Por otra parte, con este rudimentario procedimiento de “recuento”, quien efectúa la revisión se suele olvidar de cuántos resultados significativos se esperan bajo la hipótesis nula. En efecto, “si ‘sólo’ el 30 % de todos los estudios muestran resultados significativos, un revisor tradicional pudiera ser escéptico de que se apoye una determinada hipótesis. De hecho, ese 30 % supera con exceso el 5 % que se esperaría bajo la hipótesis nula...” (Green & Hall, 1984, p. 41). Si a esta tendencia *infraestimadora* de la magnitud de los efectos que se observa en las revisiones tradicionales, se le añade el hecho de que en estas revisiones se suelen “contar” como no significativos aquellos resultados que se aproximan, sin llegar, al nivel de significación habitual del 5 %, el empleo de este método narrativo conduce inevitablemente a conclusiones “conservadoras” sobre la eficacia de un determinado tratamiento.³ Además, con este procedimiento normalmente no se informa sobre la

de estudios no publicados que desgraciadamente no han visto la luz por no rechazarse la hipótesis nula (Rosenthal, 1980a; Smith, 1980).

³ Cooper & Rosenthal (1980) incluso comprobaron empíricamente que, partiendo del mismo conjunto de estudios, se obtienen conclusiones más conservadoras con revisiones tradicionales que con revisiones cuantitativas.

dirección de los resultados cuando éstos *no* alcanzan significación (Jackson, 1980; Walberg & Haertel, 1980).

3. Ineficacia

Por último, las revisiones narrativas devienen ineficaces especialmente cuando el área analizada incluye varias decenas de estudios. Dadas las limitaciones que los seres humanos tienen en el procesamiento de la información (Vázquez, 1985), parece cuestionable que un autor, aun admitiendo su buena voluntad, pueda considerar *simultáneamente* las características metodológicas, resultados, variantes técnicas, etc., de 90 ó 100 estudios. Esto supone una "sobrecarga cognitiva" (cfr. Glass, 1977) o un "ruido" (cf. Kazdin, 1985) difícilmente superable para cualquier autor, lo que suele convertir estas revisiones en un auténtico festival de parcialidades. Por otro lado, resulta difícil en este tipo de revisiones tradicionales, efectuar interpretaciones a partir de estudios bastante heterogéneos entre sí en los que, además, normalmente se hallan interacciones complejas en los resultados.

Conviene señalar que estas limitaciones, llamémoslas "cognitivas", no implican necesariamente que los autores sean más "liberales" o más propensos a aceptar la significatividad de unos resultados determinados. De hecho, parece demostrado que el empleo de criterios estadísticos en las revisiones cuantitativas resulta más sensible para descubrir efectos significativos que el mero "juicio cualitativo" de los autores (Cooper & Rosenthal, 1980).

De todos estos elementos se deriva el hecho de que en las revisiones tradicionales existe un elevado grado de subjetividad, que está alimentada por la *ausencia de reglas formales* para cada uno de los pasos que intervienen en la revisión. En consecuencia, no es infrecuente que cuando varios autores hacen revisiones diferentes sobre una misma área, e incluso analizando el mismo conjunto de estudios, lleguen a conclusiones diferentes; un ejemplo de este hecho es la polémica desatada entre Munsinger (1978) y Kamin (1978) sobre los efectos del ambiente escolar y familiar sobre el cociente intelectual infantil. Del análisis del *mismo* conjunto de estudios, Munsinger concluye que la herencia es mucho más importante que el entorno mientras que Kamin concluye exactamente lo contrario.

Ahora bien, como señalan Cook & Levinton (1980), estas dificultades no son necesariamente intrínsecas a las revisiones de tipo cualitativo; de hecho existen magníficas revisiones que son auténticamente iluminadoras sobre un área determinada de investigación (ej.: Maccoby & Jacklin, 1974; Nuechterlein, 1977; Blaney, 1986). Lo que sí es cierto es que los métodos cualitativos *favorecen* la aparición de tales dificultades, especialmente cuando el número de estudios incluidos en la revisión supera la cifra de 30 ó 40.

a) *Índices para cada estudio*

El problema de buscar un índice para evaluar la magnitud de un hallazgo empírico es de una enorme importancia. Imaginemos que una Terapia A reduce 15 puntos más que una Terapia B el grado de depresión de los sujetos en una escala de 0 a 100. O imaginemos que un grupo de niños sometidos a un programa de estimulación precoz llegan a un Nivel 7 tras unos meses de programa, en comparación con los niños de un grupo control que sólo llegan a un Nivel 5 al cabo de ese tiempo. ¿Cuál es el significado de todo esto? ¿Cómo se determina la magnitud del efecto logrado tras todas estas intervenciones?

La Psicología está cada vez más inundada de este tipo de datos y apenas se nos proporcionan herramientas para resolver estas cuestiones decisivas (Yeaton & Sechrest, 1981). A veces se hacen torpes intentos de dar cuenta de la importancia del efecto hallado. Uno de los casos más frecuentes es apelar a la probabilidad asociada al efecto obtenido; esta práctica, muy frecuente en los científicos sociales, es completamente errónea. En efecto, la probabilidad con que se acepta o rechaza una hipótesis no proporciona información alguna sobre el impacto o la magnitud del efecto hallado (Cowles & Davis, 1982; Huberty, 1987). Así, es muy habitual que en las publicaciones se mencionen frases como "los resultados alcanzaron un nivel de significación *muy elevado*", lo cual no tiene ningún significado estadístico y demuestra, por cierto, una comprensión más bien escasa de lo que quiere decir la significación estadística (Murray & Dosser, 1987; Harcum, 1989, 1990; Rosnow & Rosenthal, 1989).

En los resultados de las investigaciones empíricas, normalmente se proporciona un estadístico determinado (Pearson, F , t , etc.) junto a su correspondiente probabilidad asociada. Ésta es una práctica que se ha convertido en nuestro "modus operandi" como científicos. Sin embargo, este hábito tiene bastantes limitaciones. En efecto, las probabilidades son elementos bastante *insensibles* al tamaño de la muestra, al tipo de diseño experimental, y, en definitiva, apenas proporcionan información sobre la *fortaleza* del resultado hallado.

Por otro lado, un simple análisis de las probabilidades, es decir, comprobar si pruebas como la "t" de Student o la de "F" de Fisher alcanzan la significación, nos hace partícipes de uno de los mitos más dañinos en la investigación psicológica: la idolatría por los niveles de significación o lo que podríamos llamar "alfaismo". Ya desde los años 50, Paul Meehl, con su habitual estilo cáustico ha venido ridiculizando esta práctica "pseudocientífica" (Meehl, 1978) que, además, es posible que esté *obstaculizando* la

construcción de teorías fuertes en Psicología (Carver, 1978; Hunter et al., 1982; Dar, 1987; Huberty, 1987; Rosnow & Rosenthal, 1989).

Por el contrario, el índice más utilizado en las investigaciones meta-analíticas es el llamado "tamaño del efecto" (TE) –también conocido por "magnitud del efecto"–. Éste es un sencillo índice que nos proporciona información sobre la *magnitud de los resultados* de una investigación determinada. Propuesto originalmente por Cohen (1977) y utilizado por Smith & Glass (1977) en su trabajo pionero del meta-análisis sobre los efectos de las psicoterapias, el TE proporciona una medida estándar y por lo tanto, hace *directamente comparables* los resultados de estudios diferentes. En definitiva, reduce los resultados de cada estudio a una escala absoluta.

$$\begin{array}{ccc}
 \boxed{d = \frac{\bar{X}_t - \bar{X}_c}{S_c}} & \begin{array}{l} \bar{X}_t = 50 \\ \bar{X}_c = 30 \\ S_c = 20 \end{array} & \boxed{d = 1.00} \rightarrow \boxed{\bar{d} = 0.75}
 \end{array}$$

FIGURA 1

Cálculo del índice "d" en un procedimiento meta-analítico

El tipo de TE depende del tipo de datos con que nos encontremos al analizar los estudios originales. Uno de los TE más utilizados es el denominado "d" (delta) de Cohen (1969), basado en las medias de los grupos que deseamos comparar; dado que los valores de las medias y las desviaciones típicas figuran de modo rutinario en un alto porcentaje de investigaciones publicadas, el índice "d" es muy usado. Como puede apreciarse en la sencilla fórmula de la Figura 1, "d" no es más que una *puntuación típica*. En efecto, consiste en la puntuación típica resultante de comparar la media del grupo experimental respecto a la media del grupo control (Glass et al., 1981). Obviamente, al expresar dicha magnitud en unidades de desviación típica, se hace posible comparar directamente los efectos provenientes de estudios diferentes.

En la Figura 1 exponemos un sencillo ejemplo. Supongamos que el grupo experimental obtuvo una puntuación media de 50 tras un procedimiento terapéutico, y el grupo control obtuvo una puntuación de 30. Para hallar el índice "d" tomamos la desviación típica del grupo control ($S_c = 20$),⁴ por lo

⁴ Existe una discusión no resuelta a este respecto. Algunos prefieren el uso de S_c o S_t (Glass, 1980), mientras que otros creen más adecuado emplear la desviación típica promedio de ambos

que "d" tiene un valor de 1.00. Imaginemos que, tras la revisión de 200 estudios publicados, el efecto medio obtenido es de 0.75, por lo que podemos afirmar que la puntuación media del grupo experimental está tres cuartas partes por encima de la desviación típica del grupo control. Este valor puede traducirse a porcentajes. En efecto, dado que "d" es una puntuación típica, y asumiendo la normalidad de la distribución en la población,⁵ se puede transformar dicha puntuación en un porcentaje acudiendo a las tablas de probabilidad para las puntuaciones Z para determinar cuál es el percentil asociada al TE obtenido. Si es positivo, esto indica que el tratamiento es beneficioso, y si es negativo indica que el tratamiento es inefectivo o incluso dañino; descubrimos que a 0.75 le corresponde a un percentil del 77.3 %, lo que indica que el sujeto medio del grupo experimental tiene una mejoría mayor que el 77.3 % de los sujetos del grupo control.

De un mismo estudio se pueden obtener *varios* tamaños del efecto. Imaginemos que se han tomado 2 medidas, una a los seis meses de iniciar el tratamiento y otra a los 12 meses; en este caso, en el mismo podremos disponer de dos TE diferentes procedentes de la misma investigación. Más adelante discutiremos algunos de los problemas estadísticos que esta posibilidad plantea.

Según las pautas de interpretación propuestas por Cohen (1977) un efecto 0.2 debe considerarse pequeño, mientras que el 0.5. se corresponde a un efecto moderado y a partir del 0.8 podemos considerar que el efecto es grande. Pero naturalmente, la atribución de esa significación dependerá del tipo de estudio, el impacto social que pueda tener, etc. (Glass et al., 1981, pp. 99-106; Sechrest & Yeaton, 1982; Wortman, 1983; Huberty, 1987).

Además del índice "d", existen otros que permiten hallar un TE cuando los datos originales son correlaciones, o proporcionan sólo los valores de estadísticos como *F* o *t* (Glass et al., 1981; Rosenthal, 1978, 1980b, 1983a; Hunter et al., 1982; Gómez, 1987). Igualmente se dispone de otros índices que nos permiten efectuar estimaciones del TE cuando los datos que se ofrecen en la publicación revisada son incompletos (Hedges & Olkin, 1985; Glass et al., 1981). Asimismo, Hedges y su grupo (Hedges, 1982; Hedges & Olkin, 1985) han elaborado diversos procedimientos técnicos para obtener estimaciones no sesgadas de tamaños del efecto teniendo en cuenta para ello

grupos, es decir, S_{CH} (Hunter et al., 1982), aunque posiblemente ninguno de estos métodos es mejor *per se* (Walberg & Haertel, 1980; Glass, 1980).

⁵ Si se considera que "d" es efectivamente una estimación muestral de las diferencias de las poblaciones comparadas, es posible que dicho estadístico sea un estimador sesgado. Las innovaciones técnicas de Hedges (Hedges, 1980; Hedges & Olkin, 1985) justamente van dirigidas a proporcionar estadísticos no sesgados del TE.

el tamaño de la muestra, y Hunter et al. (1982) ofrecen diversas técnicas para corregir sesgos debidos a posibles errores en el muestreo de estudios y/o al error de las variables dependientes analizadas para hallar el TE.

Sin embargo, no debería bastar con hallar los efectos medios de un tratamiento, pues si bien nos proporcionan un índice de tendencia central, es conveniente averiguar la *distribución* de la variación de los TE provenientes de la serie de estudios analizados. Para esto se pueden emplear medidas estadísticas de variación (Rosenthal & Rubin, 1982; Hunter et al., 1982) o diversas representaciones visuales (como las sugeridas en el excelente trabajo de Light & Pillemer, 1984). La variabilidad de la magnitud del efecto puede

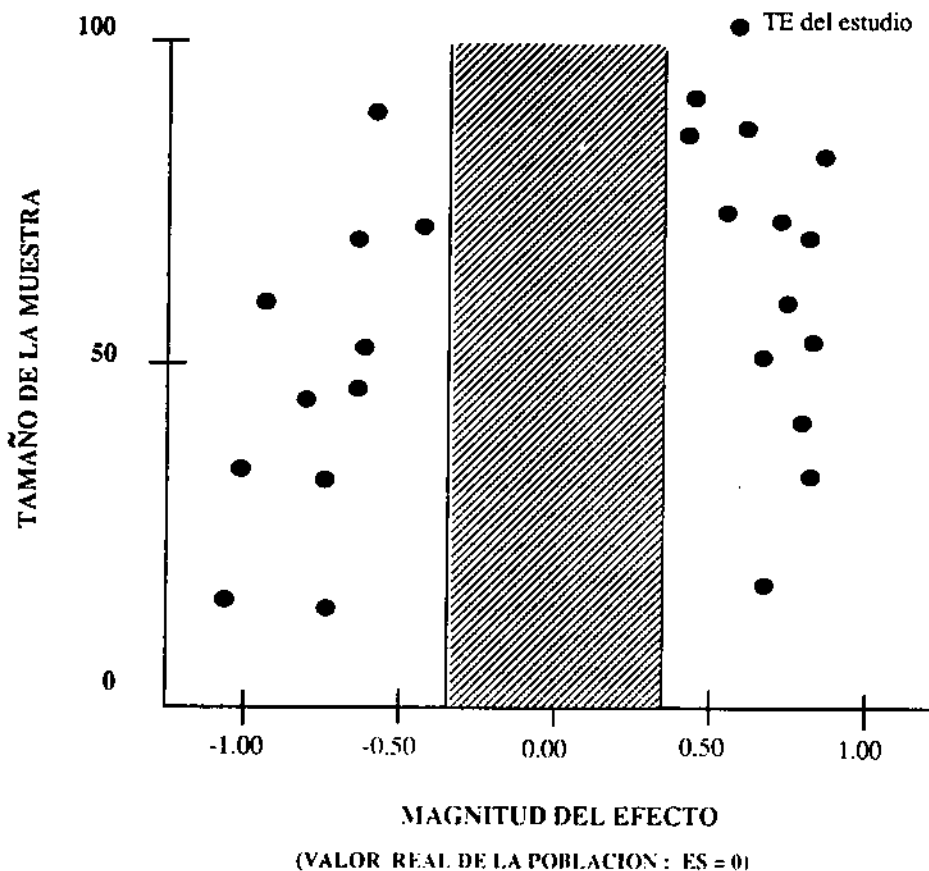


FIGURA 2

Distribución de los TE hallados en diversos estudios

proporcionar valiosa información sobre, por ejemplo, posibles *sesgos* de publicación. En efecto, imaginemos que apenas existen estudios en los que el índice TE esté próximo al 0; si suponemos que existe una distribución subyacente normal de los TE, esto puede indicar que apenas se han publicado (por las razones que sean) estudios en los que no se ha rechazado la hipótesis nula —véase un ejemplo en la Figura 2; además, en esta figura se descubre cómo los TE tienden a cero cuando los tamaños de las muestras empleadas son más grandes, lo que parece confirmar que el verdadero TE de la población es igual a cero. Con técnicas de análisis de dispersión se pueden hallar, por lo tanto, posibles sesgos en los estudios incluidos en la revisión; de modo semejante, el análisis detallado de estudios que han arrojado un TE muy alejado del TE medio de otros estudios puede proporcionar pistas valiosas sobre los factores que pudieran explicar estas diferencias (Cook & Levinton, 1980).

b) *Índices globales de agregación*

Existe cierta polémica sobre las características definitorias del MA. Para algunos lo peculiar del MA es la *cuantificación* de la información para cada estudio individual (Cooper & Arkin, 1981), mientras que para otros el elemento clave consiste en la *agregación* de resultados de diferentes investigaciones con el fin de obtener índices más globales y generalizables sobre la eficacia de un tratamiento o de un programa (Levinton & Cook, 1980). En todo caso, esta última característica es crucial en las revisiones cuantitativas, aunque dentro de la metodología del MA es más novedosa (Strube, 1985b).

¿Cómo se evalúa el impacto global de un programa o un tratamiento? Además de la mencionada técnica rudimentaria del “recuento” existen dos técnicas cuantitativas diferentes de agregación que pasamos a revisar.

Agregación de probabilidades. Consiste en *combinar* los tests de significación de diferentes pruebas y de este modo, incrementando el tamaño de la muestra, comprobar la significación del índice global (ej.: sumar las puntuaciones “Z” de cada estudio y dividir las por la raíz cuadrada del número de estudios combinados). Rosenthal, el autor que ha desarrollado esta técnica, ha descrito nueve procedimientos diferentes para agregar probabilidades provenientes de diferentes estadísticos (medias aritméticas, proporciones, coeficientes de correlación, etc.) (Rosenthal, 1978; Gómez, 1987). El problema de esta estrategia es que al tratar simplemente con la probabilidad proveniente de un estadístico, presenta serias dificultades de interpretación.⁶

⁶ Desgraciadamente, es muy habitual que se señale simplemente que se rechaza la hipótesis nula con “ $p < .05$ ” o “ $p < .01$ ” sin dar valores más exactos asociados al estadístico. Sería más

En primer lugar, la significación de una prueba, cuando el N es grande, puede cambiar con pequeños cambios en las puntuaciones (todos somos conscientes de esta limitación cuando utilizamos índices como Pearson).⁷ Utilizando este método de agregación, es muy difícil no rechazar la hipótesis nula de que el TE de la población es igual a cero (Glass et al., 1981). Además, y esto es importante reiterarlo hasta la saciedad, la significación de un estadístico —sea de un estudio aislado o sea resultante del promedio de varios estudios como en el caso que estamos discutiendo— no proporciona información alguna sobre su relevancia o significación práctica o, en otros términos, sobre la *magnitud* del efecto estudiado.

Tamaño medio del efecto. Este procedimiento, característico del meta-análisis de Glass, consiste en hallar un TE para cada estudio y a continuación, *combinar* los TE para hallar un TE global (Glass, 1980; Glass et al., 1981; Hedges & Olkin, 1985). Esta aproximación es probablemente la más utilizada y, de hecho, ha sido también incorporada por Rosenthal en sus últimos trabajos (Rosenthal & Rubin, 1986).

Si se supone que los TE de cada estudio son unidades de la población real, el TE medio evidentemente es una estimación de la media de la población. Como cualquier estadístico paramétrico, diversos factores (tamaño insuficiente de la muestra, datos originales incompletos, etc.) pueden *sesgar* dicha estimación. Hedges (Hedges 1982; Hedges & Olkin, 1985, 1986) ha derivado diversas técnicas de corrección para producir estimaciones no sesgadas del TE.

Al igual que sucedía con el TE de cada estudio, el TE medio se puede expresar el TE en unidades de proporción, lo que puede ser ventajoso. Por ejemplo, al evaluar el impacto de un programa terapéutico puede ser más interesante, o más didáctico si se prefiere, expresar la proporción de sujetos que se benefician del programa que un mero índice "TE" (Glass et al., 1981; Light & Pillemer, 1984).

El procedimiento más habitual consiste en hallar un promedio de los TE procedentes de cada estudio (véase la Figura 1). Ahora bien, hallar una simple media, sin información sobre la dispersión de los estudios, puede conducir a conclusiones infundadas. Por ejemplo, 10 estudios con un TE de 0.3 producen el mismo TE global que 9 estudios con un ES de 0 y uno con un ES de 3.0 (Light & Pillemer, 1984). Una solución para este problema puede

adecuado proporcionar el valor exacto de "p" especificando incluso milésimas (Huberty, 1987) lo que, además, favorecería las técnicas de revisión cuantitativa basadas en la acumulación de estas probabilidades (Rosenthal, 1978, 1986; Strube, 1985a, b).

⁷ Diversos autores han analizado los límites de estas prácticas estadísticas viciadas (ej.: Harcum, 1989; Dar, 1987; Hunter et al., 1982; Meehl, 1978; Huberty, 1987).

ser analizar detalladamente los estudios que se apartan mucho de la media, sin incluirlos en el cálculo del TE medio. Para este tipo de análisis minuciosos "intra-estudio", las revisiones cualitativas probablemente son más útiles y adecuadas que los procedimientos cuantitativos (Kazdin, 1985).

De cada estudio individual pueden extraerse varios tamaños del efecto, en función de las variables que se han tenido en cuenta en la codificación (ej.: tipos de grupos control, o número de evaluaciones a lo largo del tiempo si se trata de un diseño longitudinal). En este sentido, algunos autores han considerado el MA como una técnica de regresión (Kendall & Maruyama, 1985); permite averiguar qué variables (variables independientes) permiten predecir un TE basado en una medida específica (variable dependiente). Esta posibilidad plantea problemas metodológicos aún no bien resueltos. Por ejemplo, la inclusión de varios TE para cada estudio plantea el problema estadístico de la posible falta de independencia de las mismas y consecuentemente unas tasas elevadas de error Tipo I (Walberg & Haertel, 1980; Glass et al., 1981; Hunter et al., 1982; Hedges & Olkin, 1985; Strube, 1985a; Searles, 1985) o el de cuáles de los TE de cada estudio han de seleccionarse para calcular el TE global (Matt, 1989).

Respecto a la fortaleza de un TE medio dado, sirven los mismos argumentos que planteamos al exponer los TE de cada estudio individual. Es decir, aunque a veces se ha intentado crear un baremo que sirva de guía para indicar si un efecto es "fuerte" o "débil", en realidad esto es desaconsejable pues no hay bases sólidas absolutas o generales para valorar "a priori" la importancia del efecto hallado (véanse Glass et al., 1981; y, especialmente, Wortman, 1983).

LOGÍSTICA DEL META-ANÁLISIS

En este apartado señalaremos brevemente cuáles han de ser los pasos secuenciales a seguir para llevar a cabo un meta-análisis (véase la Figura 3).

1. *Selección del área de estudio.* En esta primera etapa, se ha de delimitar un área de investigación de interés en la que haya cierta confusión o contradicciones en los resultados. Es deseable que alguno de los miembros que lleven a cabo el MA sea un "experto" en el área en cuestión. En cualquier caso, si no se posee suficiente información sobre el tema elegido, debe continuarse la búsqueda de información y la lectura detallada de artículos originales pues es importante que el MA esté dirigido por hipótesis sobre la discrepancia de los resultados en las investigaciones originales (véase Light & Pillemer, 1984).

PROCESO DEL META-ANALISIS

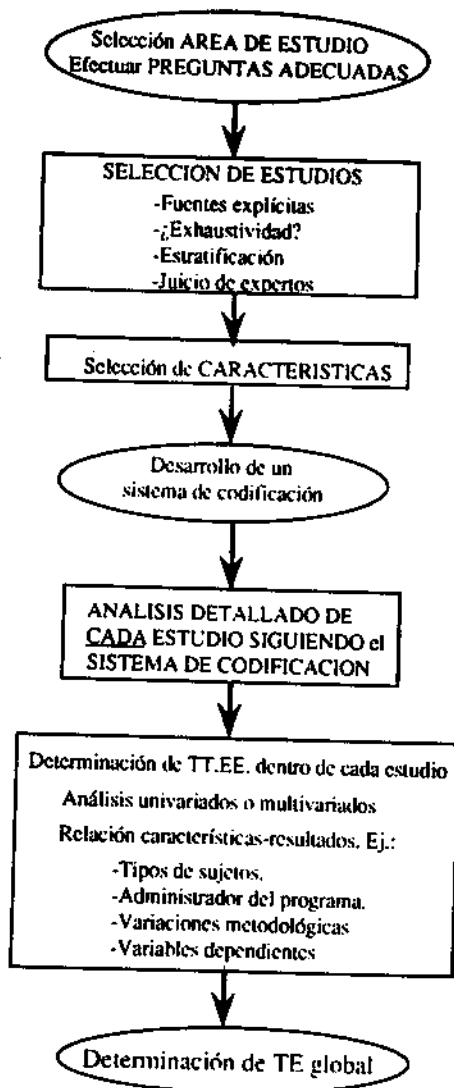


FIGURA 3

Secuencias en el desarrollo de un meta-análisis

2. *Selección de estudios originales.* Ésta es una fase importante de todo MA. Se debe hacer un rastreo lo más exhaustivo posible de los estudios publicados o no publicados sobre el tema de interés. Para ello se puede recurrir a bases de datos computarizadas (ej.: PsychInfo, PsychSCAN), Psychological Abstracts, repaso de todas las revistas del área, información de los autores más prestigiosos en el área, etc. (Hunter et al., 1982; Cooper, 1984; Light & Pillemer, 1984; Gómez, 1987). Los criterios de selección y de recogida de información han de consignarse *explícita y necesariamente* en los resultados del MA. Como señala Kazdin (1985), el MA no elimina la subjetividad pero la hace pública y manifiesta.

3. *Sistema de codificación.* Una vez leídos los principales estudios, debe efectuarse un manual de codificación para los estudios que serán incluidos en el MA. Este manual ha de ser bastante exhaustivo, para que queden consignadas las principales características de todos los estudios (ej.: tipo de sujetos, escalas utilizadas, tipo de procedimiento experimental, tipo de material utilizado, etc.). La selección de estas características dependerá del juicio que los autores del MA consideren que pudieran ser importantes para explicar la divergencia de resultados.

Es importante que este sistema sea capaz de acomodar el mayor número posible de estudios dentro del área que se quiera revisar. Si no se hace así, se corre el riesgo de incluir en el MA sólo aquellas investigaciones que se ajustan a un patrón de investigación más estándar (Green & Hall, 1984).

4. *Codificación de los estudios.* Ésta es una etapa delicada y algo tediosa. Se trata de codificar cada uno de los estudios seleccionados, siguiendo el sistema de codificación creado. Aunque habitualmente no se dan datos sobre la fiabilidad interjueces cuando hay más de un codificador, debería proporcionarse información al respecto (Kendall & Maruyama, 1985; Matt, 1989).

El sistema de codificación es el instrumento para consignar y analizar las variaciones inter-estudio.⁸ Por lo tanto, es esencial para detectar los patrones de variación en los resultados obtenidos en el MA (Hedges & Olkin, 1986). Permite identificar factores (de diseño, de tipo de sujeto empleado, etc.) asociados a los resultados. De hecho, una de las grandes ventajas de las revisiones cuantitativas sobre los procedimientos narrativos es que las primeras permiten cuantificar el papel de estos factores en los efectos hallados.

Muchos critican que en un meta-análisis se pueden incluir estudios de calidad dudosa, o incluir estudios muy antiguos, con tipos diferentes de sujetos, con tipos diferentes de diseños, etc. (Eysenck, 1978; Garfield, 1984).

⁸ Los lectores interesados pueden solicitar al autor una copia de la hoja de codificación que hemos confeccionado para la revisión de los estudios sobre estado de ánimo y memoria (Matt, Vázquez, & Campbell, 1990).

Sin embargo, esto no supone en modo alguno un problema. Por el contrario, una de las mayores virtudes del MA es que todos estos elementos se pueden codificar y cuantificar para ver si *realmente* tienen algún efecto sobre los resultados. Es decir, todas estas características (ej.: calidad metodológica del estudio) pueden ser consideradas como potenciales *variables independientes* para nuestro estudio meta-analítico.⁹ Como señalan Smith et al. (1980), “la estrategia menos productiva es intentar desechar hallazgos inconsistentes en base a prejuicios y argumentos *a priori* acerca de detalles técnicos del estudio” (p. 35).

5. *Obtención del tamaño del efecto (TE) de cada estudio.* En esta etapa se obtiene, con alguno de los procedimientos estadísticos indicados más arriba, el TE de cada estudio (o los TE, si hubiese más de uno: por ejemplo, de un estudio terapéutico con una evaluación final y un seguimiento al cabo del año, pudieran obtenerse, si lo consideramos interesante y con posibilidades de comparación con otros estudios en los que haya habido seguimientos semejantes, 2 TE diferentes). Esta etapa es crucial pues en aquí se obtiene la unidad métrica común que hace comparables los estudios originales incluidos en el MA.

6. *Obtención del TE medio.* Los TE procedentes de cada estudio se combinan y promedian, examinando además su distribución, sesgos, etc. Además, en esta etapa, se puede investigar la relación matemática existente entre las características específicas de cada estudio, recogidas en el proceso de codificación, y los TE. Por ejemplo, si en un MA de estudios de psicoterapia hemos codificado los años de experiencia de los terapeutas, se podría analizar si los estudios con psicoterapeutas con muchos años de experiencia tienen un TE diferente que los estudios que han empleado psicoterapeutas más inexpertos. Así pues, gracias al sistema de codificación empleado, se pueden estudiar las características diferenciales de subconjuntos de estudios. En este sentido se puede hablar del MA como “estrategia de regresión”, al permitir establecer la relación de determinadas características de cada estudio con el TE.

⁹ A modo de significativa anécdota, en los 475 estudios incluidos en un análisis de Smith et al., (1980) sobre los efectos diferenciales de las psicoterapias, se comprobó que la relación entre el “tamaño del efecto” y la calidad del estudio (validez interna) era prácticamente *nula* ($r = 0.03$). (Véanse algunas interpretaciones de este hallazgo en Wortman, 1983 y Kazdin, 1985).

Aunque a lo largo de este trabajo hemos ido presentando diversos problemas conceptuales y técnicos del MA, conviene resaltar algunas dificultades adicionales observables en este procedimiento.

En primer lugar, y esto es fundamental tenerlo en cuenta, el uso de términos cuantitativos en una revisión, en modo alguno supone eliminar aspectos *interpretativos*. Los aspectos cualitativos están presentes, de modo implícito o explícito, en todo estudio meta-analítico. Como muy acertadamente puntualizan Light & Pillemer (1984), la información numérica "no reduce el valor de la cuidadosa descripción del programa, los estudios de casos, los informes narrativos, o el juicio de expertos..." (p. 9) puesto que "...la precisión estadística no puede sustituir la claridad conceptual" (p. 11). El MA no obvia en modo alguno la intervención subjetiva de quien efectúa el estudio meta-analítico. Éste ha de emitir continuamente juicios y decisiones: desde la selección de estudios y características a estudiar hasta la propia *interpretación* de los resultados obtenidos. Como acertadamente afirman de nuevo Light & Pillemer (1982): "Los procedimientos formales pueden *detectar* diferencias sutiles, pero no pueden *explicarlas*. *Ofrecen un punto de partida, no una respuesta final*" (p. 12). Además, algunos autores, ya han señalado que hay que ser muy cautelosos con el empleo de técnicas meta-analíticas: su empleo correcto *no* es sencillo y, sin embargo, pueden dar una apariencia falsa de "rigor" o "cientificidad" por el simple hecho de emplear números (Cook & Gruder, 1978; Cook & Leviton, 1980; Hedges, 1987).

Una crítica ya tópica al MA es que "se mezclan manzanas con naranjas" (véase Smith et al., 1980, p. 47 y ss.); sin embargo, esto es justamente lo que hace el MA y lo que *pretende* hacer (Glass et al., 1981), si bien codificando las manzanas como manzanas y las naranjas como naranjas. La tarea de la ciencia es justamente apresar las invarianzas posibles en los resultados, descartando, o analizando con minuciosidad las discrepancias. Se trata de considerar, parafraseando la citada crítica, que ambas frutas comparten en alguna medida una característica común: la "fruteidad".

Un reto que tiene el MA es el de proporcionar cada vez respuestas cuantitativas más concretas sobre la efectividad de un programa o técnica determinados. Ya no sólo basta con saber si algo es efectivo o no, sino para *quién*, *cuándo*, y bajo qué *circunstancias* (Light & Pillemer, 1984, p. 19; Kazdin, 1985), especialmente en áreas como el estudio de la eficacia diferencial de las psicoterapias (véase Shapiro, 1985). Por otro lado, se necesitan herramientas estadísticas más sofisticadas para resolver los problemas que plantea el MA (Strube & Hartmann, 1983; Hedges & Olkin, 1985). Uno de estos problemas es precisamente la *heterogeneidad* de estudios que se

incluye dentro de una revisión. A falta de tales herramientas, es posible que los MA sean sólo capaces de ofrecer conclusiones muy generales, más bien "impresionistas", sobre los efectos de un programa o una intervención determinada. Mientras tanto, una solución a este problema puede ser el circunscribir algo más el ámbito de estudios revisados en un MA (Garfield, 1984; Shapiro, 1985).¹⁰

Asimismo, es predecible que un área importante de investigación sobre el MA será la dedicada a formular reglas básicas para analizar la calidad de un estudio dado meta-analítico, para lo que ya se han sugerido estrategias específicas (Hedges, 1987; Cooper, 1982, 1984; Jackson, 1980). Como ya hemos indicado, proporcionar índices cuantitativos no es suficiente para asegurar la calidad científica de un trabajo.

Como puede inferirse de lo anterior, el MA no está exento de problemas. Todavía no existen procedimientos estadísticos estándar para agregar resultados de diversos estudios y el fundamento matemático de tales procedimientos de hecho está desarrollándose en la actualidad (ej.: Hedges & Olkin, 1985, 1986). Asimismo, como señalamos páginas atrás, tampoco está resuelto el problema matemático de la posible falta de independencia de las observaciones cuando empleamos varios TE extraídos del mismo estudio (Glass et al., 1981; Strube & Hartmann, 1983).

Además el MA es un técnica *inapropiada* cuando el número de estudios a analizar es pequeño o cuando tales estudios son enormemente heterogéneos (Cook & Levinton, 1980; Light, 1983; Light & Pillemer, 1984); en estos casos, los estadísticos del MA tienen un alto riesgo de estar sesgados (Strube et al., 1985).

Otro problema de cierta consideración es el hecho de que la robustez del MA descansa en la calidad de los datos de los estudios originales. Sin embargo, a veces se carece de información suficiente sobre el diseño, características del estudio, e incluso sobre los datos hallados. Todo esto supone tanto problemas metodológicos como conceptuales (Strube et al., 1985; Kazdin, 1985; Searles, 1985). Por ejemplo, Shapiro & Shapiro (1983) advirtieron que una información tan necesaria y rudimentaria como los valores de las medias y las desviaciones típicas aparecían sólo en un 60 % de los estudios que incluyeron en su MA.

Junto a estos aspectos técnicos, los meta-analistas han de resolver otros aspectos más relacionados con la toma de decisiones, o juicios, que en todo

¹⁰ Por ejemplo, en nuestro estudio sobre estado de ánimo y memoria (Matt et al., 1990), nos hemos limitado a los estudios que han empleado el paradigma de "congruencia del estado de ánimo", excluyendo toda la literatura experimental fronteriza del paradigma de la "memoria dependiente del estado de ánimo" (véase Blaney, 1986).

estudio meta-analítico hay que efectuar. Como señalamos anteriormente, el MA no queda al margen de estos problemas; las reglas de inclusión para seleccionar tamaños del efecto para obtener el TE general, o el efecto que pueda tener el uso de codificadores diferentes para los estudios seleccionados (Matt, 1989), son áreas que aún requieren respuestas técnicas más precisas.

META-ANÁLISIS Y CONSTRUCCIÓN CIENTÍFICA

Ciencias "blandas" y ciencias "duras"

En nuestra opinión, uno de los mayores méritos del MA es justamente el permitirnos establecer *hechos*. Un argumento epistemológico típico para diferenciar las Ciencias Sociales de las Ciencias de la Naturaleza consiste en señalar que en estas últimas se produce un conocimiento *acumulativo* tanto empírico como teórico. Esto produce una profunda desazón en aquellos que nos hemos visto "relegados" a hacer "ciencia blanda". Pero, ¿caso décadas de investigación no han servido para nada? Desde luego no es difícil sucumbir a la tentación de pensar que las investigaciones en Psicología se suceden alocada e indisciplinadamente unas tras otras sin fuertes soportes teóricos y sin tener en cuenta los hallazgos de investigaciones previas. Aunque es posible que esta crítica no esté por completo infundada,¹¹ es asimismo posible que los "árboles" de la heterogeneidad de resultados de un experimento a otro, no nos permitan ver el "bosque" de la posible consistencia básica de los hallazgos. Si esto fuera efectivamente así, técnicas como el meta-análisis pueden tener una enorme importancia al ayudar a desvelar si los mecanismos de producción científica en Psicología dejan finalmente algún sedimento significativo de "hechos".

El uso del MA se ha popularizado extraordinariamente en sus casi quince años de existencia; la publicación de este tipo de revisiones tiene un fuerte crecimiento lineal (véase Gómez, 1987). Publicaciones como el *Psychological Bulletin*, la mejor revista existente en revisiones teóricas de la literatura psicológica, ya incorpora rutinariamente en casi todos sus números este tipo de revisiones cuantitativas. En la Tabla 2 se presentan algunos MA que, a modo de ejemplo, pueden ser ilustrativos de la expansión de estos procedimientos.

En un alentador y fascinante estudio, Hedges (uno de los metodólogos más importantes del MA) ha comparado los típicos procedimientos estadísticos del MA con las técnicas cuantitativas de revisión que, desde hace

¹¹ Véanse las duras críticas epistemológicas que Lakatos (1978), basándose fundamentalmente en algunos escritos de Meehl, efectúa al quehacer de la Psicología.

décadas, utilizan los físicos (Hedges, 1987). Lo más interesante del estudio no es sólo comprobar que ambas técnicas son matemáticamente muy *similares*, sino que la heterogeneidad de resultados en experimentos de física dura (como por ejemplo, la determinación de la masa y vida media de partículas elementales) es semejante a la heterogeneidad de resultados hallados en áreas tan importantes de investigación en Psicología como, por ejemplo, el estudio de diferencias cognitivas según el sexo, o el efecto que las expectativas del profesor tienen sobre el CI de los alumnos. Los márgenes de *error* hallados en diversas áreas de la física (ej.: la estimación de la masa de los protones o los electrones, o los valores de conductividad térmica de los elementos químicos) no tienen nada que envidiar a los hallados en muchas áreas de la Psicología. Incluso se sabe, y esto nos resultará familiar a los psicólogos, que en las investigaciones sobre Rayos X, los resultados de un laboratorio suelen diferir mucho de los resultados hallados en otros laboratorios empleando los mismos instrumentos y las mismas medidas.

TABLA 2

Algunas investigaciones meta-analíticas efectuadas en diversas áreas psicológicas

Smith & Glass (1977):	Efectividad psicoterapias (375 estudios y 850 ME).
Smith, Glass, & Miller (1980):	Efectividad psicoterapias (450 estudios y 1760 ME).
White (1982):	Nivel socioeconómico y rendimiento escolar.
Mumford, et al. (1982):	Intervenciones psicológicas en Cardiología.
Hovell (1982):	Diets e hipertensión.
Raudenbush (1982, 1983):	Efecto Rosenthal (expectativas profesor y CI).
Stoffelmayr et al. (1983):	Funcionamiento premórbido y esquizofrenia.
Straw (1983):	Desinstitucionalización de pacientes psiquiátricos.
Sweeney, et al. (1986):	Depresión y estilo Atribucional (104 estudios/15000 Ss).
Lambert, et al. (1986):	Escalas de depresión (BDI, HRS, y Zung): 1850 Ss.
Steinbrueck (1983):	Depresión unipolar.
Hazelrigg, et al. (1987):	Efectividad de la Terapia Familiar.
Eisenberg & Miller (1987):	Empatía y Conductas Prosociales.
Booth & Friedman (1987):	Predictores psicológicos de alteraciones del corazón.
Achenbach, et al. (1987):	Recogida de inform. en psicopatología infanto-juvenil.
Suls & Wan (1989):	Información previa y dolor.
Dobson (1989):	Terapia en depresión.

La Psicología ha aportado realmente hechos y, seguramente, muchos de ellos están aún *ocultos* esperando ser rescatados de entre montañas de experimentos e investigaciones heterogéneas. Un instrumento de "rescate" adecuado pudiera ser justamente las revisiones cuantitativas de la literatura. Su uso favorecerá, sin duda, la acumulación de conocimientos *empíricos*, lo que

constituye un paso imprescindible para lo que realmente carece la Psicología: la acumulación *teórica* (Silva, 1989).

Naturalmente existen muchas diferencias entre las ciencias duras y las blandas (e.g., Meehl, 1978; Kruskal, 1981; Lakatos, 1978; Dar, 1987) y no es mi propósito realizar un esfuerzo pretendidamente homogeneizador entre ambas áreas. Sin embargo, parece bastante probable que la falta de acumulación que padecemos en las ciencias blandas se debe más a *prácticas científicas viciadas* e inconsistentes, como las que anteriormente he ido señalando, que a limitaciones epistemológicas intrínsecas impuestas por el peculiar objeto de estudio que tiene la Psicología.

El papel del meta-análisis en el desarrollo científico

Por último quisiera discutir brevemente el papel que el meta-análisis puede desempeñar dentro del armazón epistemológico. Ya he señalado la importante función que puede tener como herramienta acumulativa. Esto es esencial, no sólo por razones de eficacia o de construcción progresiva del conocimiento, sino porque la *investigación* meta-analítica es un modo *legítimo* de avance científico y no menos importante que la investigación tradicional. La vieja pretensión de desarrollar "experimentos cruciales" que den respuesta inequívoca a problemas teóricos, nos parece hoy una quimera ajena a las modernas propuestas epistemológicas, que se basan más en la acumulación de conocimientos y en el desarrollo de "programas de investigación" (Rivadulla, 1987). Por otro lado, esa posibilidad acumulativa es muy necesaria dada la diáspora de procedimientos, técnicas, tipo de sujetos, etc., que se emplean en la investigación en las Ciencias Sociales.

Otra de las ventajas que tiene el MA es que puede proporcionar *respuestas* relativamente contundentes y claras sobre un área de investigación. Esto es sumamente importante para la gestión y para aquellos políticos, administradores, etc., que han de tomar decisiones (Light & Pillemer, 1982, 1984). Por ejemplo, en Australia se han empleado técnicas meta-analíticas (bajo el auspicio de agencias gubernamentales) para hallar cuáles deben ser las directrices terapéuticas, en los hospitales estatales, para el tratamiento de la depresión y la esquizofrenia (Quality Assurance Project, 1982). En este caso, el MA proporcionó respuestas bastante claras y sencillas sobre aspectos tales como el tipo de tratamiento más eficaz, el tipo de paciente que se beneficiará más del mismo, etc.

A pesar de su todavía corta vida, el MA ha ejercido una influencia notable en la investigación psicológica. Además de proporcionar información valiosa, no pocas veces origina fructíferas polémicas sobre un área de investigación específica. Por ejemplo, Glass (1978) demostró en su pionero estudio

que *todas* las psicoterapias son más efectivas que la ausencia de terapia o terapias placebo¹² lo que desencadenó una airada respuesta de Eysenck (1978), iniciándose así una disputa que aún perdura (Searles, 1985; Shapiro, 1985; Matt, 1989). Asimismo, el MA ha favorecido una mayor *sensibilidad* en los investigadores hacia determinados aspectos metodológicos. Por ejemplo, ya empieza a ser muy infrecuente que los autores no ofrezcan información sobre medias y desviaciones típicas, algo que fue originalmente observado y criticado por los teóricos del MA.¹³

Otra consecuencia del impulso de las investigaciones meta-analíticas es que cada vez se considera más la necesidad de proporcionar, en los resultados de la investigación, datos sobre la *magnitud* de los hallazgos. El MA ha puesto en tela de juicio el sentido que habitualmente se otorga a la significación estadística; en efecto, como señalamos anteriormente en este mismo trabajo, el que un estadístico alcance un nivel arbitrario de significación (ej.: el 5 % o el 1 %) no nos dice absolutamente nada sobre la *magnitud* del efecto hallado o, en otras palabras, sobre la *importancia* de los resultados obtenidos. Hay incluso autores (ej.: Rosenthal, 1983; Sechrest & Yeaton, 1982) que han propuesto que todas las investigaciones den cuenta, en la medida de lo posible, de este tipo de información; por ejemplo, sería deseable que cuando se informe de los resultados de un ANOVA, se ofrezca también información con estadísticos como, por ejemplo, "omega al cuadrado" (Ω^2), de la magnitud de las significaciones o que, cuando se muestran correlaciones Pearson, se interpreten los datos en función de la varianza explicada (r^2). Todo esto constituye, en mi opinión, un revulsivo para mejorar la calidad de nuestras investigaciones.

CONCLUSIONES

Dados todos los argumentos presentados a lo largo de estas páginas, nos parece obvia la necesidad de que los científicos sociales conozcan al menos las características básicas y las posibilidades y las limitaciones de este tipo de procedimientos: en 1984 sólo un 33 % de autores que habían publicado algún artículo en el *Psychological Bulletin* decían estar familiarizados con este tipo

¹² Esta conclusión ha sido posteriormente ratificada por Shapiro & Shapiro (1982), partiendo de una base de estudios prácticamente diferente a la de Glass. No obstante, en este estudio se demostró que las terapias conductuales eran ligeramente más eficaces que otros tipos de psicoterapia.

¹³ Esto se debió, en gran medida, al hecho de que estos estadísticos eran casi imprescindibles, para llevar a cabo los primeros estudios de meta-análisis.

de revisión (Jackson, 1984, cit. en Gómez, 1987). Los críticos de los métodos cuantitativos como el meta-análisis, han de tener muy en cuenta que la ciencia es acumulativa; por lo tanto, criticar que en el MA se incluyen estudios heterogéneos, o de dudosa calidad, es criticar el modo de actuación integrador y superador de la ciencia (Hedges & Olkin, 1986).

A pesar de las dificultades técnicas del MA, ningún científico social debería estar ajeno a sus aportaciones y, como desgraciadamente suele ser habitual, incurrir en críticas manidas e ignorantes sobre el alcance de este procedimiento. El manejo racional y fructífero de información se hace cada día más difícil por el inabarcable flujo de información al que estamos sometidos. El matemático David Hilbert ingeniosamente señalaba que "...la importancia de un trabajo científico puede medirse por el número de publicaciones previas que hace superfluo leer" (cit. en Walberg & Haertel, 1980). Quizás el MA pueda cumplir cada vez más esta función de filtro necesario para el desarrollo científico de la Psicología.

SYNTHESIS

Scientific investigation requires periodic indicators which summarize the effectiveness of a given of investigation, a basic procedure which allows for the accumulation of knowledge which characterizes what is called "hard science". The "post-Kuhnian" epistemology emphasizes how the advancement of science does not happen by sudden and striking innovative contributions (the so called "crucial experiments"), but rather progresses according to a slower, methodical accumulation of empirical data and subsequent theory development. One of psychology's basic problems is precisely the difficulty in establishing well defined "facts", that is the replication of empirical results.

Traditional literature reviews, with their grand potential for the filtering and synthesizing of data, are vulnerable to a series of factors which can ultimately limit their quality. It is particularly difficult, for example, to efficiently synthesize a large quantity of relatively heterogeneous information, and subsequently draw conclusions based on the dozens of studies typically included in such reviews. Furthermore, the authors of such literature reviews frequently commit the grave error (common in psychology in general) of misinterpreting the obtained levels of significance ("p") Given these problems, literature reviews provide only limited conclusions, and ones which often covary with the theoretical bias of the author.

Meta-analysis (MA) is a quantitative literature procedure which provides mathematical indices for the effectiveness and consistency of given findings.

Therefore, one may employ further statistical techniques to aggregate and analyse the results from various studies in a given area.

In this paper, the author presents several guidelines for the potential "consumer" of MAs. This paper does not provide a detailed analysis of the technical characteristics of MA, but rather argues for the value of such techniques in scientific development within the social sciences. MA can be a powerful epistemological tool for deducing the consistency of findings reported in a series of related but varying studies.

The MA procedure generally follows several steps: 1) To understand the basic characteristics of the area of study in question so as to be able to propose several hypotheses; 2) To select the maximum number of original studies possible for the MA; 3) To develop a coding system which allows one to quantify the basic characteristics of each individual study; and 4) To determine the importance or effect size for the findings of each individual study which are later combined via aggregation techniques to arrive at a global estimate based on the combination of findings of all of the studies included in the MA.

This MA procedure, however, can be employed according to several variations depending on the particular MA model in use, such as the employment of parametric estimates of global effect size. It is probably most useful to consider MA as a methodological procedure available to researchers in a given discipline for use in determining the utility of a given line of investigation. It is suggested that MA be considered a reasonable alternative methodological procedure which provides a more formal response to the same questions treated by classical literature reviews.

In any case, MA allows one to analyze the viability of determined research areas or investigative procedures, or more importantly, to provide information concerning a determined finding. Even given the limitations of MA, and given its current state of continuing development, the author argues that MA opens a fertile area for scientific construction in psychology.

BIBLIOGRAFÍA

- ACHENBACH, T. M.; MCCOUNAUGHY, S. H. & HOWELL, C. T. (1987). Child/Adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, 101, 213-222.
- BANGERT-DROWNS, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin*, 99, 388-399.
- BLANEY, P. (1986). Memory and mood. *Psychological Review*, 87, 220-238.
- BOOTH-KEWLEY, S. & FRIEDMAN, H. S. (1987). Psychological predictors of heart disease: A quantitative review. *Psychological Bulletin*, 101, 343-362.

- CARVER, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48, 378-399.
- COHEN, J. (1977). *Statistical power for the behavioral sciences*. (2nd. Ed.). New York: Academic.
- COOK, T. D. & GRUDER, C. L. (1978). Meta-evaluation research. *Evaluation Quarterly*, 2, 5-51.
- COOK, T. D. & LEVITON, L. C. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality*, 48, 449-472.
- COOK, T. D. & REICHARDT, C. S. (1982). *Qualitative and quantitative methods in evaluation research*. Beverly Hills, CA: Sage. (Trad. esp. en Ed. Morata, 1986).
- COOPER, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, 52, 291-302.
- (1984). *The integrative research review: A systematic approach*. Beverly Hills, CA: Sage.
- COOPER, H. M. & ARKIN, R. (1981). On quantitative reviewing. *Journal of Personality*, 49, 225-230.
- COOPER, H. M. & ROSENTHAL, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin*, 87, 442-449.
- COWLES, M. & DAVIS, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, 37, 553-558.
- DAR, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist*, 42, 145-151.
- DOBSON, K. S. (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, 57, 414-419.
- EISENBERG, N. & MILLER, P. A. (1987). The relation of empathy to prosocial and related behaviors. *Psychological Bulletin*, 101, 91-119.
- EYSENCK, H. J. (1978). An exercise of mega-silliness. *American Psychologist*, 33, 517.
- GARFIELD, S. L. (1984). Psychotherapy: Efficacy, generality, and specificity. En J. B. WILLIAMS & R. L. SPITZER (Eds.), *Psychotherapy research. Where are and where should we go?* New York: Guilford.
- GLASS, G. V. (1977). Integrating findings: The meta-analysis of research. En L. S. SCHULMAN (Ed.), *Review of Research in Education*, 5, 351-379.
- (1980). Summarizing effect sizes. En R. ROSENTHAL (Ed.), *Quantitative assessment of research domains*. San Francisco, CA: Josey-Bass.
- GLASS, G. V.; MCGAW, B. & SMITH, M. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- GLASS, G. V. & SMITH, M. L. (1978). *Meta-analysis of research on the relationship of class size and achievement*. San Francisco, CA: Far West Laboratory for Educational Research and Development.
- GÓMEZ, J. (1987). *Meta-análisis*. Barcelona: Promociones y Publicaciones Universitarias.
- GREEN, B. F. & HALL, J. A. (1984). Quantitative methods for literature reviews. *Annual Review of Psychology*, 35, 37-53.
- HARCUM, E. R. (1989). The highly inappropriate calibrations of statistical significance. *American Psychologist*, 44, 964.
- (1990). Methodological versus empirical literature: Two views on casual acceptance of the null hypothesis. *American Psychologist*, 45, 404.
- HAZELRIGG, M. D.; COOPER, H. M. & BORDUIN, C. M. (1987). Evaluating the effectiveness of family therapies: An integrative and analysis. *Psychological Bulletin*, 101, 428-442.
- HEDGES, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, 92, 490-499.
- (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist*, 42, 443-455.

- HEDGES, L. V. & OLKIN, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, 88, 359-369.
- HEDGES, L. V. & OLKIN, I. (1985). *Statistical methods for meta-analysis*. New York: Academic.
- HEDGES, L. V. & OLKIN, I. (1986). Meta Analysis: A review and a new view. *Educational Researcher*, 15, 14-21.
- HOVELL, M. F. (1982). The experimental evidence for weight-loss treatment of essential hypertension: A critical review. *American Journal of Public Health*, 72, 359-368.
- HUBERTY, C. J. (1987). On statistical testing. *Educational Researcher*, 16, 4-49.
- HUNTER, J. E.; SCHMIDT, F. L. & JACKSON, G. B. (1982). *Meta-Analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
- JACKSON, G. B. (1980). Methods for integrative reviews. *Review for Educational Research*, 50, 438-460.
- KAMIN, L. (1978). Comment on Munsinger's review of adoption studies. *Psychological Bulletin*, 85, 194-201.
- KAZDIN, A. (1985). The role of meta-analysis in the evaluation of psychotherapy. *Clinical Psychology Review*, 5, 49-61.
- KENDALL, P. C. & MARUYAMA, G. (1985). Meta-analysis: On the road to synthesis of knowledge? *Clinical Psychology Review*, 5, 79-89.
- KRUSKAL, W. (1981). Statistics in society: Problems unresolved and unformulated. *Journal of the American Statistical Association*, 76, 505-515.
- LAKATOS, I. (1978). Falsification and the methodology of scientific research programs. En J. Worrall & G. Currie (Eds.), *The methodology of scientific research programs: Imre Lakatos philosophical papers* (Vol. 1, pp. 139-167). Cambridge: Cambridge University Press.
- LAMBERT, M. J.; HATCH, D. R.; KINGSTON, M. D. & EDWARDS, C. (1986). Zung, Beck, and Hamilton Rating Scales as measures of treatment outcome: A meta-analytic comparison. *Journal of Consulting and Clinical Psychology*, 54, 54-59.
- LEVITON, L. C. & COOK, T. D. (1981). What differentiates meta-analysis from other forms of review? *Journal of Personality*, 49, 231-236.
- LIGHT, R. J. & PILLEMER, D. B. (1982). Numbers and narrative: Combining their strengths in research reviews. *Harvard Educational Review*, 52, 1-26.
- LIGHT, R. J. & PILLEMER, D. B. (1984). *Summing up. The science of reviewing research*. Cambridge, MA: Harvard University Press.
- MACCOBY, E. E. & JACKLIN, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.
- MATT, G. E. (1989). Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychological Bulletin*, 105, 106-115.
- MATT, G. E.; VÁZQUEZ, C. & CAMPBELL, K. (1990). *Mood and memory: A meta-analysis*. San Diego State University. En preparación.
- MEEHL, P. E. (1978). Theoretical risks and tabular asteriks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806-834.
- MUMFORD, E.; SCHLESINGER, H. J. & GLASS, G. V. (1982). The effects of psychological intervention on recovery from surgery and heart attacks: An analysis of the literature. *American Journal of Public Health*, 72, 141-151.
- MUNSINGER, H. (1978). Reply to Kamin. *Psychological Bulletin*, 85, 202-206.
- MURRAY, L. W. & DOSSER, D. A. (1987). How significant is a significant difference? Problems with the measurement of magnitude of effect. *Journal of Counseling Psychology*, 34, 68-72.
- NUCHTERLEIN, K. H. (1977). Reaction time and attention in schizophrenia: A critical evaluation of the data and theories. *Schizophrenia Bulletin*, 3, 373-428.
- QUALITY ASSURANCE PROJECT, THE (1982). A methodology for preparing 'ideal' treatment outlines in psychiatry. *Australian and New Zealand Journal of Psychiatry*, 16, 153-158.

- RAUDENBUSH, S. W. (1983). Utilizing controversy as a source of hypotheses for meta-analysis. The case of teacher expectancy's effects on pupil IQ. En R. J. LIGHT (Ed.), *Evaluation Studies: Review Annual (vol. 8)*, pp. 303-325. Beverly Hills, CA: Sage.
- RIVADULLA, A. (1986). *Filosofía actual de la ciencia*. Madrid: Tecnos.
- ROSENTHAL, R. (1978). Combining results of independent studies. *Psychological Bulletin*, 85, 185-193.
- (1979). The 'file drawer problem' and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- (1980a). Combining probabilities, and the file drawer problem. *Evaluation in Education*, 4, 18-21.
- (1980b). Summarizing significance levels. En R. ROSENTHAL (Ed.), *New directions for methodology of social and behavioral science (Vol. 5: Quantitative Assessment of Research domains)*. San Francisco, CA: Sage.
- (1983). Assessing the statistical and social importance of the effects of psychotherapy. *Journal of Consulting and Clinical Psychology*, 51, 4-13.
- ROSENTHAL, R. & RUBIN, D. D. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, 92, 500-504.
- ROSENTHAL, R. & RUBIN, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin*, 99, 400-406.
- ROSNOW, R. L. & ROSENTHAL, R. (1989). Statistical procedures and the justification knowledge in psychological science. *American Psychologist*, 44, 1276-1284.
- SEARLES, J. S. (1985). A methodological and empirical critique of psychotherapy outcome meta-analysis. *Behavior Research & Therapy*, 23, 453-463.
- SECHREST, L. & YEATON, W. H. (1982). Magnitudes of experimental effects in social science research. *Evaluation Review*, 6, 579-600.
- SHAPIRO, D. A. (1985). Recent applications of meta-analysis in clinical research. *Clinical Psychology Review*, 5, 13-34.
- SHAPIRO, D. A. & SHAPIRO, D. (1983). Comparative therapy outcome research: Methodological implications of meta-analysis. *Journal of Consulting and Clinical Psychology*, 51, 42-53.
- SHAPIRO, D. A. & SHAPIRO, D. (1985). Meta-analysis of comparative therapy outcome research: A replication and refinement. *Psychological Bulletin*, 92, 581-604.
- SILVA, F. (1989). *Evaluación conductual y criterios psicométricos*. Madrid: Pirámide.
- SMITH, M. L. (1980). Publication bias and meta-analysis. *Evaluation in Education*, 4, 22-24.
- SMITH, M. L. & GLASS, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, 32, 752-760.
- SMITH, M. L.; GLASS, G. V. & MILLER, T. I. (1980). *The benefits of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.
- STEINBRUECK, S. M.; MAXWELL, S. E. & HOWARD, G. S. (1983). A meta-analysis of psychotherapy and drug therapy in the treatment of unipolar depression with adults. *Journal of Consulting and Clinical Psychology*, 51, 856-863.
- STOFFELMAYR, B. E.; DILLAVOU, D. & HUNTER, J. E. (1983). Premorbid functioning and outcome in schizophrenia: A cumulative analysis. *Journal of Consulting and Clinical Psychology*, 51, 338-352.
- STRAW, R. B. (1983). Deinstitutionalization in Mental Health. A meta-analysis. En R. J. LIGHT (Ed.), *Evaluation Studies: Review Annual (vol. 8)*, pp. 253-278. Beverly Hills, CA: Sage.
- STRUBE, M. J. (1985a). Combining and comparing significance levels. *Psychological Bulletin*, 97, 334-341.
- (1985b). Power analysis for combining significance levels. *Psychological Bulletin*, 98, 595-599.

- STRUBE, M. J.; GARDNER, W. & HARTMANN, D. P. (1985). Limitations, liabilities, and obstacles in reviews of the literature: The current status of meta-analysis. *Clinical Psychology Review*, 5, 63-78.
- STRUBE, M. J. & HARTMANN, D. P. (1983). Meta-analysis: Techniques, applications, and functions. *Journal of Consulting and Clinical Psychology*, 51, 14-27.
- SULS, J. & WAN, C. K. (1989). Effects of sensory and procedural information on coping with stressful medical procedures and pain: A meta-analysis. *Journal of Consulting and Clinical Psychology*, 57, 372-379.
- SWEENEY, P. D.; ANDERSON, K. & BAILEY, S. (1986). Attributional style in depression: A meta-analytic review. *Journal of Personality & Social Psychology*, 50, 974-991.
- VÁZQUEZ, C. (1985). Limitaciones y sesgos y en el procesamiento de la información. *Estudios Psicología*, 23-24, 111-133.
- WALBERG, H. J. & HAERTEL, E. H. (1980). Research integration: An introduction and overview. *Evaluation in Education*, 4, 5-10.
- WHITE, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, 91, 461-481.
- WORTMAN, P. M. (1983). Evaluation research: A methodological perspective. *Annual Review of Psychology*, 34, 223-260.
- YEATON, W. H. & SECHREST, L. (1981). Meaningful measures of effect. *Journal of Personality & Social Psychology*, 49, 766-767.