Revista de Psicología de la PUCP. Vol. XXI, 1, 2003 Edición de aniversario

Magnitud del Efecto: Una guía para investigadores y usuarios

Robert Coe[1]
Universidad de Durham
César Merino Soto[2]
Defensoría Municipal del Niño y del Adolescente, DEMUNA – Chorrillos

El presente artículo describe un método para cuantificar la magnitud de las diferencias entre dos mediciones y/o el grado del efecto de una variable sobre un criterio, y es llamado la medida de la magnitud del efecto, *d*. Su uso en contextos de investigación y aplicados proporciona un información complementaria bastante descriptiva, mejorando la interpretación de los resultados obtenidos por los métodos tradicionales que enfatizan la significación estadística. Existen varias formas de interpretar el estadístico *d*, y se presenta un ejemplo, tomado de una investigación experimental, para aclarar los conceptos y cálculos necesarios. Este método no es robusto a ciertas condiciones que pueden distorsionar su interpretación, por ejemplo, la no normalidad de los datos entre otros; se mencionan métodos alternativos al estadístico *d*. Finalizamos con unas conclusiones que advierten sobre su apropiado uso.
Palabras clave: magnitud del efecto, meta-análisis, significancia estadística, metodología, investigación experimental

Effect Size: A guide for researchers and users
The present article describes a method to quantify the magnitude of the differences between two measures and/or the degree of the effect of a variable about criteria, and it is named like the effect size measure, *d*. Use it use in research and applied contexts provides a quite descriptive complementary information, improving the interpretation of the results obtained by the traditional methods that emphasize the statistical significance. Several forms there are of interpreting the *d*, and an example taken of an experimental research, is presented to clarify the concepts and necessary calculations. This method is not robust to some conditions that they can distort its interpretation, for example, the non normality of the data; alternative methods are mentioned to the statistical *d*. We ending with some conclusions that will notice about the appropriate use of it.
Key words: effect size, meta-analysis, statistical significance, methodology, experimental research.

[1] Profesor de la Escuela de Educación de la Universidad de Durham. Sus actividades de docencia universitaria las comparte con la capacitación a colegios y desarrollo de nuevos proyectos educativos. Actualmente es Director del Grupo de Evaluación Educativa en el Curriculum, Evaluation and Management (CEM) Centre at Durham University. Dirección:

La Magnitud del Efecto *(ME)*[3] es simplemente una manera de cuantificar la efectividad de una particular intervención, relativa a alguna com-paración. Es fácil de calcular y entender, y puede aplicarse a algún resultado medido en educación o ciencias sociales. Este concepto nos permite movernos más allá de la simple pregunta "¿el método *A* es efectivo o no? a una más sofisticada como "¿Qué tan bien funciona el método *A* en varios contextos?". Más aún, al poner énfasis en el aspecto más importante de una intervención –la magnitud del efecto– más que en su significancia estadística (que pone en conflicto a la magnitud del efecto y el tamaño de la muestra), promueve un enfoque más científico a la acumulación de conocimientos. Por estas razones, la *ME* es una herramienta importante para reportar e interpretar la efectividad de una condición específica o para describir las diferencias.

El uso rutinario de *ME*, sin embargo, se ha limitado frecuentemente al enfoque denominado Meta-análisis –que combina y compara estimaciones provenientes de diferentes estudios– y generalmente es bastante raro hallarlo en los reportes de la investigación educacional. Las fórmulas para su cálculo no aparecen en la mayoría de los libros de estadística (pero sí más bien en aquellos dedicados al meta-análisis) y raramente son enseñados en los cursos tradicionales de investigación. Por estas razones, incluso el investigador que está convencido de lo

---

Escuela de Educación, Universidad de Durham, Reino Unido. Leazes Road - Durham DH1 1TA – UK. Correo electrónico: r.j.coe@dur.ac.uk, r.j.Coe@cem.dur.ac.uk.

2 Psicólogo, licenciado, graduado en la Universidad Inca Gracilazo de la Vega (Perú). Actualmente es responsable del Servicio de Psicología de la DEMUNA – Chorrillos. Presta asistencia en la intervención y prevención del maltrato intrafamiliar y escolar. Sus investigaciones se orientan hacia la metodología psicométrica y hacia el maltrato en el contexto familiar y escolar. Dirección: Servicio de Psicología, Defensoría Municipal del Niño y del Adolescente (DEMUNA) - Av. José Olaya 166 (Casa de la Cultura) - Chorrillos – Lima 9 – Perú. Correo electrónico: sikayax@yahoo.com.ar.

3 El término *Magnitud delEfecto* corresponde a lo que en inglés significa *effect size*.

apropiado de usar medidas de *ME* y que no teme confrontar la ortodoxia de la práctica convencional, puede hallar que es bastante difícil saber cómo hacerlo.

La presente guía está escrita para no-estadísticos, aunque inevitablemente se aplicarán algunas ecuaciones y lenguaje técnico. Describiremos qué es la *ME*, qué significa, cómo se lo puede utilizar y qué potenciales problemas se asocian a su uso. En las últimas secciones, se incluyen referencias a otras medidas *ME* alternativas y conclusiones a modo de sugerencias.

¿Por qué se necesita una medida de la Magnitud del Efecto?

Consideremos el experimento conducido por Val Dowson (2000) para investigar el efecto de la hora del día sobre el aprendizaje: ¿los niños aprenden mejor en la mañana o en la tarde? Se incluyó un grupo de 38 niños en el experimento. La mitad aleatoriamente se incluyó en el grupo para escuchar una historia y responden preguntas cerca de las 9am; la otra mitad, escuchó la misma historia (en una grabación) y respondía las mismas preguntas a las 3pm. El nivel de comprensión se midió por el número de preguntas correctamente respondidas.

El puntaje promedio fue 15.2 para el grupo de la mañana y 17.9 para el de la tarde: hubo una diferencia de 2.7. Pero ¿qué tan grande es esta diferencia? Si el resultado se hubiera medido en una escala conocida, tal como las calificaciones escolares, interpretar estas diferencias no sería un problema. En un sistema escolar que utilice una calificación vigesimal (de 0 a 20), si la diferencia fuera, por decir, 4 puntos, la mayoría de las personas podrían tener una idea clara de la significancia educativa del efecto que tuvo la hora de estudio sobre la lectura en nuestro ejemplo. Sin embargo, en muchos experimentos no hay una escala conocida disponible sobre el cual registrar los resultados. El experimentador frecuentemente tiene que crear una escala o utilizar (o adaptar) uno que ya existe, pero generalmente no será uno cuya interpretación sería familiar para la mayoría de las personas.

Una forma de abordar este problema es utilizar el monto de variación en los puntajes para contextualizar la diferencia. Si no hubiera traslapamiento en todas las personas del grupo "tarde" que se desempeñó mejor en las pruebas que todos los del grupo "mañana", entonces esto podría parecer como una diferencia importante. Por otro lado, si la dispersión de los puntajes fuera grande y el traslape fuera mayor que la diferencia entre los grupos, entonces el efecto podría parecer menos significante. Debido a que tenemos una idea del monto de variación encontrada dentro de un grupo, podemos utilizarlo como una regla contra el cual comparar la diferencia encontrada. Esta diferencia es cuantificada en el cálculo de la *magnitud del efecto*. El concepto de ilustra en la Figura 1, que muestra las dos posibles maneras en que la diferencia podría variar en relación del grado de traslape existente. Si la diferencia fuera como la que aparece en el gráfico (a), tal diferencia podría ser significativa; en el gráfico (b), en cambio, difícilmente se podría reconocer una diferencia existente.

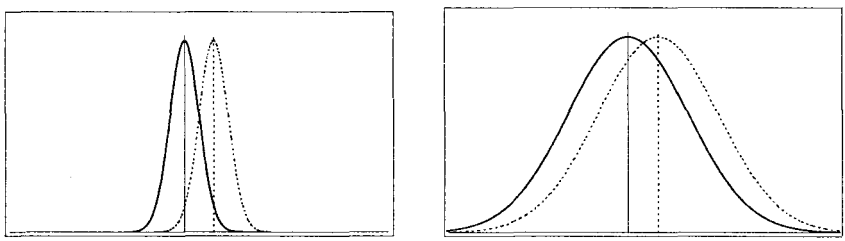

Figura 1. Traslape observado en dos distribuciones hipotéticas.

¿Cómo se calcula?

La *ME* es solo la diferencia media estandarizada entre los dos grupos. En otras palabras:

$$ME = \frac{\text{[Media del grupo experimental]} - \text{[Media del grupo control]}}{\text{Desviación Estándar}}$$

Ecuación 1

Aún si no es obvio cuál de los dos grupos es el *experimental* (es decir, al que se le aplica el *nuevo* tratamiento que se esta probando) y cuál el *control* (al que se le aplica el tratamiento estándar –o sin tratamiento– para propósitos de comparación), la diferencia se puede calcular. En este caso, *la ME* simplemente mide la diferencia entre ellos.

La desviación estándar (*DE*) en la fórmula es una habitual medida de dispersión de un conjunto de valores[4]. Aquí nos referimos a la desviación estándar de la población desde el cual los diferentes grupos de tratamiento fueron tomados. En la práctica, sin embargo, mayormente este valor nunca es conocido, así que debe ser estimado tomando la desviación estándar del grupo control o desde el valor *concentrado* de ambos grupos (ver la sección 4, más adelante, para una discusión de este aspecto).

En el experimento de Dowson, sobre la hora del día, la desviación estándar (*DE*) = 3.3, de tal modo que la magnitud del efecto es (17.9 – 15.2)/3.3 = 0.8.

¿Cómo se interpreta?

Una característica de alguna *ME* es que puede ser directamente convertido en afirmaciones sobre el traslapamiento entre dos muestras en términos de percentiles.

Una *ME* es exactamente equivalente a un puntaje *Z* de una distribución normal. Por ejemplo, una tamaño del efecto de 0.8 significa que el puntaje de la persona promedio en el grupo experimental es 0.8

[4] En los textos de estadísticas se encuentran varias fórmulas para calcular la desviación estándar, y pueden también ser construidas en hojas de cálculo como MS Excel. Una fórmula simple para un conjunto de valores, X_1, X_2, ...X_n, con media M_X es:

$$DE = \sqrt{\frac{X_1^2 + X_2^2 + \ldots + X_n^2 \quad n(M_x)^2}{n \quad 1}}$$

desviaciones estándar arriba de la persona promedio en el grupo control, y que excede los puntajes del 79% del grupo control. Con los dos grupos de 19 en el experimento de la hora del día, la persona promedio en el grupo "tarde" (es decir, aquel que podría haber estado en el puesto décimo de aquel grupo) podría haber puntuado la cuarta persona más alta en el grupo "mañana". La visualización de estos dos individuos puede dar una interpretación gráfica de la diferencia entre los dos efectos.

El Cuadro 1 muestra las conversiones de magnitud el efecto a percentiles (I_1) y el cambio equivalente en el orden de rango para un grupo de 25 (I_2). Por ejemplo, para una *ME* de 0.6, el valor de 73% indica que la persona promedio en el grupo experimental podría puntuar más alto que el 73% de un grupo control inicialmente equivalente. Si el grupo consiste de 25 personas, esto es lo mismo que decir que la persona promedio (es decir, la ubicada en el lugar 13 en el grupo) podría estar ahora al nivel de una persona ubicada en el lugar 7 en el grupo de control. Note que una *ME* de 1.6 podría elevar a la persona promedio a estar al nivel de la persona con mejor posición (primer lugar) en el grupo de control; de esta manera, la magnitud del efecto se ilustra en términos de la persona que está a la cabeza en el grupo más grande.

Otra manera de conceptualizar el traslapamiento es en términos de la probabilidad en que uno podría adivinar de qué grupo proviene una persona, basado solamente en su puntaje en la prueba. –o cualquier otro valor que está siendo comparado. Si la *ME* fuera 0 (es decir, los dos grupos son los mismos) entonces la probabilidad de una correcta adivinación podría ser exactamente la mitad, o 0.50. Teniendo una diferencia entre los dos grupos equivalente a una *ME* de 0.3, aún hay bastante traslapamiento y la probabilidad de identificar correctamente los grupos aumenta ligeramente a 0.56. Con una *ME* de 1, la probabilidad es ahora 0.69, apenas encima de dos tercios por chance. Estas probabilidades se muestran en la cuarta columna (I_3) del Cuadro 1. Es claro que el traslapamiento entre los grupos experimental (GE) y

control (GC) es sustancial (y por lo tanto la probabilidad todavía está cerca de 0.5), aún cuando la *ME* es bastante grande.

Una manera ligeramente diferente de interpretar *ME* se refiere a la equivalencia entre la diferencia media estandarizada (*d*) y el coeficiente de correlación, *r*. Si la pertenencia a un grupo es codificada como una variable "dummy" (es decir, que el grupo control es denotado con 0 y el grupo experimental con 1) y se calcula la correlación entre esta variable y la medida del resultado, se puede obtener el valor r. Haciendo algunos presupuestos adicionales, uno puede convertir *d* en *r*[5]. Rosenthal y Rubin (1982) aprovechan una interesante propiedad de *r* para sugerir una adicional interpretación. Si un resultado se reduce a una simple dicotomía (por ejemplo, si un puntaje esta debajo o encima de una valor particular, como la mediana, se podría asignar a cada valor como "fracaso" o "acierto"), *r* se puede interpretar (esto requiere, nuevamente, que se hagan algunos presupuestos estándar) como la diferencia en las proporciones de cada categoría. Por ejemplo, un *ME* de 0.2 indica una diferencia de 0.10 en estas proporciones, como podría ser el caso si el 45% del grupo control y el 55% del grupo de tratamiento alcanzan algún umbral de "éxito". Estos valores se muestran en (I_4).

Se debe advertir que los valores en el Cuadro 1 dependen del presupuesto de una distribución Normal. La interpretación de la *ME* del efecto en términos de percentiles es muy sensible a la violación de esta interpretación (ver más abajo).

[5] $r^2 = d^2 / (4+d^2)$. Ver Cohen, 1969, pp20-22 para otras fórmulas y tablas de conversión.

Cuadro 1

Interpretaciones de la magnitud del efecto

Magnitud del efecto	% personas del GC debajo del	Posición de la persona en un GC[b]	*p* para adivinar de qué grupo es alguien[c]	r equivalente [d] promedio[a]
d	I_1	I_2	I_3	$I_4 (= r)$
0.0	50	13	.50	.00
0.1	54	12	.52	.05
0.2	58	11	.54	.10
0.3	62	10	.56	.15
0.4	66	9	.58	.20
0.5	69	8	.60	.24
0.6	73	7	.62	.29
0.7	76	6	.64	.33
0.8	79	6	.66	.37
0.9	82	5	.67	.41
1.0	84	4	.69	.45
1.2	88	3	.73	.51
1.4	92	2	.76	.57
1.6	95	1	.79	.62
1.8	96	1	.82	.67
2.0	98	1[e]	.84	.71
2.5	99	1[f]	.89	.78
3.0	99.9	1[g]	.93	.83

Nota. [a]Porcentaje de personas del grupo control quienes podrían estar debajo del promedio de personas.

[b]Posición de la persona en un grupo control de 25 que podría ser equivalente a la persona promedio en el grupo experimental.

[c]Probabilidad en que uno puede adivinar a qué grupo una persona pertenece a partir del conocimiento de su puntaje.

[d]Correlación equivalente, *r* (= Diferencia en porcentaje de "éxito").

[e] de 44 [f] de 160 [g] de 740.

Otra manera de interpretar la magnitud del efecto es compararlos con alguna magnitud del efecto que sea familiar. Por ejemplo, Cohen (1969, p. 23) describe un *ME* de 0.2 como *pequeño* y lo ilustra con un ejemplo: la diferencia entre los pesos de adolescentes de 15 y 16 años de edad en EEUU corresponde a un efecto de este tamaño. Un *ME* de 0.5 lo describe como *medio* y es tan grande como para ser

visto a "simple vista". Un efecto de 0.5 corresponde a la diferencia entre los pesos de dos adolescentes de 14 y 18 años de edad. Cohen describe una *ME* de 0.8 como "bastante perceptible y por lo tanto, grande" y lo iguala a la diferencia entre los pesos de dos niñas de 13 y 18 años. Como ejemplo adicional, Cohen señala que la diferencia entre el CI de un postulante a un Ph.D. y un alumno promedio de universidad en comparable a un *ME* de 0.8.

Cohen reconoce el peligro de los términos *pequeño*, *mediano* y *grande* fuera de contexto. Glass et al. (1981, p.104) son especialmente críticos de este enfoque, argumentando que la efectividad de una intervención particular se puede interpretar solamente en relación con otras intervenciones que buscan producir el mismo efecto. Ellos también señalan que la importancia práctica de un efecto depende enteramente de sus costos y beneficios relativos. En educación, si se pudiera demostrar que al hacer una cambio pequeño y de bajo costo se podría elevar el rendimiento académico en una magnitud del efecto como de 0.1, entonces esto podría ser una mejora significativa, particularmente si la mejora es aplicada uniformemente a todos los estudiantes, y aún más si el efecto fuera acumulativo en el tiempo.

Glass et al. (1981, p. 102) dan el ejemplo que una *ME* de 1 corresponde a la diferencia de alrededor de un año de escolaridad sobre el desempeño en pruebas de rendimiento de alumnos de primaria. Sin embargo, el análisis de una prueba estandarizada de ortografía utilizada en Inglaterra, por ejemplo (Vincent y Crumpler, 1997) sugiere que el incremento en una edad ortográfica, de 11 a 12 corresponde a una *ME* de aproximadamente 0.3, pero parece que esta situación varía de acuerdo al tipo de prueba utilizada.

En la estimación del logro educacional de los alumnos de una nación o región, el estudio de los progresos entre los grados o niveles se puede obtener partiendo de una media y desviación estándar identificada. La cuantificación del monto de progreso muy bien

puede interpretarse en términos de las estimaciones de la magnitud del efecto. Esto es especialmente útil cuando de quieren medir los cambios introducidos por un currículum o sistema educativo que experimentalmente se prueba antes de su difusión.

Finalmente, se puede ayudar a la interpretación de la magnitud del efecto con unos cuantos ejemplos tomados de la literatura actual. El Cuadro 2 lista una selección de aquellos, muchos de los cuales son tomados de Lipsey y Wilson (1993). Los ejemplos citados ilustran el uso de las medidas *ME* y no tienen el propósito de hacer un juicio definitivo de la relativa eficacia de los diferentes métodos de intervención en tales ejemplos. Al interpretarlo, por lo tanto, se debería tener en mente (1) que la mayoría de los meta-análisis desde los cuales se han extraído pueden ser (y frecuentemente lo sido) criticados por una variedad de debilidades, (2) que el rango de circunstancias donde se ha estimado la *ME* puede ser limitado, y (3) que la medida de la magnitud del efecto citado es un promedio basado en valores que varían grandemente.

Perece ser que una característica de las intervenciones educativas es que muy pocos tienen efectos que podrían ser descritos por la clasificación de Cohen como *pequeña*. Parece así por los efectos en el rendimiento de los estudiantes. No hay duda que esto es parcialmente un resultado de la amplia variación encontrada en la población como conjunto, en que la estimación de la magnitud del efecto se ha calculado. También uno podría especular que el rendimiento es más difícil de influenciar que otros resultados, quizás porque la mayoría de las escuelas ya están utilizando estrategias óptimas o porque las diferentes estrategias son efectivas en diferentes situaciones. Pero estas complejidades no están suscritas directamente al simple promedio de la ME, que no refleja tales complejidades.

Cuadro 2

Ejemplos de magnitud del efecto promedio extraídos de investigaciones

Intervención	Resultado	*ME*	Fuente
Reducción de la cantidad de alumnos de 23 a 15	Rendimiento en una prueba de lectura	0.30	Finn y Achilles, (1990)
	Rendimiento en una prueba de matemáticas	0.32	
Cantidad de alumnos pequeña (<30) vs grande	Actitudes de los alumnos	0.47	Smith y Glass (1980)
	Actitudes de los profesores	1.03	
Ubicación de estudiantes: mixto vs agrupados por habilidad	Rendimiento de estudiantes (global)	0.00	Mosteller, Light y Sachs (1996)
	Rendimiento de estudiantes (para los de alto rendimiento)	0.08	
	Rendimiento de estudiantes (para los de bajo rendimiento)	-0.06	
Organización del aula: abierta ("centrada en el niño") vs tradicional	Rendimiento de los estudiantes	-0.06	Giaconia y Hedges (1982)
	Actitudes del estudiante hacia el colegio	0.17	
Educación tradicional vs especial (para niños de primaria con discapacidades)	Rendimiento	0.44	Wang y Baker (1986)
Práctica de resolver exámenes	Puntajes en la prueba	0.32	Kulik, Bangert y Kulik (1984)
Currículum de ciencias tradicional vs basada en investigación	Rendimiento	0.30	Shymansky, Hedges y Woodworth (1990)
Terapia para ansiedad de exámenes (para estudiantes ansiosos)	Ejecución en la prueba	0.42	Hembree (1988)
Retroalimentación a los profesores sobre el desempeño de los estudiantes (estudiantes con PEI)	Rendimiento de los estudiantes	0.70	Fuchs y Fuchs (1986)
Tutoría de pares	Rendimiento de tutelados	0.40	Cohen, Kulik y Kulik, (1982)
	Rendimiento de tutores	0.33	
Instrucción individualizada	Rendimiento	0.10	Bangert, Kulik y Kulik (1983)
Instrucción asistida por computadora	Rendimiento (en todos los estudios)	0.24	Fletcher-Flinn y Gravatt (1995)
	Rendimiento (en estudios bien controlados)	0.02	
Dieta especial	Hiperactividad del niño	0.02	Kavale y Forness (1983)
Entrenamiento en relajación	Síntomas médicos	0.52	Hyman *et al* (1989)
Intervenciones diseñadas para estudiantes en riesgo	Rendimiento	0.63	Slavin y Madden (1989)
Educación en abuso de sustancias basado en el colegio	Uso de sustancias	0.12	Bangert-Drowns (1988)
Programas de tratamientos para delincuentes juveniles	Delincuencia	0.17	Lipsey (1992)

¿Es tan simple como se lo define?

La Ecuación 1 da una definición básica de la *ME* que es conveniente para la mayoría de los propósitos. Sin embargo, hay unas cuantas complicaciones que se necesitan tomar en cuenta.

¿Desviación estándar concentrada o del grupo control?

El primer problema es cuál desviación estándar (*DE*) utilizar. Idealmente, el grupo control nos dará la mejor estimación de la desviación estándar pues consiste en un grupo representativo que la población que no ha sido afectado por la intervención experimental. Sin embargo, a menos que el grupo control sea muy grande, la estimación de la *verdadera* desviación estándar de la población derivada únicamente del grupo control será probablemente menos exacta que una estimación derivada de ambos grupos, la del control y el experimental. Más aún, en estudios donde no hay un verdadero grupo control (GC) (por ejemplo en el experimento de los efectos del día), puede ser arbitraria la decisión de cuál desviación estándar utilizar, y ello producirá una apreciable diferencia en la estimación de la *ME*.

Por estas razones, con frecuencia es mejor usar una estimación estándar "concentrada". La estimación concentrada es esencialmente un promedio de las desviaciones estándar de los grupos experimental y control[6]. La ecuación 2 indica la fórmula para su cálculo:

$$DE_{con} = \sqrt{\frac{(N_E \quad 1)DE_E^2 + (N_C \quad 1)DE_c^2}{N_E + N_C \quad 2}}$$

Ecuación 2

[6] Note que esto no es lo mismo que la desviación estándar de todos los valores de ambos grupos juntos concentradamente. Si, por ejemplo, cada grupo tuviera una baja *DE* pero las dos medias fueran sustancialmente diferentes, la verdadera estimación concentrada (como se calcularía con la ecuación 2) podría ser mucho más baja al concentrar todos los valores juntos y calcular el *DE* correspondiente. Las implicaciones de elegir una u otra desviación estándar son discutidas en Olejnik y Algina (2000).

(N_E y N_C es la cantidad de sujetos en el grupo experimental y control, respectivamente; DE_E y DE_C son sus desviaciones estándar)

El uso de la estimación concentrada de la desviación estándar depende del presupuesto que las dos desviaciones estándar calculadas son estimaciones del mismo valor poblacional. En otras palabras, que la desviación estándar del grupo control y experimental se diferencian solamente como resultado de la variación de muestreo. Cuando este presupuesto no se puede sostener (sea porque hay alguna razón para creer que las dos desviaciones estándar probablemente son diferentes de manera sistemática o si los valores medidos son muy diferentes), entonces la estimación concentrada no se debería usar.

En nuestro ejemplo del experimento de Dowson, sobre la hora del día, las desviaciones estándar para el grupo de la mañana y tarde fueron 4.12 y 2.10 respectivamente. Con $N_E = N_C = 19$, la ecuación 2 da una DE_{con} de 3.3, que fue el valor utilizado en la Ecuación 1 para obtener una *ME* de 0.8 (p. 3). Sin embargo, la diferencia entre las dos desviaciones estándar parece grande en este caso. Dado que una media del grupo de la tarde fue 17.9 de 20, parece verosímil que su desviación estándar puede haber sido reducido por un "efecto de techo", es decir, la distribución de los puntajes se limitó a una valor máximo de 20 solamente. Por lo tanto, en este caso podría ser apropiado utilizar la *DE* del grupo matutino como la mejor estimación disponible. Al hacer esto, se reducirá la *ME* a 0.7, y ello nos conduce a una decisión arbitraria sobre el valor *ME* que se debería utilizar. Una regla en estadística es que cuando dos métodos válidos dan diferentes respuestas, es: "si hay dudas, cita ambos".

Correcciones por Sesgo

Cualquiera que sea la versión de la *DE*, generalmente se la calcula de una muestra de valores disponible (algunas veces una pequeña muestra). Por lo tanto, su cálculo se debería considerar

solamente como una estimación del "verdadero" valor poblacional, y sujeto a un error de muestreo. Aunque se use la desviación estándar concentrada para calcular la *ME*, dando una mejor estimación que la *DE* del grupo de control, desafortunadamente aún habrá un ligero sesgo. En efecto, se puede demostrar que cuando las estimaciones de magnitud del efecto se calculan de muestras tomadas de una población conocida, en promedio los valores muestrales serán un poco más grandes que el valor de la población de la cual provienen. Idealmente, deberíamos corregir este sesgo especialmente si la muestra que tenemos es pequeña.

Hedges y Olkin (1985, p. 80) dan una fórmula con una corrección aproximada:

$$\textit{Estimación insesgada de } d \quad \textit{Valor calculado de } d \times \; 1 \; \frac{3}{(4(N_E + N_C) \;\; 9)}$$

Ecuación 3

En el experimento de Dowson, con 38 valores, el factor de corrección será 0.98, lo que evidencia una mínima diferencia, reduciendo la *ME* de 0.82 a 0.80.

¿Cuál es la relación entre "magnitud del efecto" y "significancia"?

La magnitud del efecto cuantifica el tamaño de la diferencia entre dos grupos, y por lo tanto se puede decir que es una verdadera medida de la significancia de tal diferencia. Para los resultados del experimento de Dowson, podríamos hacer la siguiente pregunta "¿Cuánto podría ser la diferencia en el aprendizaje de los niños si se les enseñara un tema particular en la tarde en vez de la mañana?". La mejor respuesta que podríamos dar vendría en términos de la *ME*.

Sin embargo, en el mundo de la estadística la palabra "significancia" se usa frecuentemente para aludir a la "significancia estadística", que es la probabilidad en que una diferencia entre dos grupos podría ser únicamente por un error de muestreo. Si tomamos dos muestras de la misma población siempre habrá una diferencia entre ellas. De manera usual, la significancia estadística se calcula como un valor *p*, es decir, la probabilidad de que una diferencia de al menos el mismo tamaño podría surgir por azar, aún si no hubiera diferencias entre las dos poblaciones. Para las diferencias entre dos grupos, este valor p se podría calcular normalmente por un procedimiento estadístico conocido como prueba *t*. Por convención, si $p < .05$ (es decir, debajo del 5%), la diferencia es considerada como "significativa"; si no, se ve como "no significativa".

Pero hay un número de problemas al usar las pruebas de significancia de esta forma. El principal es que el valor *p* depende esencialmente de dos aspectos: la magnitud del efecto *y* el tamaño de la muestra. Un investigador podría obtener un resultados significativo ya sea si el efecto sea muy grande (a pesar de tener solo una muestra pequeña) o que la muestra sea muy grande (aún si la magnitud del efecto hallada es pequeña). Es importante saber que la significancia estadística de un resultado, ya que sin ello hay el peligro de tener una firma conclusión proveniente de estudios donde la muestra es bastante pequeña como para justificar tal confianza. Sin embargo, la significancia estadística no nos dice lo más importante: la *ME*. Una manera de superar esta confusión es reportar la magnitud del efecto junto con una estimación de su "margen de error" probable o "intervalo de confianza".

¿Cuál es el margen de error de la estimación de la magnitud del efecto?

Claramente, si una *ME* se calcula desde una muestra muy grande, es probable que sea más exacta que una *ME* calculada de una pequeña muestra. Este *margen de error* puede ser cuantificado

cogiendo la idea de *intervalo de confianza*, que da la misma información proveniente de una prueba de significancia: usar un "intervalo de confianza del 95%" es equivalente a asumir un "nivel de significancia del 5%". Para calcular un intervalo del 95%, uno debe asumir que el valor que se usará (por ejemplo, la estimación de la magnitud del efecto de 0.8) es el valor "verdadero", y que se calculará el monto de variación de esta estimación si repetidamente se tomaran nuevas muestras de la misma cantidad (es decir, 38 niños). Para cada 100 de estas hipotéticas nuevas muestras, 95 de ellas podría dar una estimación de la *ME* dentro del intervalo de confianza del 95%. Si este intervalo de confianza incluye al cero, entonces es lo mismo decir que el resultado no es estadísticamente significativo. Si, por otro lado, el cero no está dentro de tal intervalo, entonces se tiene un nivel estadísticamente significativo del 5%. El uso del intervalo de confianza es una mejor modo de llevar esta información, pues mantiene el énfasis en la *ME* –que es una información importante– más que en el valor *p*.

Una fórmula para calcular el intervalo de confianza proviene de Hedges y Olkin (1985, p. 86). Si la estimación de la *ME* de la muestra es *d*, ello se distribuye normalmente, con una desviación estándar igual a:

$$\sigma[d] = \sqrt{\frac{N_e + N_c}{N_e . N_c} + 2\left(\frac{d^2}{N_e . N_c}\right)}$$

Ecuación 4

(N_E y N_C son la cantidad de sujetos en el grupo experimental y control, respectivamente)

De aquí, un intervalo de confianza del 95%[7] para *d* estaría desde

[7] El valor "1.96" en esta fórmula proviene del valor crítico a dos colas de la distribución normal. Otros valores provenientes de esta misma distribución pueden ser sustituidos para obtener diferentes intervalos de confianza.

$$d - 1.96 * F[d] \qquad \text{hasta} \qquad d + 1.96 * F\,[d]$$

Ecuación 5

Tomando los resultados del experimento de la hora del día, $N_E = N_C = 19$ y $d = 0.8$, entonces, F $[d]$ = r(0.105 + 0.008) = 0.34. Entonces, el intervalo de confianza al 95% es [0.14, 1.46]. Esto se podría interpretar[8] como el verdadero efecto de la hora del día está muy probablemente entre 0.14 y 1.46. en otras palabras, ocurre realmente un efecto (la tarde es mayor que la mañana) y la diferencia es bastante grande.

¿Se puede combinar la información sobre la magnitud del efecto?

Una de las principales ventajas del uso de la *ME* es que cuando un experimento en particular ha sido replicado, las diferentes *ME* de cada estudio pueden ser combinados para dar una mejor estimación global del monto del efecto. Este proceso que sintetiza los resultados experimentales en una simple estimación de la *ME* es conocido como *meta-análisis*. Este método fue desarrollado en su forma actual por un estadístico educacional, Gene Glass[9] y ahora es ampliamente utilizado no solamente en educación, sino también en medicina y en las ciencias sociales. Una introducción breve y accesible del meta-análisis se puede encontrar en Fitz-Gibbon (1984).

El meta-análisis, sin embargo, puede hacer mucho más que simplemente dar una *ME* promedio general, aunque tal cosa sea impor-

[8] En efecto, esta interpretación solamente se justifica con la incorporación de algo adicional, como la frase "siendo lo demás igual" y un enfoque Bayesiano que la mayoría de los estadísticos explícitamente usan como regla. Sin embargo, la interpretación de lo que precisamente proporciona las pruebas de significancia es controversial y a menudo confuso. Ver Oakes (1986) para una iluminada discusión de este aspecto.

[9] Ver Glass, McGaw y Smith, 1981. Las raíces del meta-análisis pueden ser rastreadas en la literatura, por ejemplo, en Lepper et al. (1999).

tante. Si, para una intervención particular, algunos estudios producen grandes efecto y otros producen pequeños efectos, podría ser de limitado valor hacer una simple combinación de ellos y decir que el efecto promedio es *moderado*. Mucho más útil podría ser examinar los estudios originales para ver las diferencias entre aquellas con grandes y pequeños efectos , y tratar de comprender qué factores podrían haber originado las diferencias. El mejor meta-análisis, por lo tanto, implica buscar las *ME* y características de la intervención, el contexto y diseño de estudio de donde provienen[10].

Las importancia de la replicación para obtener evidencia sobre lo que funciona, no puede ser nunca sobrestimada. En el experimento de Dowson, se encontró un gran efecto educacional y estadísticamente significativo. Debido que sabemos que los estudiantes fueron puestos aleatoriamente en cada grupo, podemos estar confiados que las diferencias por azar iniciales entre los dos grupos son factores con mucha improbabilidad de haber influenciado en las diferencias significativas. Además, el pretest de ambos grupo antes de la intervención lo hace aún menos probable. Pero uno no puede controlar la posibilidad de que las diferencias halladas provengan de algunas características peculiares a los niños en este experimento en particular. Por ejemplo, si ninguno de ellos pudo desayunar ese día, esto podría haber originado el pobre desempeño del grupo de la mañana. El resultado podría no ser presumiblemente generalizado a la amplia población de estudiantes, en que la mayoría que tiene un desayuno servido. Alternativamente, el efecto podría depender de la edad de los estudiantes. Los estudiantes de Dowson tuvieron una edad de 7 u 8; es bastante posible que el efecto podría disminuir o invertirse con alumnos de más (o menos) edad. Esto ilustra el peligro de implementar alguna política sobre la base de un simple experimento. La confianza en la generalidad de un resultado solo puede se puede seguir mediante replicaciones.

[10] Rubin (1992). Ver también Lepper et al. (1999) para una discusión de los problema que ocurren y algunas otras limitaciones de la aplicabilidad del meta-análisis.

Una importante consecuencia de la capacidad del meta-análisis para combinar resultados es que aún pequeños estudios pueden hacer una contribución importante a la suma de conocimientos. El tipo de experimento que un profesor de escuela puede realizar podría haberse efectuado con solo 30 estudiantes. A menos que el efecto sea grande, un estudio de esta magnitud es bastante improbable que produzca resultados significativamente importantes. De acuerdo con los conocimientos estadísticos convencionales, tal experimento no tendría valor. Sin embargo, si los resultados de varios experimentos se combinan con el meta-análisis, los resultados globales pueden ser probablemente altamente significativos. Más aún, tendrán la fortaleza de ser obtenidos desde varios contextos (lo que incrementa su confianza en su generalización) y desde la práctica de trabajar en contextos de la vida real (y por lo tanto, hace que las políticas educativas se basen en estas experiencias y pueden ser implementadas auténticamente).

Debemos hacer una advertencia final sobre el riesgo de combinar resultados inconmensurables. Dados dos (o más) números, uno siempre puede calcular un promedio. Pero, si hay *ME* de experimentos que difieren de manera importante en términos de las mediciones de sus resultados, los resultados pueden ser menos significativos. Puede ser muy tentador que, una vez que hemos calculado la magnitud del efecto, los tratemos todos de la misma manera y perdamos de vista sus orígenes. Ciertamente, hay abundantes ejemplos en el meta-análisis en los que la yuxtaposición de *ME* son cuestionables.

Al comparar (o combinar) *ME*, uno debería considerar cuidadosamente si ellos se refieren a los mismos resultados. Este consejo no se aplica solamente al meta-análisis, sino también a alguna comparación de *ME*. Además, debido a la sensibilidad de las estimaciones del *ME* a aspectos como la confiabilidad y la restricción del rango, también deberíamos preocuparnos si las medidas de resultados se derivan de los mismos (o al menos suficientemente similares) instrumentos y de las mismas (o al menos suficientemente similares) poblaciones.

También es importante dar una mirada crítica a los términos o conceptos utilizados en los tratamientos para crear las diferencias que fueron medidas. En la literatura educacional, frecuentemente los mismos nombres se ponen a intervenciones que son diferentes; por ejemplo, hay situaciones en que la definición es distinta o, simplemente, no están bien definidos como para quedar claro si se refieren a los mismos conceptos. También podría suceder que diferentes estudios han desarrollado tratamientos bien operacionalizados y bien definidos pero que se implementaron de modo diferente; o que el mismo tratamiento pueda haber tenido diferentes niveles de intensidad en diferentes estudios. En cualquiera de estos casos, parece que no tiene sentido promediar sus efectos.

¿Qué otros factores afectan la magnitud de efecto?

Aunque la *ME* es una medida simple y de sencilla interpretación sobre la efectividad de una intervención, es sensible a varias influencias espurias, y por ello de debe estar alerta.

Restricción del Rango

Supongamos que el experimento que venimos citando (efecto de la hora del día), fuera repetido con los mejores alumnos del colegio y con, nuevamente, un grupo de habilidad mixta. Si los estudiantes fueran asignados aleatoriamente a los grupos de la mañana y tarde, las respectivas diferencias podrían ser las mismas en cada caso; las medias en los de mejor habilidad pueden haber sido altas o las diferencias similares como en la situación original. Sin embargo, es improbable que las desviaciones estándares sean las mismas. El grupo superior del colegio es un grupo bastante selectivo y la distribución de puntajes en ellos podría ser mucho menos que en el grupo combinado de la población, por ejemplo, una clase con grupos variables en habilidad. Esta situación, de hecho, podría tener un importante impacto en el cálculo de la magnitud del efecto. Con una elevada res-

tricción del rango proveniente del grupo superior, la magnitud del efecto puede ser más grande que lo que se encontraría en un grupo con una distribución de puntajes fuera *normal*.

Idealmente, para calcular la *ME* uno debería usar la desviación estándar de la población completa, para que las comparaciones sean equitativas. Sin embargo, habrán muchos casos en que valores no restringidos no son posibles, sea en la práctica o en principio. Por ejemplo, al considerar el efecto de una intervención a estudiantes universitarios o a alumnos con dificultades en la lectura, uno debe recordar que estos son poblaciones restringidas. Al reportar la *ME*, uno debería llevar su atención a este hecho; si el monto de restricción puede ser cuantificado, entonces puede ser posible hacer concesiones. Cualquier comparación con *ME* calculadas de una población de rango completo se debe hacer con gran precaución.

Distribuciones No-Normales

Las interpretaciones de la *ME* en el Cuadro 1 depende del presupuesto de que el grupo experimental y control tienen una distribución normal, es decir, la familiar curva simétrica mostrada, por ejemplo, en la Figura 1. No es necesario decir que si este presupuesto no es verdad, la interpretación se alterará y, en particular, puede ser difícil hacer una imparcial y justa comparación entre una magnitud del efecto basada sobre distribuciones normales y otra basada en distribuciones no normales.

Una ilustración de lo que decimos se muestra en la Figura 2, en que se ve las curvas de frecuencia para dos distribuciones, una de ellas normal y la otra similar en forma en la parte central de su rango, pero con extremos prolongados. En efecto, la última parece un poco más dispersa que la distribución normal, pero su desviación estándar es realmente 50% más grande. La consecuencia de esto en cuanto a las diferencias de la *ME* se muestra en la Figura. Ambos gráficos muestran distribuciones que difieren por una *ME* igual a 1, pero el resultado gráfico de la es más bien disímil. En el gráfico (b),

la separación entre el grupo experimental y control parece mucho mayor, aún cuando la *ME* es la misma para las distribuciones normales en el gráfico (a). En términos del grado de traslapamiento, en el gráfico (b) el 94% del grupo *experimental* está arriba de la media del grupo control, comparado con el valor de 84% de la distribución normal del gráfico (a) (como ocurre en el Cuadro 1). Esto es una diferencia bastante sustancial e ilustra el peligro de utilizar los valores de la Cuadro 1 cuando la distribución no es normal.

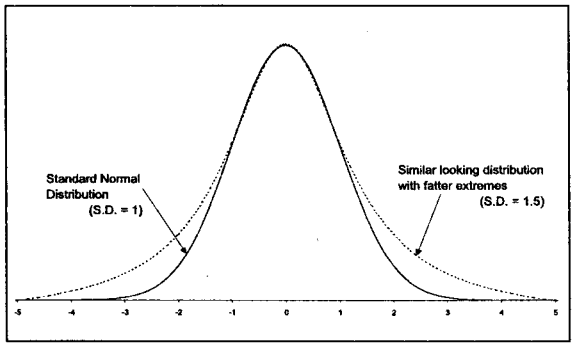

Figura 2. Comparación de distribuciones Normal y No Normal.

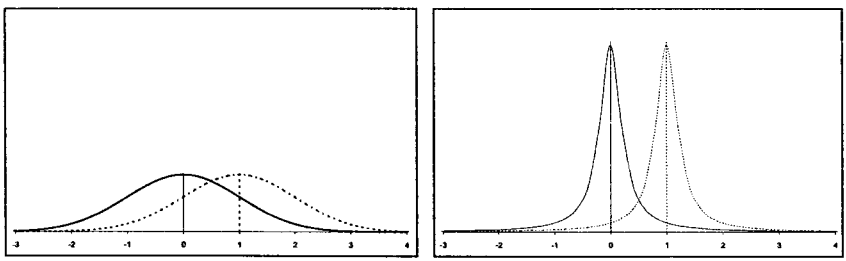

Figura 3. Distibución Normal y No Normal con magnitud del efecto = 1.

Confiabilidad de la Medición

Un tercer factor que espuriamente puede afectar un *ME* es la confiabilidad de la medición sobre el cual está basada. La confiabilidad se refiere a la exactitud, la estabilidad y robustez de una medición, y frecuentemente se la define en función del grado en que dos mediciones del mismo atributo (o una medición repetida) producen el mismo resultado[11]. En otras palabras, cualquier medida de un particular resultado puede ser considerada como un subyacente valor *verdadero* junto con un componente de error. El problema es que el monto de variación en los puntajes obtenidos para una particular muestra (es decir, su desviación estándar) dependerá de la variación en los puntajes subyacentes y del monto de error en su medición.

Para dar un ejemplo, imagine el experimento de la "hora del día" que fue conducido con dos muestras hipotéticamente idénticas de estudiantes. En la primera versión, la prueba evaluó la comprensión con solo 10 ítemes y sus puntajes fueron convertidos en un porcentaje. En una segunda versión, se utilizó una prueba con 50 ítemes y sus puntajes se convirtieron nuevamente en un porcentaje. Las dos pruebas fueron de igual dificultad y el actual efecto de la diferencia en la hora del día fue la misma en cada caso, así que los respectivos porcentajes promedios del grupo la mañana y de la tarde fueron los mismos para ambas versiones. Sin embargo, siempre se da al caso que la prueba más larga será más confiable y la desviación estándar de los porcentajes de la prueba de 50 ítemes será más bajo que la desviación estándar de la prueba de 10 ítemes. De esta forma, aunque el verdadero efecto fue el mismo, el *ME* calculado será diferente.

En la interpretación de *ME*, es importante, por lo tanto, conocer la confiabilidad de la medición desde cual ha provenido su cálculo.

[11] Para una más detallada definición y discusión del concepto de confiabilidad, ver Nunnally y Bernstein (1995).

Esto es una razón de por qué la confiabilidad de alguna medición se debe reportar. Teóricamente es posible hacer una corrección por no-confiabilidad (llamado algunas veces *atenuación*), que nos da una estimación del *ME* que podría haberse obtenido si la confiabilidad hubiese sido mejor. Sin embargo, en la práctica el efecto de esto es de extrema precaución pues mientras peor fue la prueba, más se incrementará la estimación del *ME*.

¿Hay medidas de magnitud del efecto alternativas?

Se han propuesto otras técnicas estadísticas como medidas de la magnitud del efecto alternativas, diferentes a la diferencia media estandarizada. Se considerarán algunas de ellas.

Proporción de la varianza explicada

Si la correlación entre dos variables es r, el cuadrado de este valor (frecuentemente denotado con la letra R^2) representa la proporción de la varianza que es "explicada por" la otra variable. En otras palabras, es la proporción por el que la varianza de los residuales de una ecuación de regresión. Se puede extender esta idea a la regresión múltiple (donde representa la proporción de la varianza explicada por todas las variables independientes juntas) y tiene una analogía cercana en ANOVA (donde regularmente se le llama *eta cuadrado*, 0^2). El cálculo de r (y de aquí R^2) para el tipo de situación experimental que hemos estado considerando, ya ha sido referido anteriormente.

Debido que R^2 (u otras medidas alternativas de varianza explicada) tiene esta flexibilidad de conversión, algunas veces se lo considera como una medida universal de *ME* (Thompson, 1999). Una desventaja de este enfoque es que las medidas de *ME* basadas en la varianza explicada sufren algunas limitaciones técnicas, como su sensibilidad a la violación de presupuestos (heterogeneidad de la varianza, diseños balanceados) y sus errores estándar pueden ser grandes

(Olejnik y Algina, 2000). En general, también pueden ser más complejos estadísticamente y por lo mismo menos fácilmente comprendidos. Sin embargo, hay una objeción más fundamental para el uso de lo que es esencialmente una medida de asociación para indicar la *fuerza de un efecto*. Expresar diferentes medidas desde el punto de vista de la misma estadística puede ocultar importantes diferencias entre ellos; en efecto, estas *medidas del efecto* diferentes son fundamentalmente diferentes y no se deberían confundir.

La crucial diferencia entre un *ME* calculado desde un experimento y otro calculado desde una correlación está en la naturaleza causal de lo que se afirma con ello. Además, la palabra *efecto* tiene una inherente implicación de causalidad: hablar de "el efecto de A sobre B" sugiere una relación causal más que solo una asociación. Desgraciadamente, sin embargo, la palabra *efecto* es frecuentemente utilizada cuando no se hace una afirmación causal explícita, sino que su implicación es algunas veces permitida que salga a flote dentro y fuera del significado, aprovechando la ventaja de la ambigüedad de sugerir un vínculo causal donde realmente nada se justifica para ello.

Este tipo de confusión es tan extendido en educación que aquí recomendamos que la palabra *efecto* (y por lo tanto *magnitud del efecto*) no se debería usar a menos que una afirmación deliberada y explícita se esté haciendo. Cuando tal afirmación no se hace por ningún lado, podemos hablar de la "varianza explicada por" (R^2) o *la fuerza de asociación* (r), o simplemente –y quizás más informativamente– solo citar el coeficiente de regresión (Tukey, 1969). Si se está haciendo una afirmación de tipo causal, se debería hacerlo explícito junto con una justificación de ello. Por sí mismo, r sólo indica el grado de asociación pero no señala causalidad, y no es una información suficiente para inferir que A causa B (Brown, 1980).

Otras medidas de la Magnitud del Efecto

Se ha demostrado que la interpretación de la *diferencia media estandarizada* como medida de la magnitud del efecto es muy sensi-

ble a las violaciones del presupuesto de normalidad. Por esta razón, ha aparecido un número de alternativas más robustas. Un ejemplo de estos se encuentran en Cliff (1993). También hay medidas de *ME* para resultados multivariados; una detallada explicación se puede encontrar en Olejnik y Algina (2000).

Finalmente, un índice común de *ME* ampliamente utilizado en medicina es el "odds ratio". Este es apropiado cuando un resultado es dicotómico, como del tipo positivo o negativo. Las explicaciones referidas a este índice se pueden hallar en varios textos de estadísticas aplicada a la medicina, como Altman (1991).

En la línea de la estadística no paramétrica, la prueba *U* de Mann-Whitney ha sido útil también para estimar la magnitud del efecto en situaciones en que la medición de resultados se describe como una variable ordinal (Grissom, 1994a , 1994b). En situaciones en que, además de la naturaleza ordinal de los datos, los supuestos estadísticos no se cumplen (normalidad, homocedastidad), medidas de *ME* basadas en estadísticas no paramétricos parece atrayente. Aún cuando estadísticos no paramétricos como la U de Mann-Whitney no muestran robustez ante la heterocedastidad (Zimmerman y Zumbo, 1993), esta debilidad es compensada por la medida de ME alternativas, por ejemplo la propuesta por Grissom (Grissom, 2001).

Una medida de la magnitud del efecto, menos conocida, en términos de la fuerza de asociación y expresada como correlación por rangos también ha aparecido en el contexto de las pruebas no paramétricas. Esta es la presentada por May, Masson y Hunter (1990), a la que identificaron como coeficiente de correlación biserial por rangos de Glass.

Otra índice cuyo uso lo ha vuelto popular en la literatura de las ciencias conductuales es el eta cuadrado (0^2) y omega cuadrado (T^2); generalmente, el primero es el más frecuentemente reportado en el

análisis de varianza y es interpretado en términos del monto de variabilidad explicada (Cooksey, 1997).

Finalmente, la elección de medidas alternativas para describir la magnitud del efecto en el estudio en mano, o aún aquella descrita con amplitud en el presente artículo, debe estar apoyada por el buen juicio del investigador después de obtener la información pertinente para su uso e interpretación.

Conclusiones

Proponemos algunos consejos sobre el uso de las medidas de *ME*:

- La *ME* es una medida libre de escala, estandarizada del monto relativo del efecto de una intervención. Es particularmente útil para cuantificar los efectos medidos en una escala arbitraria o poco familiar, y para comparar los relativos montos de efectos provenientes de diferentes estudios.
- La interpretación de la *ME* depende generalmente del presupuesto de que los valores del grupo *experimental* y *control* están normalmente distribuidos y tienen las mismas desviaciones estándar. Se lo puede interpretar en términos de percentiles o ranks en el que dos distribuciones se traslapan, en términos de la probabilidad de identificar la fuente de donde procede un valor o con referencia a efectos o resultados conocidos.
- Utilizar una *ME* con un intervalo de confianza conduce a la misma información que una prueba estadística de significancia pero con el énfasis puesto en el efecto más que en el tamaño de la muestra.
- Se debería calcular y reportar la *ME* (con referencia a los intervalos de confianza) en estudios primarios tanto como en los meta-análisis.
- La interpretación de la *ME* estandarizado puede ser problemática cuando la muestra tiene un rango restringido o no proviene de

una distribución normal, o no se conoce la confiabilidad la medición de la cual se ha derivado.

- El uso de la diferencia de media *no estandarizada* (es decir, la diferencia simple entre dos grupos, junto con un intervalo de confianza) puede ser preferible cuando:
  - el resultado es medido en una escala familiar o conocida
  - la muestra tiene un rango restringido
  - la población de referencia es significativamente no normal
  - los grupos experimental y control poseen una apreciable diferencia en sus desviaciones estándar.
  - la escala en que se miden los resultados tiene una baja o desconocida confiabilidad

- Se debe ser precavido al comparar o agregar magnitudes del efecto basados en diferentes resultados, diferentes operacionalizaciones del mismo resultado, diferentes tratamientos o niveles del mismo tratamiento, o mediciones derivadas de diferentes poblaciones.
- La palabra *efecto* lleva la implicación de causalidad, y no se debería utilizar la expresión *magnitud del efecto* en tal sentido a menos que esta implicación sea el propósito del estudio y pueda ser justificada.

Los lectores interesados pueden contactar electrónicamente a los autores para obtener un archivo, en MS Excel, que facilita los cálculos de la magnitud del efecto descrita en este artículo.

Referencias

Altman, D. G. (1991). *Practical Statistics for Medical Research.* Londres: Chapman and Hall.

Brown, F. G. (1980). *Principios de la medición en Psicología y educación.* México, DF: El Manual Moderno.

Bangert, R. L., Kulik, J. A. y Kulik, C.C. (1983). Individualised systems of instruction in secondary schools. *Review of Educational Research, 53*, 143-158.

Bangert-Drowns, R. L. (1988). The effects of school-based substance abuse education: a meta-analysis. *Journal of Drug Education, 18*, 3, 243-65.

Cliff, N. (1993). Dominance Statistics – ordinal analyses to answer ordinal questions *Psychological Bulletin, 114*, 3. 494-509.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences.* NuevaYork: Academic Press.

Cohen, P. A., Kulik, J. A. y Kulik, C. C. (1982). Educational outcomes of tutoring: a meta-analysis of findings. *American Educational Research Journal, 19*, 237-248.

Cooksey, R. W. (1997). *Statistics for behavioural and social research: A descriptive handbook.* Armidale NSW, Australia: University of New England.

Dowson V. (2000). Time of day effects in school-children's immediate and delayed recall of meaningful material. *TERSE Report.* Documento en línea: http://www.cem.dur.ac.uk/ ebeuk/research/terse/library. htm.

Finn, J. D. y Achilles, C. M. (1990) Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27*, 557-577.

Fitz-Gibbon C. T. (1984). Meta-analysis: an explication. *British Educational Research Journal, 10*, 2, 135-144.

Fletcher-Flinn, C. M. y Gravatt, B. (1995). The efficacy of Computer Assisted Instruction (CAI): a meta-analysis. *Journal of Educational Computing Research, 12*(3), 219-242.

Fuchs, L. S. y Fuchs, D. (1986). Effects of systematic formative evaluation: a meta-analysis. *Exceptional Children, 53*, 199-208.

Giaconia, R. M. y Hedges, L. V. (1982). Identifying features of effective open education. *Review of Educational Research, 52*, 579-602.

Glass, G. V., McGaw, B. y Smith, M. L. (1981). *Meta-Analysis in Social Research.* Londres: Sage.

Grissom, R. J. (1994). Probability of the superior outcome of the one treatment over another. *Journal of Applied Psychology, 79*(2), 314-316.

Grissom, R. J. (1994b). Statistical análisis of ordinal categorical status after therapies. *Journal of Consulting and Clinical Psychology, 62*(2), 281-284.

Grissom, R. J. (2001). Review of assumptions and problems in the appropriate conceptualisation of effect size. *Psychology Methods, 6*(2), 135-146.

Hedges, L. y Olkin, I. (1985). *Statistical methods for meta-analysis.* Nueva York: Academic Press.

Hembree, R. (1988). Correlates, causes effects and treatment of test anxiety. *Review of Educational Research, 58*(1), 47-77.

Hyman, R. B, Feldman, H. R., Harris, R. B., Levin, R. F. y Malloy, G. B. (1989). The effects of relaxation training on medical symptoms: a meta-analysis. *Nursing Research, 38*, 216-220.

Kavale, K. A. y Forness, S. R. (1983). Hyperactivity and diet treatment: a meta-analysis of the Feingold hypothesis. *Journal of Learning Disabilities, 16*, 324-330.

Kerlinger, F. N. (1986). *Foundations of behavioral research.* Nueva York: Holt, Rinehart and Winston.

Kulik, J. A., Kulik, C. C. y Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Education Research Journal, 21*, 435-447.

Lepper, M. R., Henderlong, J. y Gingras, I. (1999). Understanding the effects of extrinsic rewards on intrinsic motivation - Uses and abuses of meta-analysis: Comment on Deci, Koestner, and Ryan. *Psychological Bulletin, 125*, 6, 669-676.

Lipsey, M. W. (1992). Juvenile delinquency treatment: a meta-analytic inquiry into the variability of effects. En T. D. Cook, H. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis y F. Mosteller (Eds.), *Meta-analysis for explanation* (pp. 83-127). Nueva York: Sage.

Lipsey, M. W. y Wilson, D. B. (1993). The efficacy of psychological,

educational, and behavioral treatment: confirmation from meta-analysis. *American Psychologist, 48*, 12, 1181-1209.

May, R. B., Masson, M. E. J. y Hunter, M. A. (1990). *Application of statistics in behavioural research*. NuevaYork: Harper & Row.

Mosteller, F., Light, R. J. y Sachs, J. A. (1996). Sustained inquiry in education: lessons from skill grouping and class size. *Harvard Educational Review*, 66, 797-842.

Nunnally, J. C. y Bernstein, I. J. (1995). *Teoría psicométrica* (3ra ed.). México, DF: McGraw-Hill.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. Nueva York: Wiley.

Olejnik, S. y Algina, J. (2000). Measures of effect size for comparative studies: applications, interpretations and limitations. *Contemporary Educational Psychology*, 25, 241-286.

Rosenthal, R. y Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166-169.

Rubin, D. B. (1992). Meta-analysis: literature synthesis or effect-size surface estimation. *Journal of Educational Statistics, 17*, 4, 363-374.

Thompson, B. (1999). *Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap*. Documento invitado para la reunión anual del American Educational Research Association, Montreal.

Shymansky, J. A., Hedges, L. V. y Woodworth, G. (1990). A reassessment of the effects of inquiry-based science curricula of the 60s on student performance. *Journal of Research in Science Teaching, 27*, 127-144.

Slavin, R. E. y Madden, N. A. (1989). What works for students at risk? A research synthesis. *Educational Leadership, 46*(4), 4-13.

Smith, M. L. y Glass, G. V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal, 17*, 419-433.

Vincent, D. y Crumpler, M. (1997). *British Spelling Test Series Manual 3X/Y.* Windsor: NFER-Nelson.

Wang, M. C. y Baker, E. T. (1986). Mainstreaming programs: Design features and effects. *Journal of Special Education, 19,* 503-523.

Zimmerman, D. M. y Zumbo, B. D. (1993). Rank transformations and the power of the Student t test and Welch t test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology, 47,* 523-539.