RiiTE
N.º 13 Diciembre 2022 pp. 5-40 ISSN: 2529-9638 DOI: https://doi.org/10.6018/riite.545451

Revisiones sistemáticas y meta-análisis en Educación: un tutorial

Systematic reviews and meta-analyses in Education: A tutorial

Julio Sánchez-Meca

Universidad de Murcia (ESpaña)
jsmeca@um.es

Recibido: 02/11/2022 Aceptado: 23/11/2022 Publicado:

RESUMEN

Las revisiones sistemáticas (RSs) y los meta-análisis (MAs) constituyen una metodología consolidada en las Ciencias Sociales y de la Salud. Su propósito es sintetizar los resultados de estudios empíricos para dar respuesta a alguna pregunta de interés. Tomando como base una revisión comprehensiva de la literatura sobre RSs y MAs, se presenta en este artículo un tutorial sobre cómo hacer este tipo de investigaciones. Para ello, se describe su desarrollo siguiendo siete etapas: (1) formulación de la pregunta de interés; (2) definición de los criterios de selección de los estudios; (3) búsqueda de los estudios mediante el uso de fuentes formales e informales; (4) extracción de las características de los estudios; (5) definición del resultado de los estudios, haciendo hincapié en los índices del tamaño del efecto (ej., familia *d* y familia *r*); (6) métodos de síntesis, distinguiendo entre síntesis meta-analítica y otros métodos de síntesis, y (7) publicación o redacción de la RS/MA. También se presentan recomendaciones sobre cómo hacer lectura crítica de RSs/MAs hechas por otros y se presentan checklists y guías orientativas sobre cómo redactarlas, tales como los checklists PRISMA, AMSTAR-2, MOOSE o REGEMA, entre otros. Finalmente, se discuten las ventajas

y las limitaciones de las RSs/MAs y se alcanzan algunas reflexiones finales, centrando la atención en la importancia de valorar posibles sesgos en los resultados de este tipo de investigación.

PALABRAS CLAVE

Meta-análisis; revisión sistemática; Educación Basada en la Evidencia; tamaño del efecto.

ABSTRACT

Systematic reviews (SRs) and meta-analyses (Mas) have become a consolidated methodology in Social and Health Sciences. Their purpose is to synthesize the results of empirical studies in order to offer an answer to a given question of interest. Based on a comprehensive review of the literature on SRs and Mas, in this article is presented a tutorial on how to conduct this kind of research. With this purpose, the development of a SR/MA is presented following seven phases: (1) formulating the question of interest; (2) defining the selection criteria of the studies; (3) searching for the studies by using formal and informal search sources; (4) extracting study characteristics; (5) defining the study results relevant for the question of interest and emphasizing effect size indices (ej., *d* family and *r* family); (6) synthesis methods, distinguishing

between meta-analytic synthesis and other synthesis methods, and (7) publishing or writing the SR/MA. In addition, recommendations on how to make critical reading of SRs/Mas conducted by other researchers are described, as well as checklists and guidelines on how to write them, such as PRISMA, AMSTAR-2, MOOSE, or REGEMA, among others. Finally, the advantages and limitations of SRs/Mas are discussed and some final reflections are stated, specially focusing on the need of making critical appraisal for potential biases in the results of this kind of research.

Julio Sánchez-Meca
RiiTE, Núm. 13 (2022), 5-40

KEYWORDS

Meta-analysis; systematic review; Evidence-Based Education; effect size.

CITA RECOMENDADA:

Sánchez-Meca, J. (2022). Revisiones sistemáticas y meta-análisis en Educación: Un tutorial. *RiiTE Revista interuniversitaria de investigación en Tecnología Educativa, 13*, 5-40. https://doi.org/10.6018/riite.545451

Principales aportaciones del artículo y futuras líneas de investigación:

- Las RSs/MAs son una metodología de investigación consolidada en Educación.
- La realización de una RS/MA siguen una serie de etapas rigurosas que le confieren carácter científico.
- El correcto reporte y lectura crítica de RSs/MAs se beneficia de la aplicación de numerosos chekclists. propuestos en la literatura.
- La posible existencia de sesgos en los resultados de una RS/MA debe analizarse de forma rigurosa.

1. INTRODUCCIÓN

Las revisiones sistemáticas (RSs) y los meta-análisis (MAs) se han convertido en una metodología imprescindible para la correcta acumulación del conocimiento científico en cualquier disciplina. Las Ciencias de la Educación no han sido ajenas al auge que esta metodología ha experimentado en los últimos 20 años, de forma que actualmente es habitual encontrar MAs publicados en revistas del ámbito educativo.

Antes de la existencia de los MAs, las revisiones de la literatura científica se llevaban a cabo de forma narrativa, subjetiva y sin aplicar una metodología científica. La consecuencia de esta deficiencia fue una pobre acumulación del conocimiento científico, de forma que era habitual encontrar revisiones

narrativas firmadas por diferentes autores que llegaban a conclusiones divergentes (Hunt, 1997). Los MAs surgen a principios de la década de 1980 como una metodología capaz de dotar de rigor científico al proceso de revisión de la investigación, posibilitando una mejor acumulación del conocimiento.

En el ámbito de la Educación el primer MA que se llevó a cabo fue el de Glass y Smith (1978) sobre la relación existente entre el tamaño de la clase y el rendimiento académico de los estudiantes. Desde entonces, se ha publicado una miríada de MAs y RSs en este ámbito. Fue precisamente Gene V. Glass (1976) quien acuñó el término 'meta-análisis' (meta-analysis) para referirse a este método de investigación como "el análisis estadístico de una gran colección de resultados de trabajos individuales con el propósito de integrar los hallazgos obtenidos" (p. 7). Y fue este autor quien publicó el primer libro sobre MA (Glass, McGaw y Smith, 1981).

1.1. Revisiones sistemáticas y meta-análisis: concepto

Una RS es un tipo de investigación en el que se formula una pregunta con toda claridad y, para responder a la misma, se lleva a cabo una búsqueda comprehensiva de los estudios empíricos que han tratado de dar respuesta a dicha pregunta, se seleccionan los estudios, se extrae la

información relevante de los mismos y se lleva a cabo una síntesis de los resultados. Mediante la síntesis acumulativa de los resultados de los estudios se alcanza una respuesta a la pregunta inicialmente formulada (Cooper, 2016; Petticrew y Roberts, 2006). Si en el proceso de síntesis de los estudios los resultados de los mismos son cuantificados mediante el cálculo de índices del tamaño del efecto y se aplican técnicas de análisis estadístico sobre éstos, entonces la RS se convierte en un MA. Es

decir, un MA es un tipo de RS en el que los tamaños del efecto de los estudios se analizan estadísticamente para obtener una estimación conjunta del efecto medio y comprobar la consistencia u homogeneidad de dichos tamaños del efecto. En el caso de que los tamaños del efecto de los estudios exhiban heterogeneidad, entonces se procede a analizar el influjo de características de los estudios (variables moderadoras) que puedan dar cuenta de al menos parte de dicha variabilidad (Botella y Sánchez Meca, 2015). Cuando en los resultados de los estudios no se sintetizan mediante técnicas de análisis estadístico, entonces tenemos una RS que no es un MA. Por tanto, todo MA es un tipo especial de RS, pero no toda RS tiene que ser un MA.

Cualquier pregunta sobre la que se hayan realizado estudios cuantitativos es susceptible de una RS o de un MA. En Ciencias de la Educación el tipo de MA más frecuente es sobre la eficacia de programas, intervenciones, tratamientos, técnicas, etc. para resolver, mejorar o prevenir algún problema

relacionado con la educación. Así, por ejemplo, Cheng et al. (2019) realizaron un MA para estimar la eficacia de las 'clases invertidas' ('flipped clasrooms') como estrategia de aprendizaje en el aula para mejorar el rendimiento académico. Seleccionaron 55 estudios empíricos que abordaron dicha pregunta y la síntesis estadística de los resultados resultó en un tamaño del efecto medio g_+ = 0.193 (IC95%: 0.113 −0.274)[1] en favor de las clases invertidas como recurso didáctico, una magnitud del efecto baja. Menos frecuente es la realización de RSs que no son MAs, es decir, sin realizar una síntesis estadística de los tamaños del efecto. Tal es el caso de la RS publicada por Piñeiro-López et al. (2022) sobre la eficacia de las intervenciones educativas dirigidas a mejorar conductas prosociales y empatía en alumnado con altas capacidades. Estos autores seleccionaron 16 estudios empíricos y, mediante una síntesis cualitativa de sus resultados, concluyeron que estos programas logran mejorar el desarrollo socioemocional e integral de este alumnado.

El MA también se puede aplicar para sintetizar

estudios correlacionales acerca de la relación existente entre dos constructos o variables. Por ejemplo, Scherer y Shiddiq (2019) realizaron un MA para estimar el sentido y la fuerza de la relación existente entre el estatus socioeconómico y el grado de conocimientos sobre las nuevas tecnologías de la información y la comunicación (TIC). Para ello, localizaron 11 estudios con muestras de estudiantes de Educación Primaria y Secundaria y obtuvieron un coeficiente de correlación promedio r_+ = 0.214 (IC95%: 0.1840.244), indicando la existencia de una asociación positiva de magnitud moderada entre estatus socioeconómico y conocimientos de las TIC.

Otros MAs se han dirigido a comparar alguna característica o sintomatología de una población con algún problema frente a población comunitaria o normalizada. Por ejemplo, Patros et al. (2016) realizaron un MA en el que comprobaron si los niños y adolescentes con trastorno por

---

[1]El índice del tamaño del efecto representado por la letra '*g*' hace referencia a la diferencia de medias

estandarizada con una corrección para muestras pequeñas propuesta por Hedges (1981). La abreviatura IC95% se refiere al intervalo de confianza asumiendo un nivel de confianza del 95%.

déficit de atención con hiperactividad (TDAH) puntúan más alto que los normalizados en impulsividad de elección. Estos autores sintetizaron los resultados de 26 estudios obteniendo un efecto medio $g_+ = 0.47$ (IC95%: 0.40 – 0.54), de magnitud moderada, indicando la existencia de una mayor impulsividad de elección en TDAH.

En el ámbito de la Educación son numerosos los MAs que se han llevado a cabo para comprobar la existencia de diferencias de género en diferentes habilidades educativas o rendimientos académicos. Así, Lindberg et al. (2010) realizaron un MA para comprobar si, en población infantojuvenil, existen

diferencias entre varones y mujeres en rendimiento en Matemáticas. Mediante la integración de los resultados de 242 estudios, estos autores obtuvieron un tamaño del efecto medio $d_+ = 0.05$,[2] indicando ausencia de diferencias de género en rendimiento en Matemáticas.

Menos frecuente en Educación es el uso del MA para estimar la prevalencia de alguna enfermedad, trastorno o déficit en población comunitaria. Por ejemplo, Bahadivand et al. (2021) realizaron un MA en el que sintetizaron 37 estudios con muestras de adolescentes iraníes que aportaban datos sobre prevalencia de conductas de alto riesgo, tales como consumo de drogas, alcohol y tabaco, obteniendo estimaciones medias de dichas prevalencias del 4% (IC95%: 3%5%), 9% (IC95%: 6% - 10%) y 9% (IC95%: 7% - 10%), respectivamente.

Por último, otro tipo especial de MA es el que se centra en analizar las propiedades psicométricas de tests y escalas de uso habitual en contextos educativos (Schmidt y Hunter, 2015). Los denominados 'meta-análisis psicométricos' consisten en sintetizar coeficientes de validez (MA de

generalización de la validez) o coeficientes de fiabilidad (MA de generalización de la fiabilidad) obtenidos al aplicar un determinado test en diferentes muestras de participantes a lo largo de los estudios (Badenes-Ribera et al., 2020; Sánchez-Meca et al., 2013; Sánchez-Meca y López-Pina, 2008). Así, Breidbord y Croudace (2013) realizaron un MA de generalización de la fiabilidad con el propósito de caracterizar el error de medida de la 'Escala de Valoración del Autismo Infantil' (*Childhood Autism Rating Scale*, CARS). Estos autores lograron sintetizar los coeficientes de fiabilidad de 36 estudios, obteniendo un coeficiente alfa de Cronbach promedio igual a $\alpha_+ = 0.896$ (IC95%: 0.877 – 0.913) y un grado de acuerdo inter-jueces igual a $r_+ = 0.796$ (IC95%: 0.7360.844).

1.2. El enfoque de la Educación Basada en la Evidencia

Estrechamente vinculado a las RSs y los MAs está el enfoque de la Práctica Basada en la Evidencia que, en el ámbito de la Educación, pasó a denominarse

'Educación Basada en la Evidencia' (*Evidence-Based Education*; EBE).[3] Este enfoque surgió en el ámbito de la Medicina, donde se denominó 'Medicina Basada en la Evidencia' (*Evidence-Based Medicine*) y posteriormente se ha ido extrapolando a todas las ciencias sociales y de la salud. En el contexto de la Educación, el enfoque de la EBE representa un modo de pensamiento en el que se preconiza que la práctica profesional en Educación debe aplicar los programas, métodos de aprendizaje, técnicas de

---

[2]El índice del tamaño del efecto representado con la letra '*d*' se refiere a la diferencia de medias estandarizada propuesta por Cohen (1988). A diferencia del índice *g* de Hedges, la *d* de Cohen sufre un ligero sesgo positivo con muestras pequeñas. No obstante, el efecto medio de un meta-análisis está basado en un tamaño muestral muy elevado, por lo que dicho sesgo puede considerarse despreciable.

[3]La traducción más correcta al castellano del término inglés 'evidence-based education' sería 'Educación basada en pruebas', no 'en evidencias'. No obstante, dado que se ha extendido el uso del término 'evidencias' en

lugar de 'pruebas', hemos preferido mantenerlo en este artículo.

mejora, de rehabilitación, técnicas de diagnóstico, etc. que hayan recibido las mejores evidencias científicas (Davies, 1999; Deeker y Meeter, 2022). En lugar de basar el desempeño profesional exclusivamente en la opinión de expertos, el EBE recomienda que el profesional de la Educación guíe su desempeño profesional basándose en los estudios científicos que demuestran qué métodos y técnicas son los más eficaces. Todo ello con la ayuda de las TIC. Conocer cuáles son los mejores métodos educativos requiere una actualización continua por parte de los profesionales de la Educación. Esa actualización sería inviable si no fuera gracias a la existencia de las RSs y los MAs, ya que este tipo de investigación sintetiza los resultados

de un conjunto de estudios sobre un mismo problema o pregunta. Gracias a los MAs el profesional sólo necesita leer un artículo (el MA) en lugar de tener que leerse los estudios individuales, facilitando la puesta en práctica del enfoque de la EBE. Es por ello que los MAs van ligados a este enfoque tan útil para los profesionales de la Educación. La utilidad del enfoque de la EBE no sólo afecta a los maestros, profesores, educadores, etc. que están en contacto directo con los escolares, universitarios y demás receptores de los procesos educativos, sino que también es aplicable en niveles superiores donde se toman decisiones en materia de políticas educativas, tanto autonómicas como nacionales y supranacionales.

1.3. Objetivos

El propósito de este artículo es ofrecer un tutorial sobre qué son las RSs y los MAs, cómo se hacen y cómo se interpretan sus resultados. Este propósito general se puede desglosar en dos objetivos más específicos. En primer lugar, se pretende ofrecer

una panorámica de cómo se lleva a cabo un MA, qué fases o etapas se siguen en su desarrollo y qué decisiones tiene que afrontar un investigador durante su realización. En segundo lugar, se pretende ofrecer pautas para una correcta lectura de los MAs, indicios para valorar si un MA está bien realizado o si, por el contrario, presenta deficiencias metodológicas que puedan comprometer la validez de sus conclusiones.

2. MÉTODO

Para alcanzar los objetivos de este artículo, se ha realizado una revisión comprehensiva de la literatura meta-analítica, que incluye más de una centena de libros publicados hasta la fecha sobre esta metodología y multitud de artículos que abordan problemas, decisiones, estrategias y guías para la realización de un MA y para hacer lectura crítica de RSs y MAs.

3. RESULTADOS

Las orientaciones y recomendaciones que se presentan en este tutorial se estructuran en dos partes. En primer lugar, se abordan las fases que en la práctica se siguen para realizar una RS y un MA, con especial detenimiento en las decisiones que, a lo largo del proceso, el investigador tiene que tomar. En segundo lugar, se presentan guías, checklists y recomendaciones para hacer una correcta lectura crítica de RSs y MAs, haciendo especial hincapié en los diferentes checklists que se han propuesto en la literatura para diferentes tipos de MA (ej., para MAs sobre la eficacia de intervenciones, MAs de estudios no experimentales, MAs sobre la precisión de pruebas diagnósticas, MAs de generalización de la fiabilidad, etc.).

3.1. Fases de una revisión sistemática / meta-análisis

La realización de un MA conlleva las mismas fases que la de cualquier investigación primaria, aunque con una serie de peculiaridades que le dan un carácter especial. La principal diferencia entre un estudio primario y un MA está en la unidad de análisis. Mientras que en los estudios primarios la unidad de análisis suelen ser los participantes,[4] en un MA la unidad de análisis está formada por los estudios primarios. Las fases o etapas que se siguen para llevar a cabo un MA se pueden estructurar en las siguientes: (1) formulación de la pregunta, (2) definición de los criterios de selección de los estudios, (3) búsqueda y localización de los estudios, (4) extracción de la información de los estudios, (5) medida del resultado de los estudios, (6) síntesis e interpretación de los resultados y (7) redacción del MA (Borenstein et al., 2019; Cooper, 2016; Cooper et al., 2019; Lipsey y Wilson, 2001; Sánchez-Meca, 2010; Sánchez-Meca y Marín-Martínez, 2010).

3.1.1. Formulación de la pregunta

La primera tarea en la realización de un MA es formular la pregunta o preguntas de interés, o los objetivos que se pretenden alcanzar. La pregunta de interés tiene que formularse con total concreción, sin ambigüedades. Entre otros aspectos, la correcta formulación de la pregunta en un MA pasa necesariamente por la definición conceptual y operativa de los constructos, conceptos y variables implicados en la misma. También es preciso definir cuál es la población, o poblaciones, de interés (ej., población clínica vs. comunitaria, población infanto-juvenil vs. adulta, etc.). Si el MA tiene como objetivo investigar la eficacia de programas, tratamientos o intervenciones, deben definirse con meticulosidad tales programas.

Por ejemplo, en el MA de Cheng et al. (2019) sobre la eficacia de las 'clases invertidas' (flipped classrooms) sobre el rendimiento académico, los autores plantearon la siguiente pregunta: "¿Qué efectos tienen las clases invertidas sobre los resultados de aprendizaje de los estudiantes en comparación con las clases tradicionales y qué

variables moderan tales efectos?" (Chang et al., 2019, p. 799).

Una vez que se plantea la pregunta de interés, es muy recomendable elaborar un *Protocolo* de la RS o del MA que se pretende llevar a cabo. Dicho Protocolo debe incluir todos los detalles de la investigación que se pretende realizar, es decir, los objetivos y la metodología que se va a aplicar. El Protocolo debe hacerse público mediante su publicación en algún repositorio o registrador con objeto de garantizar la máxima transparencia en el desarrollo de la investigación y, por ende, la reproducibilidad de la misma. La recomendación de publicar, previa la realización de una investigación, del Protocolo de la misma surge de la corriente actual de pensamiento en el ámbito científico basada en el *Enfoque de la Ciencia Abierta* ('Open Science Framework'; ej., Nosek et al., 2015). Repositorios públicos online en los que se puede hacer público el Protocolo de una RS o un MA son el de la organización 'Open Science Framework' o el de 'PROSPERO' del 'National Institute for Health Research' de la Universidad de York en el Reino

Unido (pueden consultarse los sitios web de estas dos organizaciones en el apartado 'Enlaces').

---

[4]Si bien en Ciencias de la Educación es muy frecuente que la unidad de análisis en los estudios primarios sean personas (ej., niños, adolescentes, profesores, estudiantes, padres, etc.), caben otras unidades de análisis como, por ejemplo, la díada estudiante-profesor, la díada estudiante-padre o madre, la unidad familiar, el aula, el centro escolar, el distrito escolar, etc.). Por simplicidad, nos referimos a las personas como unidad de análisis por ser la más frecuente.

3.1.2. Criterios de selección de los estudios

Una vez planteada la pregunta de interés, la siguiente fase consiste en definir los criterios de selección de los estudios, es decir, qué características

debe cumplir un estudio primario para que pueda ser incluido en el MA. Los criterios de selección de un MA dependerán de la pregunta de interés. No obstante, se pueden ofrecer varias pautas para su definición. Un criterio obligatorio es que los estudios que se integran en un MA tienen que ser necesariamente estudios empíricos de naturaleza cuantitativa, es decir, deben aportar datos estadísticos que permitan obtener una estimación numérica del resultado.[5]

Cuando el objetivo de un MA es examinar la eficacia de programas, tratamientos o intervenciones, es recomendable utilizar el acrónimo PICOS para no olvidar una serie de criterios importantes que deben tenerse en cuenta a la hora de definir tales criterios de selección (cf., ej., Higgins et al., 2022):

- *Participants* (Participantes): debe especificarse cuál es la población o poblaciones de participantes de interés (ej., adolescentes escolarizados pertenecientes a población comunitaria).

- *Interventions* (Intervenciones): debe concretarse el programa o programas cuya efectividad se está interesado en investigar (ej., programas de tutorización parental).
- *Comparison group* (Grupo de comparación): debe concretarse qué tipos de grupos de comparación se van a aceptar junto con el grupo experimental (ej., grupos de control inactivos, grupos de control activos, grupos de control que implementan el programa tradicional, grupos que reciben un programa alternativo al de interés).
- *Outcomes* (Variables de resultado): debe especificarse cuáles son las variables dependientes sobre las que se pretende comprobar la eficacia de los programas objeto de interés (ej., ansiedad, depresión, autoconfianza, calificaciones del rendimiento académico, etc.).
- *Study design* (Diseño del estudio): debe concretarse qué diseños se aceptarán en el MA, pudiendo tratarse de diseños experimentales, diseños cuasi-experimentales,

diseños de cohortes, diseños de casos y controles, diseños transversales, diseños correlacionales, etc.

Si el MA no tiene como propósito investigar la eficacia de programas e intervenciones, sino asociaciones entre variables o efectos de factores de exposición sobre variables de resultado, entonces el acrónimo PICOS se sustituye por de PECOS, donde la única dimensión que cambia es la 'I' de 'Intervención' por la 'E' de 'Exposure factor' (Factor de exposición), es decir, en este caso debe definirse cuál fue el factor de exposición o variable cuyo posible efecto pretende investigarse sobre la(s) variable(s) de resultado (ej., un factor de exposición sobre el rendimiento académico posterior podría ser haber asistido a guardería frente a no haber asistido). Los acrónimos PICOS y PECOS pueden también utilizarse para formular la pregunta objeto de interés en la primera fase de un MA.

---

[5]Si en lugar de un MA estamos interesados en realizar

una RS, entonces los resultados de los estudios pueden extraerse de forma cuantitativa o cualitativa. Además, en una RS que no es un MA se pueden incluir tanto estudios empíricos de naturaleza cuantitativa como cualitativa.

Junto con estos criterios de selección cabe mencionar otros adicionales. Es preciso determinar la franja temporal en la que tienen que haberse realizado (o publicado) los estudios primarios para el MA. Es posible que estudios previos a un determinado año se consideren ya obsoletos, por lo que un criterio de selección sería que los estudios estén publicados después de esa fecha. Otro aspecto que conviene especificar en los criterios de selección es que los estudios tienen que aportar algún índice del tamaño del efecto, o bien aportar los datos estadísticos necesarios para que el meta-analista pueda calcularlo por su cuenta. Otro criterio que se debe especificar es el idioma o

idiomas en los que tiene que estar escrito el estudio para que pueda ser incluido en el MA. Ello se debe a que los investigadores que pretenden hacer un MA no dominarán todos los idiomas. En cualquier caso, el idioma que no puede faltar en un meta-análisis es el inglés, ya que actualmente ésta es la lengua vehicular de la ciencia.

Por último, otro aspecto que debe especificarse en los criterios de selección es si se van a aceptar sólo estudios publicados (en revistas o libros) o si se permitirá la inclusión de estudios no publicados (ej., contribuciones en congresos, tesis doctorales no publicadas, informes técnicos, trabajos fin de máster, etc.). Lo más recomendable es aceptar a priori la inclusión tanto de estudios publicados como no publicados. Esta recomendación se debe a la existencia del problema del sesgo de publicación, un fenómeno muy extendido en muchas disciplinas según el cual, es más fácil publicar en las revistas científicas estudios con resultados estadísticamente significativos ($p < .05$) y, por tanto, con tamaños del efecto de mayor magnitud, que estudios con resultados estadísticamente no

significativos (p > .05), que suelen estimar efectos de menor magnitud. Un MA que se nutre exclusivamente de estudios publicados puede ofrecer una sobreestimación del verdadero efecto en la población objeto de interés (Rothstein et al., 2005).

Por ejemplo, el MA de Cheng et al. (2019, Tabla 2, p. 801) sobre la eficacia de las clases invertidas presenta la relación de criterios de inclusión y de exclusión que tenían que cumplir los estudios para ser aceptados en dicho MA. Entre otros, estos autores incluyeron como criterios de selección el idioma del estudio, el tipo de diseño, la franja temporal en que se tenían que haber realizado, las variables de resultado, etc.

3.1.3. Búsqueda de los estudios

Una vez fijados los criterios de selección de los estudios, el siguiente paso consiste en intentar localizar en la literatura los estudios que cumplan con dichos criterios. Dado que la búsqueda de los estudios en un MA debe ser lo más comprehensiva

posible, lo más recomendable es utilizar varias fuentes de búsqueda. Cabe clasificar las fuentes de búsqueda en dos categorías: *fuentes formales* y *fuentes informales* (Glanville, 2019). Las *fuentes formales* son aquéllas que nos permiten localizar principalmente estudios publicados (en revistas o como libros o capítulos de libro). Entre las fuentes formales la más importante en un MA es la consulta de bases electrónicas, tales como ERIC en el campo de la Educación, PsycInfo en Psicología o Medline en Medicina y ciencias de la salud afines. Una correcta consulta en dichas bases electrónicas requiere definir previamente la combinación de palabras-clave que mejor pueden ayudar a localizar estudios que pudieran cumplir con los criterios de selección. Las palabras-clave deben combinarse convenientemente utilizando los conectores 'and' y 'or', especificando si la búsqueda debe realizarse sólo en el título del artículo, en el abstract o a texto completo. Otra fuente formal consiste en revisar de forma más detenida los artículos publicados en revistas que se

sabe son prolíficas en la publicación de estudios que pudieran cumplir con los criterios de selección. Otra

fuente formal consiste en revisar las referencias de los trabajos citados en aquellos estudios que cumplen con los criterios de selección del MA, con objeto de localizar posibles estudios elegibles para el mismo. Si se encuentran en la literatura RSs o MAs previamente publicados sobre la pregunta de interés, o sobre preguntas similares, otro procedimiento formal de búsqueda consiste en revisar la lista de estudios incluidos en esas revisiones o MAs para intentar localizar estudios que pudieran cumplir con los criterios de selección.

Como se comentó más arriba, para contrarrestar el problema del sesgo de publicación es aconsejable incluir en un MA estudios no publicados. Para intentar localizar estudios no publicados se utilizan otros procedimientos de

búsqueda denominados 'fuentes informales'. Este tipo de fuentes obedecen a estrategias que podrían definirse como 'detectivescas' y no tan sistemáticas como las de las fuentes formales. No en vano, los estudios no publicados suelen denominarse 'literatura fugitiva' (*fugitive literature*) o 'literatura gris' (*grey literature*). Métodos informales de búsqueda incluyen la consulta de sitios web, a través de internet, de asociaciones, organismos o entidades que sabemos han investigado sobre la pregunta de interés del MA. Cabe también mencionar la revisión de las actas de congresos celebrados sobre el tema, o repositorios de tesis doctorales no publicadas. Otra fuente informal muy aconsejable es elaborar una lista de autores prolíficos en la realización de estudios sobre el tema y escribirles un correo electrónico solicitándoles que envíen estudios no publicados sobre el tema que hayan realizado (Cooper, 2016; Giustini, 2019).

El resultado del proceso de búsqueda, cribado y selección definitiva de los estudios que cumplen con todos los criterios de selección del MA se debe

reportar mediante un diagrama de flujo (*flow chart*) que resuma dicho proceso. Este diagrama de flujo ayuda a facilitar la replicabilidad del proceso de búsqueda realizado en el MA (Higgins et al., 2022). De todos los modelos de diagrama de flujo disponibles en la literatura, el que se utiliza con mayor frecuencia es el diagrama de flujo PRISMA, que ha sido recientemente actualizado (PRISMA 2020; consúltese en el epígrafe 'Enlaces' el sitio web desde el que puede descargarse este diagrama de flujo). En el proceso de cribado de los estudios, facilita la tarea el uso de algún gestor bibliográfico, tales como Mendeley o Zotero. Dado que un mismo estudio puede estar recogido en diferentes bases electrónicas, el uso de un gestor bibliográfico permite identificar y excluir de la base los estudios repetidos. Por último, el proceso de selección de los estudios que cumplen los criterios de inclusión de un MA debe someterse a un análisis de su fiabilidad. Para ello, al menos dos investigadores del equipo de meta-analistas debería realizar el proceso de selección de estudios de forma independiente y,

posteriormente, comprobar el grado de acuerdo entre ellos, pudiéndose aportar alguna evidencia de fiabilidad mediante el cálculo de un coeficiente de acuerdo inter-codificadores, tales como kappa de Cohen.

Por ejemplo, en el meta-análisis de Cheng et al. (2019, p. 799) sobre la eficacia de las clases invertidas, los autores realizaron búsquedas electrónicas en los siguientes recursos: EBSCOhost, ProQuest, Web of Science y World Cat, entre otros. También describieron los términos de búsqueda con la combinación de palabras-clave elegidas, así como el período de búsqueda (20002015). Estos autores no utilizaron ningún otro método de búsqueda, ni formal ni informal. Este MA también presenta un diagrama de flujo en el que se describe el proceso de selección de los estudios en sus diferentes fases (Cheng et al., 2019, p. 802, Figura 2).

3.1.4. Extracción de la información de los estudios

Una vez seleccionados los estudios que cumplen con los criterios de inclusión del MA, la siguiente etapa consiste en extraer la información relevante para la pregunta de interés. Cabe distinguir dos tipos de datos e información relevantes: sobre las características de los estudios y sobre los resultados obtenidos en los mismos respecto de la pregunta a la que pretende responder el MA. En esta fase (4) del MA centramos la atención en la extracción de las características de los estudios, mientras que la fase (5) se dedica a presentar cómo obtener índices cuantitativos del resultado de los estudios.

En relación con las características de los estudios, cabe distinguir, a su vez, entre *características sustantivas* y *características metodológicas* (Cooper, 2016; Lipsey, 2019). Las *características sustantivas* son aquéllas que tienen que ver con la pregunta del MA. Entre la información a extraer de los estudios, no pueden faltar las características

sociodemográficas de las muestras de participantes, tales como la edad promedio de la muestra, la distribución por género (ej., porcentaje de mujeres) o la distribución étnica. Dependiendo de los objetivos del MA, puede ser también de interés extraer características clínicas de la muestra (ej., el tipo de trastorno, deficiencia o problema, la gravedad del problema, etc.), o características educativas (ej., curso, nivel o grado cursado por los participantes de la muestra). Otras características sustantivas relevantes pueden referirse al contexto de procedencia, aplicación o realización del estudio (ej., si la muestra procedía de centros públicos o privados; si el programa de intervención se aplicó en un centro educativo, en un gabinete privado o en el propio hogar del estudiante; el país de realización del estudio, etc.).

Cuando el propósito del MA es examinar la eficacia de programas, tratamientos o intervenciones, dentro de las características sustantivas no pueden faltar variables relacionadas con cómo se aplicó el programa. Por ejemplo, podrá resultar de interés extraer de los estudios la duración del programa (ej.,

número de semanas o de meses), la intensidad del programa (ej., número de horas por semana recibidas por cada participante), modo de aplicación del programa (ej., de forma individual o grupal, si implicó a los padres o no, si el formato fue online o presencial, etc.), o diferentes versiones o adaptaciones del programa.

Además de características sustantivas, se deberán extraer *características metodológicas*, que son aquéllas que tienen que ver con cómo se llevó a cabo la investigación. Los aspectos metodológicos permiten valorar si el estudio presentó buena calidad metodológica o si, por el contrario, adolece de deficiencias que comprometen la validez interna de sus resultados. Estudios con pobre calidad metodológica pueden dar lugar a estimaciones sesgadas de los efectos investigados, arrojando una imagen distorsionada del verdadero efecto en la población. Un MA que integra estudios con pobre calidad metodológica puede alcanzar conclusiones erróneas. Es por ello que todo MA debe incorporar alguna escala o checklist de valoración de la calidad metodológica de los estudios primarios

incluidos en el mismo. Se han propuesto en la literatura un extenso número de escalas, especialmente en ciencias de la salud, si bien su aplicación es fácilmente extrapolable al ámbito de la Educación (Conn y Rantz, 2003; Saunders et al., 2003).

Independientemente de qué escala metodológica se decida utilizar en un MA, hay una serie de ítems que no deben faltar. Uno de ellos es el tipo de diseño del estudio (ej., diseño experimental, cuasi-experimental, transversal, de cohortes, etc.). Otro aspecto metodológico a extraer es la mortalidad experimental, es decir, el porcentaje de participantes que inician el estudio pero que

no lo finalizan. Otra característica es si los evaluadores o aplicadores de las pruebas y escalas estaban enmascarados respecto del grupo de

procedencia de los participantes (para evitar sesgos de las expectativas del experimentador). Otra característica tiene que ver con la posible presencia de sesgo de reporte, es decir, si se reportaron los resultados de todas las variables dependientes que en la sección Método se informaron o si, por el contrario, algunas variables dependientes recogidas en el Método no se reportaron en la sección de Resultados. Otra característica metodológica es, caso de que hubiera pérdida de participantes, si se realizaron análisis estadísticos por intención de tratar o si sólo se analizaron los que completaron la investigación.

De los múltiples checklists de calidad metodológica que se han propuesto, la elección del más apropiado para los estudios empíricos de un MA dependerá del tipo de diseños de dichos estudios. Puede consultarse el sitio web 'Equator Network' para seleccionar la escala o checklist metodológico más apropiado a las características de los estudios primarios de un MA (véase en el epígrafe 'Enlaces' la dirección web). No obstante, cabe recomendar algunas escalas en concreto. Para MA sobre la

eficacia de programas, una escala muy recomendable para valorar la calidad metodológica de los estudios primarios es la escala PEDro (Verhagen et al., 1998; véase en el epígrafe 'Enlaces' la dirección web). Para estudios primarios con diseños no experimentales, tales como diseños de cohortes o de casos y controles, es muy recomendable la escala de Newcastle-Ottawa (NOS; Wells et al., 2000; en el epígrafe 'Enlaces' se recoge el sitio web de esta escala). Por último, para estudios transversales es muy recomendable la escala AXIS propuesta por Downes et al. (2016).

Además de las características sustantivas y metodológicas, otro conjunto de posibles variables de interés son las *características extrínsecas* (Lipsey y Wilson, 2001). Éstas reciben su nombre porque son características de los estudios que no deberían afectar a los resultados de una investigación, porque no tienen nada que ver con la aplicación del método científico. Sin embargo, en ocasiones se observa que algunas de estas características afectan a los resultados debido a sesgos provocados por las expectativas del investigador. Cabe mencionar

en este bloque de características el hecho de que la fuente de financiación de un estudio empírico sea pública o privada, o la posible existencia de conflicto de intereses entre los autores del estudio. También puede ser de interés extraer la formación de base del investigador principal del estudio (ej., pedagogo, educador social, psicólogo educativo, trabajador social, etc.), porque en ocasiones se puede observar un efecto del 'corporativismo' sobre los resultados del estudio. Si el MA ha sido capaz de incluir tanto estudios publicados como no publicados, esta característica también debe registrase en la base de datos, con objeto de comprobar posteriormente si existe sesgo de publicación. Por último, incluso el sexo del investigador principal puede ser una característica relevante, sobre todo cuando el propósito del MA es investigar diferencias entre hombres y mujeres en determinadas habilidades, competencias, destrezas, rendimientos, actitudes, aptitudes, etc.

El proceso de extracción de las características de los estudios puede sufrir errores. Es por ello que se recomienda elaborar un 'Protocolo' de extracción

de las mismas, junto con un 'Manual de Codificación' en el que se detalle cómo se debe codificar cada una de las características de los estudios objeto de interés. También es muy recomendable realizar un análisis de la fiabilidad de este proceso. Para ello, al menos dos investigadores del equipo de meta-analistas deben llevar a cabo la extracción de las características de los estudios, utilizando el Protocolo preestablecido y aplicando el Manual de Codificación. De esta forma, se pueden aportar evidencias del grado de

acuerdo inter-codificadores, tales como coeficientes kappa de Cohen para las variables categóricas (ej., tipo de diseño, país de realización del estudio) o coeficientes de correlación intraclase para las continuas (ej., la duración del programa, la edad media de la muestra). Las inconsistencias

entre los codificadores se deben resolver por consenso o mediante la intervención de un tercer investigador (ej., Vevea et al., 2019; Wilson, 2019).
Por ejemplo, en el meta-análisis de Cheng et al. (2019) sobre la eficacia de las clases invertidas sobre el rendimiento académico, se extrajeron características tales como la población de referencia de los participantes (según franja de edad), materias sobre las que se valoró el rendimiento académico (ej., matemáticas, lengua, ciencias sociales, etc.), duración del programa, año de realización del estudio y el país del estudio, entre otras.

3.1.5. Medida del resultado de los estudios

Para la realización de una RS o de un MA es imprescindible extraer el resultado de cada estudio en relación a la pregunta de interés. Si, por ejemplo, el objetivo de un MA es comprobar si hombres y mujeres de una determinada franja de edad difieren o no en habilidades para las Matemáticas, el resultado que interesa extraer de

cada estudio debe informar sobre el grado en que ambos sexos difirieron en dicha habilidad. Siempre que sea posible, es preferible obtener dicho resultado de forma cuantitativa. Los índices estadísticos más apropiados para cumplir con esta misión son los denominados *índices del tamaño del efecto*. Kelley y Preacher (2012) definen el tamaño del efecto como "una representación cuantitativa de la magnitud de un fenómeno que se utiliza para responder a una pregunta de interés" (p. 140).

Existen muchos índices del tamaño del efecto susceptibles de ser utilizados en una RS o un MA, tantos que incluso algunos de ellos se agrupan en familias. La elección del índice del tamaño del efecto más apropiado para un MA está en función de tres factores. En primer lugar, de la pregunta de interés. Así, no es lo mismo que el objetivo del MA sea comparar dos poblaciones (ej., hombres y mujeres) sobre alguna variable continua, en cuyo caso se podría utilizar una diferencia de medias o una diferencia de medias estandarizada, que el objetivo sea estimar el sentido y la fuerza de la relación entre dos variables continuas (ej., rendimiento académico

y expectativas de autoeficacia), en cuyo caso el índice del tamaño del efecto más apropiado sería un coeficiente de correlación. El segundo factor que condiciona la elección del índice del tamaño del efecto es el tipo de diseño, o diseños, que los estudios deben tener para ser incluidos en el MA. Así, en diseños experimentales o cuasi-experimentales con asignación a grupos pueden utilizarse diferencias de medias, mientras que en estudios con diseños correlacionales lo más habitual será utilizar coeficientes de correlación. El tercer factor a tener en cuenta es la naturaleza de las variables implicadas en la pregunta de interés para el MA. Así, en estudios con dos grupos se podría utilizar una diferencia de medias si la variable dependiente es continua, mientras que si ésta es dicotómica habrá que utilizar una diferencia de proporciones, una razón de proporciones o una razón de ventajas. En función de estos factores, el meta-analista tendrá que elegir a priori el índice del tamaño del efecto idóneo para cuantificar el resultado de cada estudio incluido en el MA. En Educación cabe mencionar especialmente dos familias de índices del tamaño

del efecto de uso más frecuente: la *familia d* y la *familia r* (Rosenthal, 1991). Existen otros índices del tamaño del efecto susceptibles de ser utilizados en una RS o un MA, pero su tratamiento excede el alcance de este artículo. Pueden consultarse diversas fuentes para profundizar en dichos índices (cf., ej.,

Borenstein y Hedges, 2019; Card, 2012; Cortina y Nouri, 2000; Cumming, 2012; Grissom y Kim, 2012; Ellis, 2010; Rosenthal et al., 2000; Sánchez-Meca, 2008; Sapp, 2017; White et al., 2021)

La *familia d* aglutina un conjunto de índices del tamaño del efecto que se caracterizan por la comparación de dos medias mediante el cálculo de su diferencia. Ello implica que para su aplicación la variable dependiente tiene que ser cuantitativa y, en consecuencia, susceptible de obtener medias y desviaciones típicas. De todos los

índices de la familia *d*, el de uso más frecuente es la *diferencia de medias estandarizada*, que se define como la diferencia entre las dos medias dividida por un promedio de las desviaciones típicas de los dos grupos. Este índice fue propuesto por Glass et al. (1981), aunque se le conoce más comúnmente como el índice *d* de Cohen (1988). Hedges y Olkin (1985; véase también Hedges, 1981) demostraron que el índice *d* de Cohen exhibe un sesgo positivo con muestras pequeñas, proponiendo una diferencia de medias estandarizada que corrige dicho sesgo y se conoce como el índice *g* de Hedges.[6] El hecho de dividir por una desviación típica permite que estudios que han aplicado escalas diferentes para medir un mismo constructo o variable (ej., ansiedad a las Matemáticas) puedan medirse sus resultados en una métrica común (unidades estándar). Que los resultados de los estudios incluidos en un MA estén cuantificados en una métrica común es un requisito imprescindible para poder aplicar posteriormente técnicas de análisis estadístico sobre ellos. Dada la gran diversidad de escalas e instrumentos para medir una misma variable en

Educación, el uso de la diferencia de medias estandarizada permite solventar el problema de la heterogeneidad de los instrumentos de medida y posibilitar así la síntesis estadística propia del MA.

Dentro de la familia *d* caben otros índices del tamaño del efecto aplicables en un MA. Cuando los estudios primarios de un MA han aplicado diseños pretest-postest con un solo grupo (i.e., sin grupo de control) y la variable dependiente es continua, una adaptación de la diferencia de medias estandarizada es el *índice de cambio medio estandarizado*, que consiste en calcular la diferencia entre las medias del pretest y el postest y dividirla por una desviación típica, que puede ser la desviación típica del pretest, la del postest, un promedio de ambas, o la desviación típica de las puntuaciones de cambio. En realidad, dividir por diferentes desviaciones típicas hace que estemos definiendo índices estadísticos diferentes que estiman a parámetros diferentes (aunque relacionados entre sí). Es decisión del meta-analista la elección de cuál de estos índices se desea aplicar en el MA (Becker, 1988; Borenstein et al., 2019; Morris y DeShon, 2002). Otra adaptación de

la diferencia de medias estandarizada es el índice de la *diferencia de cambios medios estandarizados*. Este índice es aplicable cuando los estudios primarios del MA utilizan diseños de dos grupos (ej. experimental vs. control) con medidas pretest y postest. En este caso, el índice más apropiado consiste en calcular la diferencia entre los cambios medios pretest-postest de ambos grupos y dividirla por una desviación típica, que puede ser el promedio de las desviaciones típicas del pretest de los dos grupos, de las del postest, o de las desviaciones típicas de las puntuaciones de cambio (Morris, 2008). Los índices de la familia *d* hasta aquí descritos, basados en la diferencia de medias estandarizada, se caracterizan por utilizar un 'estandarizador' en el denominador (una desviación típica), con objeto de traducir a una métrica común los resultados de los estudios primarios que han utilizado

---

[6]Glass et al. (1981) propusieron otra diferencia de medias estandarizada diferente a la *d* de Cohen y a la *g* de Hedges, que se conoce como *Delta* de Glass. Este índice, a diferencia de los dos anteriores, divide por la desviación

típica del grupo de control, en lugar de por la desviación típica promedio de los dos grupos.

diferentes escalas para medir un mismo constructo o variable. Si un MA incluye estudios primarios, todos los cuales han utilizado la misma escala o instrumento para medir la variable dependiente de interés, entonces puede utilizarse el índice *diferencia de medias* (sin estandarizar). En este caso, ya no es necesario dividir la diferencia de medias por ninguna desviación típica, ya que tales diferencias de medias están en la misma escala métrica (Botella y Sánchez-Meca, 2015).

La otra familia de índices del tamaño del efecto de uso más frecuente en MA es la *familia r*, que incluye cualquier coeficiente de correlación (de Pearson, ordinal de Spearman, biserial, biserial puntual, phi,

etc.). Estos tamaños del efecto son aplicables cuando los estudios primarios del MA han utilizado diseños correlacionales o asociativos y la pregunta de interés tiene que ver con el sentido y la fuerza de la relación entre dos variables. La elección de uno u otro coeficiente de correlación está en función del nivel de medida de las dos variables implicadas en su cálculo, si bien el de uso más frecuente es el *coeficiente de correlación de Pearson*. Con objeto de normalizar la distribución y estabilizar las varianzas de los coeficientes de correlación, en un MA se suelen transformar las correlaciones a la Z de Fisher, de forma que la síntesis estadística se lleva a cabo con las Zs de Fisher en lugar de con las correlaciones directamente (Borenstein et al., 2019).

Además de los índices del tamaño del efecto recogidos en la *familia d* y la *familia r*, otros muchos son susceptibles de ser utilizados en MA, aunque con menor frecuencia. Por ejemplo, si el objetivo del MA es estimar la prevalencia de una discapacidad o diversidad educativa en una determinada población (ej. la prevalencia del TDAH

en la adolescencia), el resultado cuantitativo que interesa extraer de cada estudio primario para su incorporación al MA es la prevalencia de dicho diagnóstico observada en la muestra de participantes, es decir, la proporción de personas con ese diagnóstico dividida por el tamaño de la muestra. Vemos cómo en este caso el índice del tamaño del efecto puede ser una simple proporción. Si el tipo de MA que se pretende realizar es un MA de generalización de la fiabilidad de un determinado test o escala, entonces el índice del tamaño del efecto que interesará extraer de cada estudio primario que haya aplicado la escala en cuestión será el coeficiente de fiabilidad reportado en el estudio (ej., un coeficiente alfa de Cronbach, un coeficiente Omega, una correlación test-retest, etc.). Por tanto, son muchos los índices estadísticos que pueden actuar como índices del tamaño del efecto para los propósitos de un MA.

A diferencia de los contrastes de hipótesis, que informan de la existencia o no de un resultado estadísticamente significativo, los índices del

tamaño del efecto informan de la significación práctica o real del resultado de una investigación. La interpretación de los índices del tamaño del efecto requiere del juicio razonado del meta-analista. Algunos índices del tamaño del efecto permiten hacer una interpretación directa de su relevancia práctica (ej., una diferencia entre dos medias sobre una variable con significado práctico), pero los índices del tamaño del efecto estandarizados (ej., diferencias de medias estandarizadas) y los coeficientes de correlación necesitan de indicios o guías orientativas para informar de su relevancia práctica. Se han propuesto en la literatura guías orientativas con este fin. La más conocida y utilizada es la propuesta por Cohen (1988) para las diferencias de medias estandarizadas y para los coeficientes de correlación. Así, este autor propuso diferencias de medias estandarizadas de 0.20, 0.50 y 0.80 (en valor absoluto) como reflejando magnitudes del efecto de relevancia baja, moderada y alta, respectivamente. Respecto de los coeficientes de correlación, este autor propuso valores de 0.10, 0.30 y

0.50 (en valor absoluto) con el mismo fin. Otra estrategia para interpretar el tamaño del

efecto obtenido en un estudio empírico es compararlo con los tamaños del efecto obtenidos en estudios previos similares. Si existe algún MA publicado sobre la pregunta en cuestión, el tamaño del efecto del estudio empírico puede compararse con el efecto medio reportado en dicho MA. En cualquier caso, es en el juicio crítico, pero razonado, del meta-analista donde debe fundamentarse la interpretación de los tamaños del efecto (Kline, 2019; Valentine et al., 2019).

Como ya se ha comentado, para la realización de un MA es imprescindible la obtención de índices del tamaño del efecto de los estudios primarios para su posterior síntesis estadística. En una RS que no es un MA la obtención de índices del tamaño del efecto es

en general recomendable, pero no es obligatoria. Existen diversas razones para no obtener o calcular índices del tamaño del efecto en una RS. Una de ellas es cuando los estudios primarios ni aportan los índices del tamaño del efecto ni los estadísticos necesarios para que el meta-analista pueda calcularlos por su cuenta (ej., medias, desviaciones típicas, tamaños muestrales, proporciones, etc.), ni los autores de los estudios primarios responden a nuestra solicitud de información. Otra razón puede ser que los diseños de los estudios primarios incluidos en la RS son tan dispares que no es posible traducir sus resultados a un índice del tamaño del efecto común a todos ellos (ej., algunos estudios aplican diseños experimentales, mientras que otros aplican diseños correlacionales). Cuando no es posible, o recomendable, utilizar tamaños del efecto en una RS, entonces puede recurrirse a otras formas de extraer los resultados de los estudios primarios, unas son también cuantitativas, mientras que otras son cualitativas. Estos métodos alternativos no son tan válidos y eficientes como los índices del tamaño del efecto, por lo que sólo tiene

sentido utilizarlos allá donde no sea posible, o adecuado, utilizar tamaños del efecto (McKenzie y Brennan, 2022).

Un índice cuantitativo alternativo al tamaño del efecto es extraer de cada estudio primario el valor de probabilidad unilateral exacto, *p*, asociado al resultado del contraste de hipótesis de interés para la RS. Si disponemos de dichos valores *p* para los estudios primarios, es posible llevar a cabo una síntesis estadística de éstos, como alternativa a la síntesis de los tamaños del efecto. Si los estudios primarios no reportan el valor *p* exacto del contraste de hipótesis de interés para el MA y no es posible calcularlo, puede utilizarse la dirección del resultado del estudio. En este caso, es importante tener en cuenta que el resultado que se extrae de cada estudio es si el estudio en cuestión obtuvo un resultado (en términos de medias, proporciones, etc.) de acuerdo con la hipótesis del MA o en contra de ésta (no del valor *p*). De esta forma, es posible clasificar los estudios primarios en dos categorías: a favor de la hipótesis y en contra de la hipótesis de interés. Y estos datos son susceptibles de una síntesis estadística

posterior, pero menos eficiente que si se sintetizaran los tamaños del efecto. Una última estrategia alternativa al uso de los tamaños del efecto consiste en extraer los resultados de los estudios primarios de forma narrativa, es decir, no cuantitativa, sino con frases que describan los hallazgos alcanzados. Este método es el menos aceptable o recomendable, ya que, debido a su fuerte componente de subjetividad, es propenso a sufrir sesgos de interpretación.

Por último, del mismo modo que en la extracción de las características de los estudios primarios, el cálculo o la obtención de los resultados de los estudios, ya sean tamaños del efecto o de cualquier otra índole, debe someterse a un análisis de la fiabilidad. Para ello, dos investigadores del equipo de meta-analistas deben realizar los cálculos o la extracción de los resultados de los estudios de forma independiente y contrastar sus datos, pudiéndose aportar evidencias de fiabilidad en términos de grado de acuerdo inter-codificadores mediante el cálculo de coeficientes kappa de Cohen o correlaciones intra-clase.

Siguiendo con el ejemplo del MA de Cheng et al. (2019) sobre la eficacia de las clases invertidas sobre el rendimiento académico, estos autores decidieron utilizar como índice del tamaño del efecto la diferencia de medias estandarizada, en concreto, el índice *g* de Hedges (Cheng et al., 2019, pp. 803-805).

3.1.6. Métodos de síntesis e interpretación

Una vez extraídas las características de los estudios primarios y sus resultados (en forma de tamaños del efecto o de otra índole), el siguiente paso consiste en llevar a cabo la síntesis de dichos resultados. Para ello, se debe construir una base de datos informatizada donde las filas las conforman los estudios (normalmente, una fila por estudio) y las columnas son las características de los estudios (ej.,

tipo de diseño, edad media de la muestra, duración del programa, etc.) y los resultados (ej., diferencias de medias estandarizadas, valores *p* unilaterales, etc.). Cabe distinguir dos métodos de síntesis: síntesis de los tamaños del efecto, que es propia de los MAs, y otros métodos de síntesis, que es propia de las RSs sin aplicar MA.

3.1.6.1. Métodos de síntesis meta-analítica

Cuando se lleva a cabo un MA, el *método de síntesis estadística* involucra a los tamaños del efecto obtenidos en los estudios primarios. Para realizar una síntesis estadística de tamaños del efecto tienen que darse ciertas condiciones: (a) requiere disponer de un conjunto de estudios de los que se ha podido obtener al menos un índice del tamaño del efecto y su correspondiente varianza muestral; (b) el resultado de cada estudio se ha podido convertir en un mismo índice del tamaño del efecto (ej., todos los tamaños del efecto de los estudios son diferencias de medias estandarizadas, o todos son coeficientes de correlación, etc.) y (c) se requiere

que los estudios primarios exhiban una razonable homogeneidad en cuanto a sus características metodológicas y sustantivas.

Cuando se cumplen estas condiciones, el siguiente paso que tiene que dar el meta-analista es decidir el modelo estadístico que mejor se ajusta a las características de los estudios primarios del MA. Cabe distinguir dos modelos estadísticos en MA: el modelo de efecto fijo y el modelo de efectos aleatorios (Borenstein et al., 2019; Botella y Sánchez-Meca, 2015). El *modelo de efecto fijo* ('fixed-effect model') asume que existe un tamaño del efecto paramétrico (poblacional) común a todos los estudios, μ, de forma que cada estudio primario aporta un tamaño del efecto, T_i, que estima a dicho efecto paramétrico: $T_i = \mu + e_i$. Bajo este modelo, la única fuente de variación entre los tamaños del efecto de los estudios individuales es la debida al error de muestreo (e_i), es decir, al hecho de cada estudio utiliza una muestra de participantes diferente, aunque todas ellas sean representativas de la misma población de referencia. El *modelo*

de efectos aleatorios ('random-effects model') asume que el tamaño del efecto de cada estudio primario, T_i, estima a un efecto paramétrico diferente, μ_i, y que los efectos paramétricos son una muestra representativa de una población mayor de potenciales tamaños del efecto poblacionales, una distribución que tiene su correspondiente efecto medio, μ, y varianza, τ^2 (denominada 'varianza inter-estudios): $T_i = \mu_i + \varepsilon_i + e_i$. Bajo este modelo se asumen dos fuentes de variabilidad y no sólo una como ocurre en el modelo de efecto fijo: la variabilidad intra-estudio, debida al error de muestreo, e_i, y la variabilidad inter-estudios, ε_i, debida al muestreo de efectos paramétricos.

Las consecuencias de asumir uno u otro modelo afectan a los cálculos estadísticos y a la interpretación de los resultados del MA. Es por ello que el meta-analista debe justificar las razones de su

elección. El meta-analista debe elegir el modelo estadístico que, según las características de los estudios, parece que mejor se ajusta a la realidad. El modelo de efecto fijo sólo sería aplicable en un MA donde los estudios primarios son muy homogéneos entre sí en cuanto a sus características (ej., población de referencia, tipo de programa, tipo de diseño, etc.), de forma que dichos estudios pueden considerarse como réplicas idénticas el uno del otro. Sólo bajo esas circunstancias es realista asumir que la única fuente de variabilidad entre los tamaños del efecto es el error de muestro intra-estudio, e_i. Cuando los estudios primarios exhiben una clara heterogeneidad en sus características, entonces el modelo más realista es el de efectos aleatorios. Dado que en las ciencias sociales y de la salud no se suelen hacer réplicas idénticas de estudios previos, lo esperable es que los estudios incluidos en un MA, aun siendo parecidos, exhibirán una clara heterogeneidad en sus características. En consecuencia, el modelo más realista, en general, es el modelo de efectos aleatorios.

Tanto en el modelo de efecto fijo como en el de efectos aleatorios, los análisis estadísticos se realizan aplicando técnicas de ponderación, de forma que los estudios con mayor tamaño muestral (i.e., con menor varianza de error) ejerzan un mayor peso en los cálculos, en comparación con los estudios con tamaños muestrales pequeños (i.e., con mayor varianza de error). Pero el factor de ponderación es diferente dependiendo de qué modelo estadístico se ha elegido (Cooper et al., 2019; Egger et al., 2022).

Independientemente del modelo estadístico asumido, la síntesis estadística de un MA comienza con la construcción de un gráfico denominado 'forest plot', que presenta de forma numérica y visual el tamaño del efecto de cada estudio primario y su IC95% (Light y Pillemer, 1984). El forest plot es, pues, un gráfico que representa una especie de 'fotografía' de los resultados de los estudios incluidos en el MA y, por tanto, muy útil para ilustrar las magnitudes del efecto encontradas en los estudios y el grado en que varían entre sí. Junto con el forest plot, se calcula el tamaño del efecto

medio y un intervalo de confianza al 95% para estimar el efecto promedio en la población (Sánchez-Meca y Marín-Martínez, 2008).

Para comprobar si los tamaños del efecto de los estudios primarios son homogéneos o heterogéneos entre sí, se aplican varios procedimientos. Uno de ellos consiste en aplicar el estadístico *Q* de heterogeneidad. Si el estadístico *Q* alcanza un resultado estadísticamente significativo ($p < .05$), ello será evidencia de excesiva heterogeneidad entre los tamaños del efecto. El resultado del estadístico *Q* se suele complementar con el cálculo del índice I^2 de heterogeneidad, que aporta un valor entre 0 y 100% y se interpreta como el porcentaje de heterogeneidad relativa que exhiben los tamaños del efecto. Adicionalmente, cuando se asume un modelo de efectos aleatorios, otro estadístico que ayuda a valorar la existencia de heterogeneidad entre los tamaños del efecto es la estimación de la varianza inter-estudios, τ^2, o la desviación típica inter-estudios, τ. Por último, se recomienda también

construir un intervalo de predicción al 95% en torno al tamaño del efecto medio. Si bien el intervalo de confianza ofrece una estimación de entre qué valores se estima que se encontrará el efecto promedio poblacional, el intervalo de predicción ofrece una conjetura de entre qué valores se encontrará el tamaño del efecto si se realizara un nuevo estudio empírico sobre la pregunta de interés. Si el intervalo de predicción es muy ancho, ello será evidencia de que existe heterogeneidad entre los tamaños del

efecto y, en consecuencia, el tamaño del efecto esperado en un nuevo estudio será muy variable (IntHout et al., 2016).

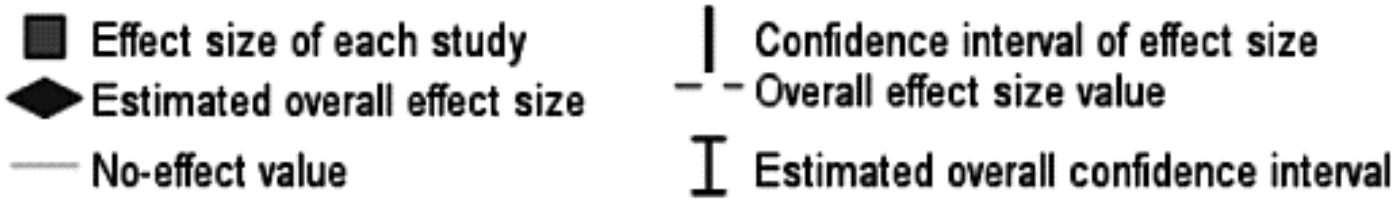
Como ejemplo, obsérvese en la Figura 1 un forest plot de los tamaños del efecto del MA de Patros et al. (2015) sobre diferencias en impulsividad de elección entre niños y adolescentes con TDAH y normales. El índice del tamaño del efecto utilizado en este MA

fue la diferencia de medias estandarizada, *g* de Hedges, y se calculó de forma que valores positivos indicaron una mayor impulsividad en las muestras TDAH en comparación con las muestras comunitarias. En la parte inferior del forest plot se recoge el tamaño del efecto medio y su intervalo de confianza al 95%: $g_+ = 0.47$, IC95% = 0.40 – 0.54. Según la guía orientativa de Cohen (1988), este tamaño del efecto refleja una relevancia práctica moderada (próximo a 0.50). El análisis de la heterogeneidad también queda recogido en dicho forest plot (véase parte inferior del gráfico). En primer lugar, el estadístico *Q* de heterogeneidad arrojó el valor $Q(27) = 29.97$, $p = .32$, un resultado que evidenció ausencia de heterogeneidad relevante entre los tamaños del efecto ($p > .05$). Además, el índice I^2 fue nulo, $I^2 = 0\%$, indicando un 0% de heterogeneidad relativa. Así mismo, la varianza interestudios fue nula: $\tau^2 = 0.00$. Todos estos resultados apuntan hacia la ausencia de heterogeneidad entre los tamaños del efecto de este MA. En este ejemplo, al ser nula la

varianza de heterogeneidad, el intervalo de predicción al 95% coincide con el intervalo de confianza arriba descrito.

Figura 1.

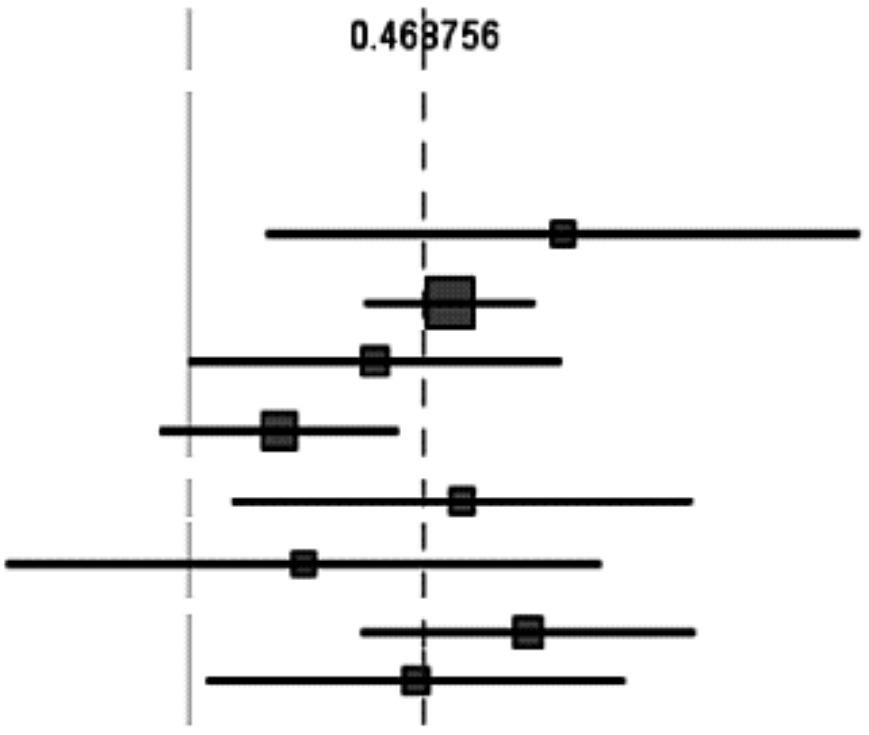
Forest plot de los 28 tamaños del efecto del MA de Patros et al. (2015) sobre impulsividad de elección y TDAH en niños y adolescentes. Este forest plot se ha construido con el módulo de MA que incorpora el programa IBM SPSS 28 (fuente: elaboración propia).

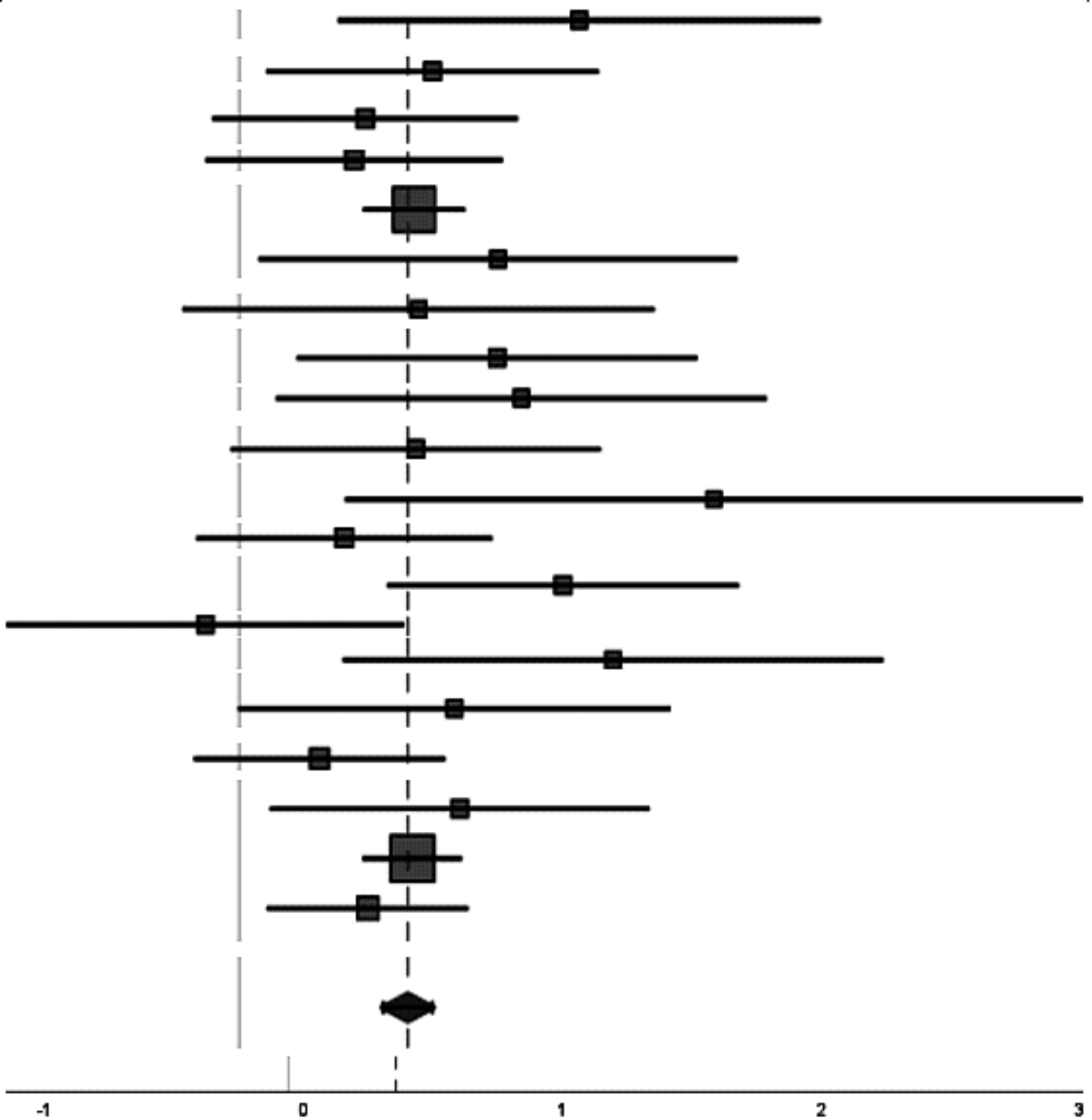


ID	Effect Size	Std. Error	Lower	Upper
Antrop et al (2006)	0.75	0.30	0.15	1.34
Banaschewski et al (2012)	0.52	0.09	0.35	0.69
Barkley et al (2001)	0.37	0.19	-0.00	0.74
Bidwell et al (2007)	0.18	0.12	-0.06	0.42
Bitsakou et al (2009) ...	0.55	0.23	0.09	1.00
Bitsakou et al (2009) ...	0.23	0.30	-0.36	0.82
Coghill et al (2014)	0.68	0.17	0.34	1.01
Costa Dias et al (2013)	0.45	0.21	0.04	0.87
Dalen et al (2004)	0.94	0.34	0.27	1.62
Demurie et al (2013)	0.54	0.24	0.07	1.00
Gawrilow et al (2011)	0.35	0.22	-0.07	0.77
Karalunas et al (2011)	0.32	0.21	-0.09	0.73

Kuntsi et al (2010)	0.49	0.07	0.34	0.63
Max et al (2010) 8-12 ...	0.72	0.34	0.05	1.38
Max et al (2010) 13-16...	0.50	0.33	-0.16	1.15
Metin et al (2013)	0.72	0.28	0.16	1.27
Paloyelis et al (2010)	0.78	0.35	0.10	1.47
Scheres et al (2010)	0.49	0.26	-0.02	1.00
Schweitzer et al (1995)	1.32	0.52	0.29	2.34
Sjöwall et al (2013)	0.29	0.21	-0.12	0.70
Solanto et al (2001)	0.90	0.25	0.41	1.39
Solanto et al (2007)	-0.09	0.28	-0.65	0.46
Sonuga-Barke et al (1992)	1.04	0.38	0.29	1.79
Vloet et al (2010)	0.60	0.31	-0.00	1.20
Wahlstedt et al (2009)	0.22	0.18	-0.13	0.57
Wilson et al (2011)	0.61	0.27	0.09	1.14
Wood et al (2011)	0.48	0.07	0.34	0.62
Yang et al (2011)	0.36	0.14	0.08	0.64
Overall	0.47	0.03	0.40	0.54

Forest Plot





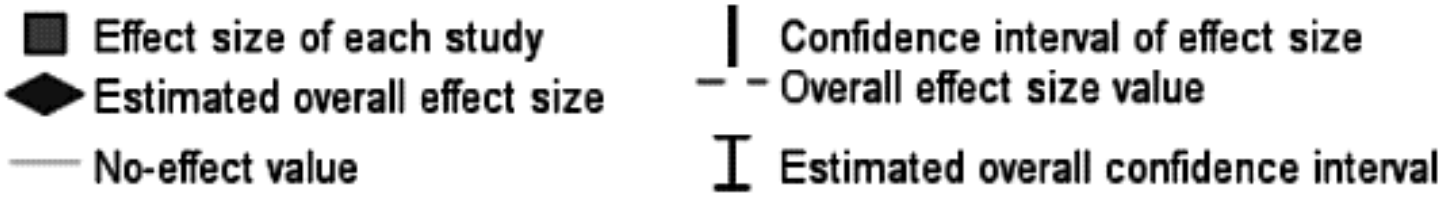
Model: Random-effects model
Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00
Homogeneity: Q = 29.97, df = 27, p-value = 0.32
Test of overall effect size: t = 14.15, df = 27, p-value = 0.00

Veamos otro ejemplo en el que se observa una clara heterogeneidad entre los tamaños del efecto. En el MA de Erion (2006) sobre la eficacia de los programas de tutorización parental en casa para mejorar el rendimiento escolar de sus hijos, este autor integró los tamaños del efecto de 32 estudios utilizando la diferencia de medias estandarizada como índice del tamaño del efecto (*g* de Hedges). En la Figura 2 se presenta un forest plot de sus resultados. Se observó un efecto medio $g_+ = 0.73$, IC95% = 0.43 – 1.02. Según la guía de Cohen (1988), este efecto medio reflejaría una relevancia práctica entre moderada y alta (entre 0.50 y 0.80). El estadístico de heterogeneidad alcanzó la significación estadística, $Q(31) = 117.17$, $p < .001$, el índice I^2 fue de alta magnitud, $I^2 = 81\%$ y la varianza inter-estudios alcanzó el valor $\tau^2 = 0.41$ ($\tau = 0.64$). Además, el intervalo de predicción al 95% fue tan ancho que incluso incluyó tamaños del efecto negativos: IP95% = -0.61 – 2.06. Esto significa que un futuro estudio que pusiera a prueba la eficacia de los programas de

tutorización parental podría alcanzar un resultado tanto favorable como desfavorable a su eficacia.

Figura 2.

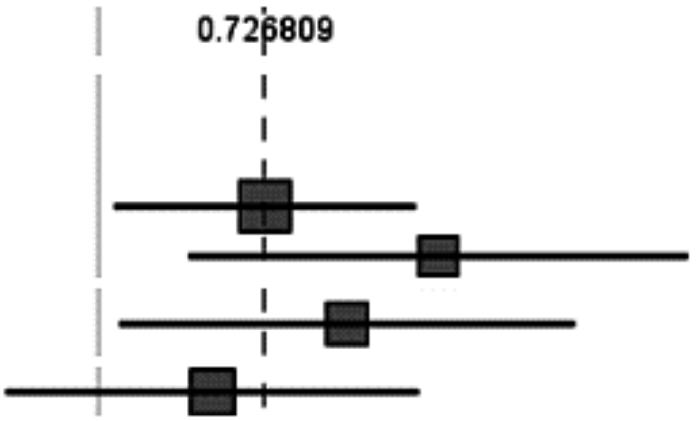
Forest plot de los 32 tamaños del efecto del MA de Erion (2006) sobre la eficacia de los programas de tutorización parental para mejorar el rendimiento escolar de sus hijos. Este forest plot se ha construido con el módulo de MA que incorpora el programa IBM SPSS 28 (fuente: elaboración propia).

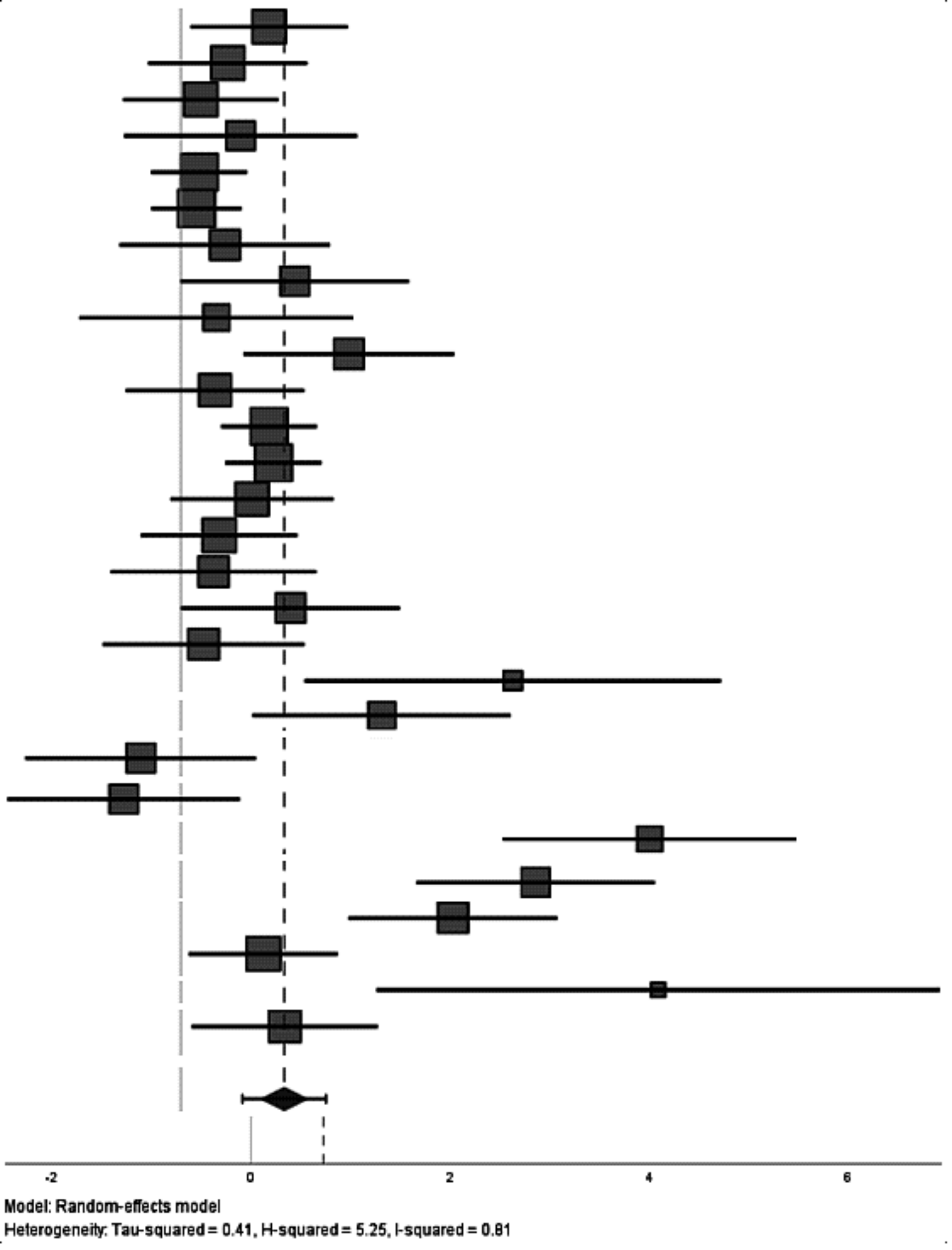


ID	Effect Size	Std. Error	Lower	Upper
Bickerton-Ross (1988)	0.73	0.34	0.07	1.39
Casey (1987) 1	1.49	0.56	0.40	2.58
Casey (1987) 2	1.09	0.51	0.09	2.09
Casey (1987) 3	0.50	0.46	-0.40	1.40
Colvin (1980) 1	0.62	0.28	0.07	1.17
Colvin (1980) 2	0.33	0.28	-0.23	0.89
Ellis (1996)	0.14	0.28	-0.40	0.68
Erion (1994)	0.42	0.42	-0.40	1.24
Fox (1982)	0.13	0.17	-0.21	0.47
Hannon (1987)	0.11	0.16	-0.20	0.42
Izzo (1976) 1	0.31	0.38	-0.43	1.05
Izzo (1976) 2	0.80	0.41	0.00	1.60
Jacobwitz (1979)	0.25	0.49	-0.71	1.21
Johnson (1981) 1	1.18	0.38	0.44	1.92

Johnson (1981) 2	0.24	0.32	-0.39	0.87
Mahoney (1985) 1	0.62	0.17	0.28	0.96
Mahoney (1985) 2	0.65	0.17	0.31	0.99
Mehran & White (1988)	0.50	0.29	-0.07	1.07
Meteyer (1998)	0.27	0.28	-0.28	0.82
Miller & Narrett (1995) 1	0.23	0.37	-0.49	0.95
Miller & Narrett (1995) 2	0.77	0.39	0.00	1.54
Miller & Narrett (1995) 3	0.16	0.36	-0.55	0.87
Minner (1989)	2.33	0.75	0.87	3.79
Nielson (1991)	1.41	0.46	0.50	2.32
Powell-Smith et al. (2...	-0.28	0.41	-1.09	0.53
Powell-Smith et al. (2...	-0.40	0.42	-1.22	0.42
Reagal & Elliott (1971) 1	3.29	0.52	2.26	4.32
Reagal & Elliott (1971) 2	2.49	0.43	1.65	3.33
Reagal & Elliott (1971) 3	1.91	0.37	1.18	2.64
Reagal & Elliott (1971) 4	0.58	0.27	0.06	1.10
Stevenson (2001)	3.35	1.01	1.37	5.33
Vinograd (1986)	0.73	0.33	0.08	1.38
Overall	0.73	0.14	0.43	1.02

Forest Plot




-2
0
2
4
6
Model: Random-effects model
Heterogeneity: Tau-squared = 0.41, H-squared = 5.25, I-squared = 0.81

Homogeneity: Q = 117.17, df = 31, p-value = 0.00
Test of overall effect size: t = 5.07, df = 31, p-value = 0.00

El siguiente paso en los análisis estadísticos típicos de un MA consiste en valorar la posible existencia de sesgo publicación. Con este propósito, puede construirse un gráfico denominado 'funnel plot', que presenta en un diagrama de dispersión la relación existente entre los tamaños del efecto y sus errores estándar (Light y Pillemer, 1984). Si el funnel plot exhibe la forma de un

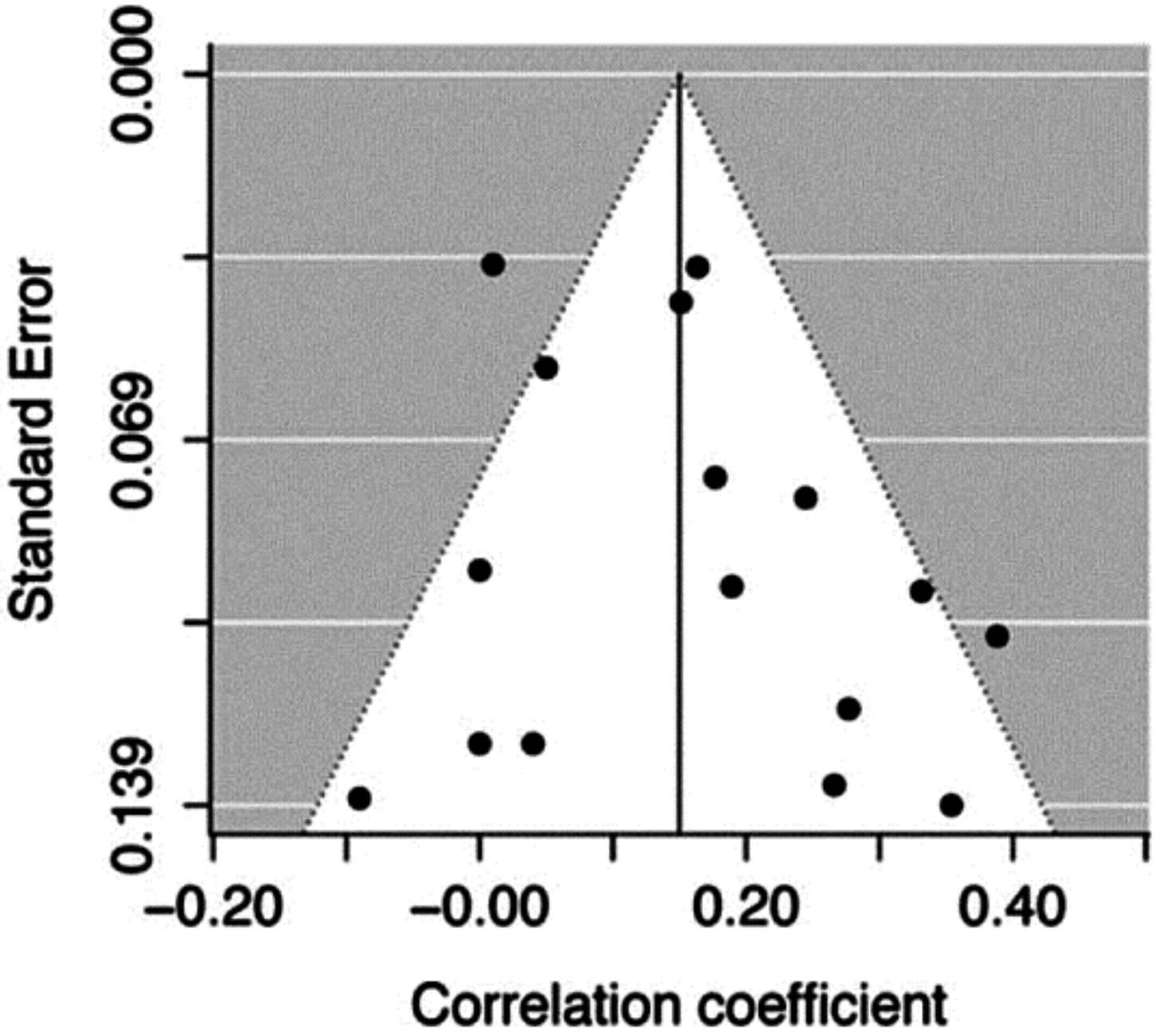
'embudo invertido' o 'triángulo', entonces ello sería evidencia de que el sesgo de publicación no es una amenaza contra la validez de los resultados del MA. Si, por el contrario, el funnel plot presenta ausencia de efectos en la zona baja del gráfico y de los valores contrarios a la hipótesis predominante en el campo, entonces ello sería evidencia de sesgo de publicación y, por tanto, de una posible sobreestimación del

verdadero fenómeno estudiado en la población. La Figura 3 presenta dos ejemplos de funnel plot. El etiquetado como (A) exhibe ausencia de asimetría, lo que se puede interpretar como que el sesgo de publicación no es una amenaza contra la validez de los resultados de ese MA. El etiquetado como (B) exhibe una clara asimetría, provocada por la ausencia de efectos en la zona baja del gráfico y, en concreto, en los tamaños del efecto negativos o contrarios a la hipótesis preponderante, y con tamaños muestrales pequeños. En este segundo caso, el funnel plot evidencia que los resultados del MA pueden verse afectados por un problema de sesgo de publicación, lo que se traduce principalmente en que el efecto medio del MA está sobreestimando el verdadero efecto en la población.

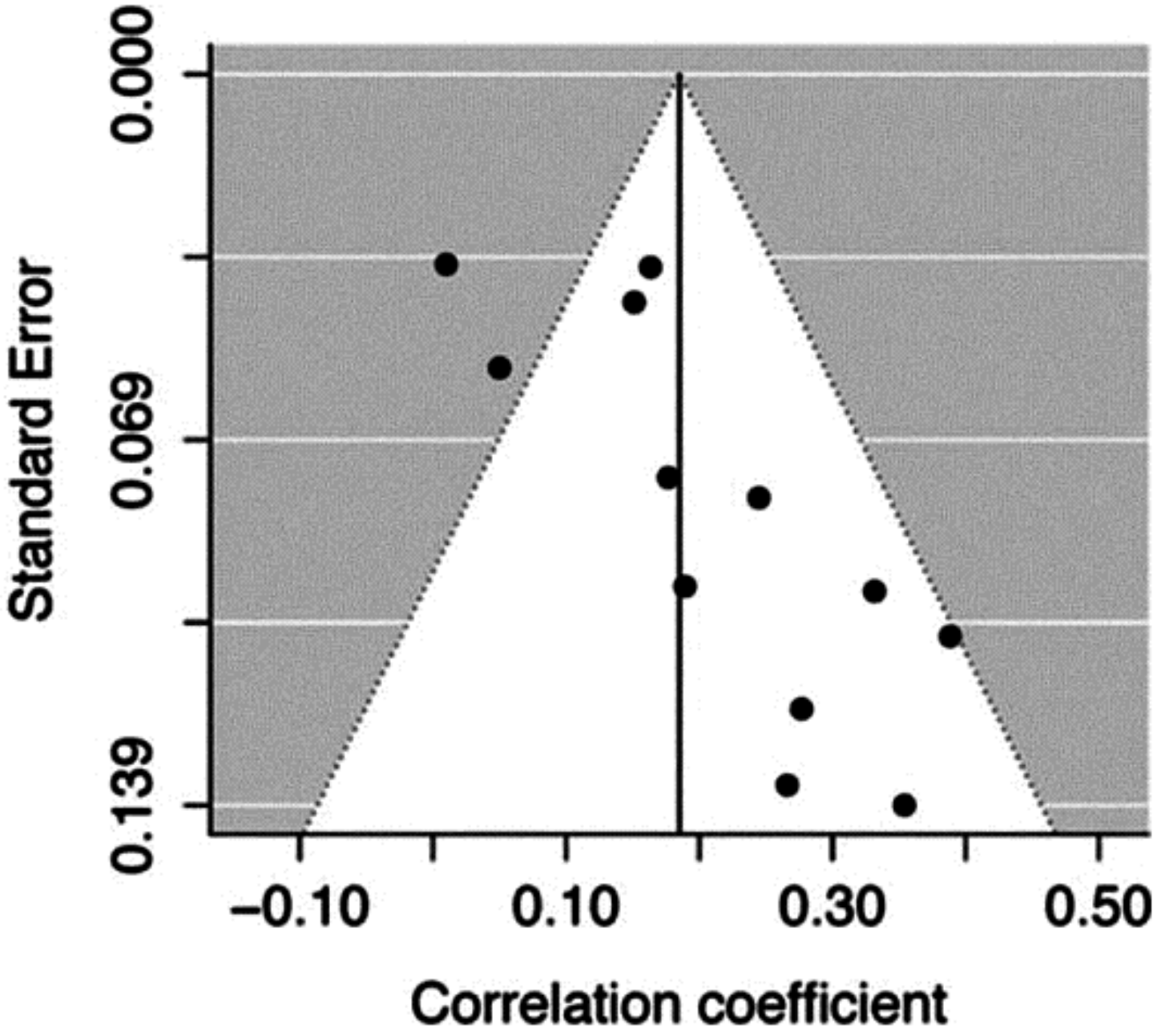
Figura 3.

Ejemplos de funnel plot: (A) funnel plot que refleja ausencia de una clara asimetría, indicando que el sesgo de publicación no es una amenaza a la validez de los resultados del MA, y (B) funnel plot con una clara asimetría provocada por la ausencia de efectos negativos o contrarios a la hipótesis y con tamaños muestrales bajos.

A


Standard Error
0.000
0.069
0.139
–0.20
–0.00
0.20
0.40
Correlation coefficient

B



Junto con la construcción de un funnel plot, es aconsejable aplicar alguna prueba estadística que permita tomar una decisión sobre si los tamaños del efecto de un MA exhiben un problema de sesgo de publicación. El test de Egger cumple ese objetivo. Este test contrasta la significación estadística de la

intercepción (b_0) de un modelo de regresión simple (no ponderada), tomando como predictor los errores estándar de los tamaños del efecto, y como variable dependiente los propios tamaños del efecto (debidamente estandarizados). Un resultado estadísticamente no significativo de la intercepción de este modelo ($p > .05$) evidencia ausencia de sesgo de publicación, mientras que un resultado significativo ($p < .05$) indicaría asimetría del funnel plot y, en consecuencia, evidencia de sesgo de publicación (Egger et al., 1997).[7]

Cuando el aspecto del funnel plot revela cierta asimetría y el test de Egger alcanza la significación estadística, ello evidencia que el tamaño del efecto medio del MA puede estar sobreestimando

---

[7]Cuando el número de estudios del MA es pequeño ($k < 20$), entonces se aconseja utilizar como nivel de significación el 10% en lugar del habitual 5% ($p < .10$ en lugar de $p < .05$), para contrarrestar la baja potencia estadística que exhibe el test de Egger bajo estas condiciones.

el verdadero efecto en la población. En estas circunstancias, se recomienda aplicar el método 'trim-and-fill' de Duval y Tweedie (2000) para 'simetrizar' el aspecto del funnel plot imputando tamaños del efecto adicionales, supuestamente no publicados (o no localizados por el metaanalista), con resultados contrarios a la hipótesis dominante en el campo. El tamaño del efecto medio ajustado mediante la adición de tamaños del efecto 'imputados' al MA arrojará un valor inferior al del tamaño del efecto medio original. Si el efecto medio ajustado es muy inferior al del efecto medio original (ej., es un 10%, o más, inferior al original), entonces ello sería evidencia de sesgo de publicación en la estimación del verdadero efecto poblacional. En estos casos, se recomienda aportar el efecto medio ajustado como una estimación más adecuada del

efecto poblacional, en lugar del efecto medio original. Otras técnicas y métodos adicionales para comprobar el sesgo de publicación en los MAs pueden consultarse en Rothstein et al. (2005) y en Vevea et al. (2019).

Para ilustrar la técnica 'trim-and-fill' utilizamos el MA de Rosa-Alcázar et al. (2008) sobre la eficacia de los tratamientos psicológicos del trastorno obsesivo-compulsivo. Este MA integró los tamaños del efecto (diferencias de medias estandarizadas, *g* de Hedges) de 24 estudios, obteniendo un tamaño del efecto medio $g_+ = 1.07$, con IC95% = 0.84 – 1.31, de elevada magnitud y estadísticamente significativo. El test de Egger reveló un resultado estadísticamente significativo: $b_0 = 1.801$, $t(22) = 2.282$, $p = .032$, indicando la existencia de sesgo de publicación. La Figura 4 presenta un funnel plot de los tamaños del efecto de este MA. La imputación de 4 tamaños del efecto, mediante la técnica 'trim-and-fill', para simetrizar el aspecto del funnel plot dio lugar a un tamaño del efecto medio ajustado $g_{aj} = 0.92$ (IC95% = 0.66 – 1.18). El descenso

del efecto medio ajustado respecto del efecto medio original fue del 14%: (0.92 – 1.07)/1.07 = -0.14. Al ser superior al 10% dicho descenso, es recomendable reportar el efecto medio ajustado en lugar del original como estimación del verdadero efecto en la población de los tratamientos psicológicos sobre este trastorno.

Cuando los tamaños del efecto de un MA exhiben heterogeneidad, se hace preciso investigar las fuentes de dicha heterogeneidad. Para ello, se ponen en relación las características de los estudios con los tamaños del efecto. Cuando la característica, o variable moderadora, es categórica, se aplican 'análisis de subgrupos' ('subgroup analyses'), que equivalen a ANOVAs especiales, ya que implican ponderar cada tamaño del efecto en función de su precisión. Cuando la variable moderadora es continua se aplican modelos de meta-regresión ('meta-regression models'), que son el equivalente a modelos de regresión linear ponderada (Botella y SánchezMeca, 2015; Rubio-Aparicio et al., 2020).

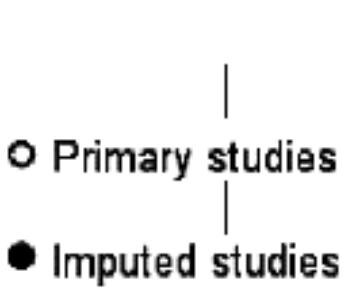
Para variables moderadoras categóricas (ej., tipo de

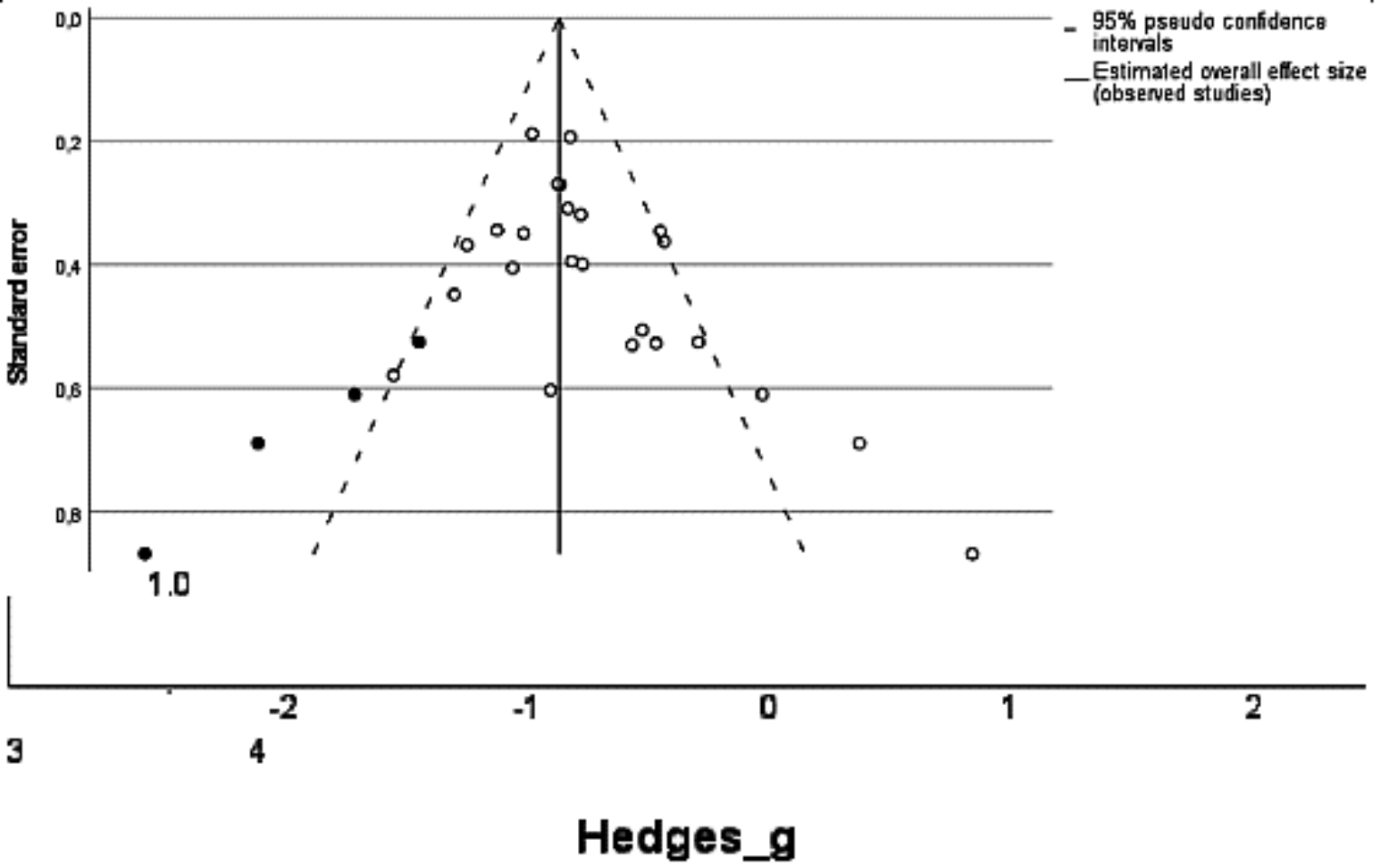
tratamiento, tipo de diseño), el análisis de subgrupos implica obtener el tamaño del efecto medio (y su IC95%) para cada categoría del moderador, comprobar la significación estadística del moderador (con el estadístico Q_B), estimar la proporción de varianza explicada por el moderador (con el índice R^2) y comprobar la especificación del modelo (con el estadístico Q_W). Como ejemplo, la Tabla 1 presenta los resultados del análisis de subgrupos de la variable moderadora 'duración del programa' en el MA de Erion (2006) sobre la eficacia de los programas de tutorización parental. Aunque la duración del programa es una variable continua, el autor de este MA la clasificó en tres categorías (corta, media y larga). Los efectos medios obtenidos para las tres duraciones fueron g_+ = 0.46, 0.69 y 1.15 para duraciones corta, media y larga, respectivamente. Se observó un incremento de la eficacia cuanto mayor era la duración del programa, alcanzando una significación estadística

marginal dicho efecto: $Q_B(2) = 5.50$, $p = .064$ (i.e., $p < .10$), con un 10% de varianza explicada ($R^2 = 0.10$), si bien el modelo estaba mal especificado: $Q_W(29) = 111.48$, $p < .001$. Por tanto, la duración del programa fue una variable moderadora relevante, que explicó el 10% de la heterogeneidad de los tamaños del efecto.

Figura 4.

Funnel plot con la técnica 'trim-and-fill' de tamaños del efecto imputados aplicada al MA de Rosa-Alcázar et al. (2008) sobre la eficacia del tratamiento psicológico del trastorno obsesivo-compulsivo. Los puntos negros representan los 4 tamaños del efecto 'imputados' con esta técnica. Los puntos blancos representan los tamaños del efecto de los 24 estudios originales (gráfico realizado con el programa IBM SPSS 24; fuente: elaboración propia).





Para variables moderadoras categóricas (ej., tipo de tratamiento, tipo de diseño), el análisis de subgrupos implica obtener el tamaño del efecto medio (y su IC95%) para cada categoría del moderador, comprobar la significación estadística del moderador (con el estadístico Q_B), estimar la proporción de varianza explicada por el moderador (con el índice R^2) y comprobar la especificación del modelo (con el estadístico Q_W). Como ejemplo, la Tabla 1 presenta los resultados del análisis de

subgrupos de la variable moderadora 'duración del programa' en el MA de Erion (2006) sobre la eficacia de los programas de tutorización parental. Aunque la duración del programa es una variable continua, el autor de este MA la clasificó en tres categorías (corta, media y larga). Los efectos medios obtenidos para las tres duraciones fueron g_+ = 0.46, 0.69 y 1.15 para duraciones corta, media y larga, respectivamente. Se observó un incremento de la eficacia cuanto mayor era la duración del programa, alcanzando una significación estadística marginal dicho efecto: $Q_B(2) = 5.50$, $p = .064$ (i.e., $p < .10$), con un 10% de varianza explicada ($R^2 = 0.10$), si bien el modelo estaba mal especificado: $Q_W(29) = 111.48$, $p < .001$. Por tanto, la duración del programa fue una variable moderadora relevante, que explicó el 10% de la heterogeneidad de los tamaños del efecto.

Tabla 1.

Resultados del análisis de subgrupos (ANOVA) para la duración del programa de tutorización parental sobre los tamaños del efecto en el MA de Erion (2006).

Duración del programa	*k*	g_+	IC al 95% Li	IC al 95% Ls
Corta	12	0.46	0.26	0.66
Media	10	0.69	0.27	1.11
Larga	10	1.15	9.58	1.71

Resultados del ANOVA: $Q_B(2) = 5.50$, $p = .064$; $R^2 = 0.10$

$Q_W(29) = 111.48$, $p < .001$

k = número de estudios. g_+ = tamaño del efecto medio (*g* de Hedges). Li y Ls = límites confidenciales inferior y superior del IC al 95%, respectivamente. Q_B = prueba de significación de la diferencia entre los efectos medios de las categorías. R^2 = proporción de varianza explicada por el moderador. Q_W = prueba de especificación del modelo (análisis realizados con el módulo de MA del programa IBM SPSS 28; fuente: elaboración propia).

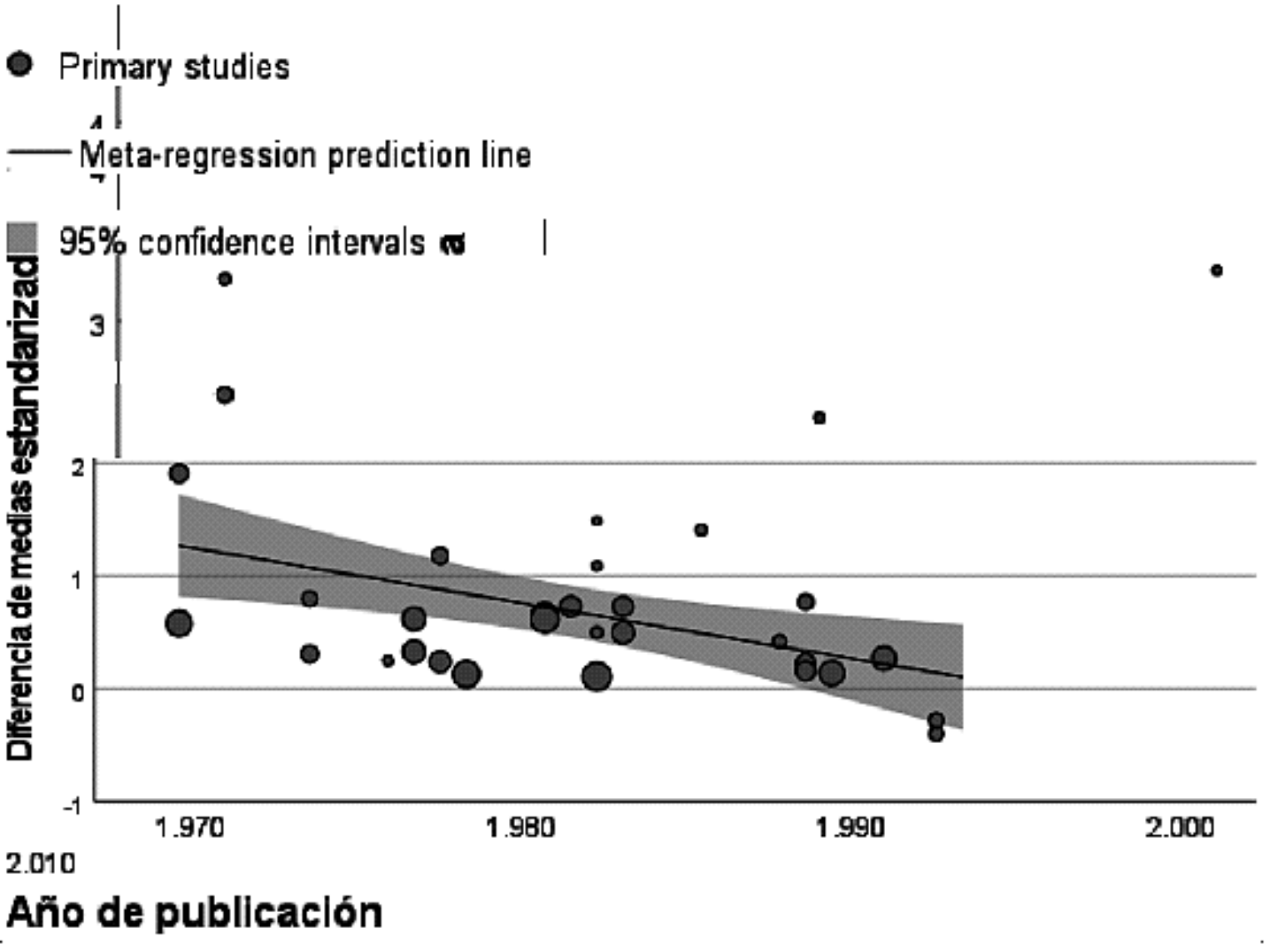
El análisis de moderadores continuos se lleva a cabo mediante modelos de meta-regresión,

tomando el moderador como variable predictora y los tamaños del efecto como variable dependiente (López-López et al., 2014). Los principales resultados de este análisis consisten en reportar el coeficiente de regresión (b_j, cuyo signo revela el sentido de la relación entre el moderador y los tamaños del efecto), la significación estadística del moderador (con el estadístico *t* o con el estadístico Q_R), la proporción de varianza explicada por el moderador (con el índice R^2) y la prueba de especificación del modelo (con el estadístico Q_E). Como ejemplo, la Figura 5 presenta un diagrama de dispersión ('bubble plot') de la relación entre el año de realización del estudio y los tamaños del efecto en el MA de Erion (2006) sobre la eficacia de los programas de tutorización parental. Se observa en el gráfico un descenso de los tamaños del efecto en los estudios más recientes, respecto de los más antiguos. Esta relación queda cuantificada con la obtención de un coeficiente de regresión de signo negativo (b_j = -0.039), que alcanzó la significación estadística: $t(30) = -2.98$, $p = .006$,

explicó un 36.9% de la varianza de los tamaños del efecto (R^2 = 0.369), si bien el modelo estaba mal especificado: $Q_E(30) = 95.98$, $p < .001$.

Figura 5.

Diagrama de dispersión ('bubble plot') de los tamaños del efecto en función del año de realización del estudio en el MA de Erion (2006) sobre la eficacia de los programas de tutorización parental (gráfico realizado con el programa IBM SPSS 28; fuente: elaboración propia).



3.1.6.2. Otros métodos de síntesis

Son varias las razones que se pueden dar para no aplicar una síntesis meta-analítica: (a) por la existencia de excesiva heterogeneidad metodológica (ej., diferentes tipos de grupos de control, diferentes tipos de diseños, diferentes medidas de resultado), estadística (i.e., presencia de heterogeneidad entre los tamaños del efecto que no puede explicarse por el influjo de ninguno de los moderadores analizados) o sustantiva (ej., poblaciones de referencia muy diferentes, tratamientos muy distintos) entre los estudios; (b) por no disponer de los tamaños del efecto de los estudios o, disponiendo de éstos, no se dispone de sus varianzas muestrales; ó (c) por sospecha de fuerte sesgo de publicación. Cuando no se cumplen las condiciones necesarias para realizar un MA,

entonces se pueden aplicar otros métodos de síntesis, si bien estos métodos son menos eficientes y rigurosos que los del MA (McKenzie y Brennan, 2022).

En todos estos métodos se recomienda presentar la información extraída de los estudios (características y resultados) mediante tablas en las que se muestren los datos de cada estudio individual (ej., identificador del estudio, riesgo de sesgo, tamaño muestral, tipo de diseño, población de referencia, método de medida de la variable de resultado, tamaño del efecto, dirección del efecto, valor *p*, etc.). Dependiendo del método elegido, se pueden utilizar representaciones gráficas (ej., forest plot, box plots, stem-and leaf displays, etc.) y pruebas de significación estadística (ej., el método de Fisher de acumulación de valores de probabilidad, *p*, la prueba de los signos, etc.).

Si se dispone de los tamaños del efecto y de sus varianzas muestrales, una estrategia alternativa a la síntesis meta-analítica consiste en reportar los tamaños del efecto de los estudios individuales con

sus intervalos de confianza al 95%, por ejemplo, mediante un forest plot, pero sin reportar el tamaño del efecto medio ni realizar otros análisis estadísticos antes descritos, que son propios de la síntesis meta-analítica. Si se dispone de los tamaños del efecto, pero no de sus varianzas muestrales (ej., debido a falta de información en los estudios), se pueden reportar los tamaños del efecto en una tabla sin ofrecer los intervalos de confianza, pero junto con las características principales de los estudios. Opcionalmente, se pueden calcular medias o medianas y reportar los tamaños del efecto mínimo y máximo.

Si de cada estudio no disponemos del tamaño del efecto, pero sí del valor de probabilidad exacto, *p*, asociado al resultado de la prueba de contraste de hipótesis objeto de interés, puede aplicarse el método de Fisher de acumulación de valores de probabilidad. No obstante, los valores *p* reportados en los estudios (que suelen ser bilaterales), deben transformarse a valores *p* unilaterales antes de aplicar dicho método de acumulación de valores

p.

Cuando no disponemos de los tamaños del efecto, pero sí de la dirección del efecto encontrado en cada estudio, es posible definir una métrica estandarizada dicotómica consistente en clasificar cada estudio en una de dos categorías: (a) efecto a favor de la hipótesis objeto de estudio y (b) efecto en contra de dicha hipótesis. También puede utilizarse cuando los tamaños del efecto de los estudios están en métricas diferentes, de forma que no es posible transformarlos a una métrica común. Sobre los resultados dicotómicos del efecto observado en cada estudio, se puede aplicar la prueba de los signos, con objeto de comprobar si es más probable un resultado favorable o desfavorable a la hipótesis de interés. Es muy importante tener en cuenta que en este método la categorización de la dirección del efecto no se basa en el hecho de que el valor *p* haya alcanzado o no la significación estadística (ej., $p < .05$), sino en el hallazgo de un resultado

favorable o desfavorable a la hipótesis de interés (independientemente de que haya alcanzado o no la significación estadística).

Todos estos métodos de síntesis alternativos a la síntesis meta-analítica deben ir acompañados de una síntesis narrativa que describa y explique los resultados obtenidos. Es posible encontrar RSs que sólo han aplicado una síntesis narrativa sin apoyarse en ninguno de los métodos de síntesis alternativos aquí descritos. En ocasiones, la síntesis narrativa se aplica basándose en los resultados de los contrastes de hipótesis, resaltando aquéllos que obtuvieron un valor *p* estadísticamente significativo a favor, o en contra, de la hipótesis de interés. Esta estrategia no es recomendable. Es preferible basarse en la dirección del efecto (independientemente de si el resultado fue estadísticamente significativo o no) y aplicar la prueba de los signos, tal como se describe más arriba.

Al igual que con los métodos de síntesis anteriores,

la síntesis narrativa debe ir acompañada de tablas en las que se recojan las principales características de los estudios (ej., población de referencia, edad media de la muestra, distribución por género, tipo de diseño, tamaño muestral, calidad metodológica o riesgo de sesgo, etc.) y los resultados (valores *p*, dirección del efecto, medias, etc.).

El uso de la síntesis narrativa como único recurso para integrar los resultados de los estudios de una RS es el método menos aceptable, o recomendable, ya que es propenso a sufrir sesgos subjetivos en la interpretación de los mismos (Campbell et al., 2019).

3.1.7. Publicación

La redacción de una RS o un MA sigue los mismos apartados que la de un estudio empírico: Introducción, Método, Resultados y Discusión/Conclusiones (Cooper, 2016; Page et al., 2021; Rubio-Aparicio et al., 2018; Sánchez Martín et al., 2022; Wilson y Grant, 2019). La *Introducción* debe recoger la misma información que en cualquier investigación: una fundamentación del

problema objeto de estudio, definir conceptual y operativamente los constructos, conceptos y variables implicados en la pregunta de interés, revisar la literatura previa de forma narrativa y formular la pregunta, los objetivos y, en su caso, las hipótesis objeto de estudio.

El apartado de *Método* presenta ciertas peculiaridades que lo hacen diferente de los apartados típicos de un estudio empírico. En primer lugar, deben especificarse los criterios que los estudios tenían que cumplir para ser incluidos en la RS/MA y, en su caso, los criterios de exclusión. En segundo lugar, se deben describir los procedimientos de búsqueda de los estudios, cuáles fueron las fuentes formales (ej., bases electrónicas consultadas) y, en su caso, informales, especificando la secuencia de palabras-clave utilizadas en las búsquedas. Debe también presentarse un diagrama de flujo que describa el resultado del proceso de búsqueda, cribado y selección de los estudios. En tercer lugar, debe describirse el proceso de extracción de las características de los estudios, así como los resultados del análisis de la fiabilidad de dicho

proceso, en términos de grado de acuerdo inter-codificadores. En cuarto lugar, debe describirse cómo se extrajeron los resultados de los estudios relativos a la pregunta de interés: mediante índices del tamaño del efecto, en cuyo caso debe especificarse cuál fue el índice del tamaño del efecto utilizado (ej., diferencia de medias estandarizada, coeficiente de correlación, etc.) o mediante otros métodos (ej., valores de probabilidad unilaterales, dirección del efecto, de forma narrativa). En quinto lugar, deben hacerse explícitos los métodos de síntesis de los resultados: mediante una síntesis

estadística de los tamaños del efecto, en el caso de un MA, o mediante otros métodos de síntesis, en el caso de una RS. En el caso de una síntesis meta-analítica, debe especificarse el programa de software estadístico aplicado (ej., metafor, IBM SPSS 28,

JAMOVI, JASP, CMA 3.3, RevMan 5.5, MetaWin, etc.).
La sección de *Resultados* debe presentar una descripción de las principales características de los estudios incluidos en la RS/MA, seguido de los resultados de la síntesis estadística, para el caso de un MA, o de otros métodos de síntesis, para el caso de una RS. Los resultados deben presentarse mediante el uso de tablas, gráficos (ej., forest plots, funnel plots, etc.) que ayuden a describir los principales resultados de la síntesis.
El apartado de *Discusión* y *Conclusiones* no difiere del de un estudio empírico. Los resultados de la síntesis deben ponerse en relación con los objetivos o con las preguntas objeto de interés. También deben discutirse los resultados obtenidos con los de otras investigaciones (ej., otras RSs u otros MAs previos similares) o con los modelos teóricos explicativos del fenómeno de interés. En la Discusión debe reflexionarse sobre las implicaciones prácticas que tienen los resultados de la RS/MA para la práctica profesional (educativa, social, clínica, etc.). También debe hacerse un ejercicio de autocrítica y revelar las limitaciones o dificultades de la RS/MA. Por último,

deben plantearse orientaciones para la investigación futura en el campo.

En la sección de *Referencias* se deben incluir las referencias bibliográficas de los estudios empíricos incluidos en la RS/MA. Para facilitar su identificación, lo habitual es preceder con un asterisco dichas referencias. Finalmente, se deben incorporar tablas con toda la información extraída de cada estudio de la RS/MA, mediante la incorporación de *anexos*, *apéndices* o, en su caso, *archivos suplementarios*.

3.2. Lectura crítica de revisiones sistemáticas / meta-análisis

Del mismo modo que se han propuesto en la literatura científica numerosos listados, checklists y guías orientativas sobre cómo debe reportarse un estudio empírico, así también se han propuesto para la correcta redacción de RSs y MAs. Los checklists y recomendaciones para la redacción de RSs y MAs cumplen una doble función (Pigott y Polanin, 2019; Rubio-Aparicio et al., 2018). Por una parte, ayudan a

la correcta redacción del informe escrito de una investigación de este tipo, con objeto de que el revisor o meta-analista no olvide reportar ningún aspecto relevante. Por otra parte, los checklists de reporte de RSs y MAs pueden utilizarse para hacer lectura crítica de alguna RS o MA publicada por otros y comprobar si la RS/MA está bien hecha o reportada, o si presenta deficiencias que pueden comprometer la validez de sus resultados.

La organización PRISMA ha desarrollado un amplio número de checklists para el correcto reporte de RSs y MAs (puede consultarse el sitio web de esta organización en el apartado 'Enlaces'). Así, para una adecuada redacción del Protocolo de una RS/MA, se puede utilizar el checklist PRISMAP (Moher et al., 2015). Para MAs sobre la eficacia de intervenciones puede utilizarse el checklist PRISMA 2020 (**P**referred **R**eporting **I**tems for **S**ystematic Reviews and **M**eta-**A**nalyses; Page et al., 2021). Una descripción de cómo aplicar el checklist PRISMA 2020 en el ámbito educativo puede consultarse en Sánchez-Serrano et al. (2022). Para la correcta redacción del Abstract de una RS/MA puede utilizarse el checklist PRISMA-A

(Page et al., 2021). La organización PRISMA también ha desarrollado checklists para otros tipos específicos de RSs/MAs, tales como para MA sobre la precisión de pruebas diagnósticas (PRISMA-DTA; Salameh et al., 2020), MA en red ('network

meta-analysis'; PRISMA-NMA; Hutton et al., 2015) o MA con datos de participantes individuales ('individual participant data meta-analysis'; PRISMA-IPD; Stewart et al., 2015).

Otro checklist, alternativo al checklist PRISMA 2020, elaborado para valorar la calidad de los MAs sobre la eficacia de intervenciones, es el checklist AMSTAR-2 (A MeaSurement Tool to Assess systematic Reviews-2; Shea et al., 2017; véase el sitio web en el apartado 'Enlaces'), que está más enfocado a valorar calidad metodológica que calidad del reporte de una RS o un MA. En el caso de que la RS o el MA en cuestión no sea sobre la eficacia de

intervenciones, sino que sintetiza estudios de naturaleza observacional, correlacional o asociativa, es muy recomendable el uso del checklist MOOSE (Meta-analysis of ObservatiOnal Studies in Epidemiology; Stroup et al., 2008). Para el reporte o lectura crítica de un MA de generalización de la fiabilidad, se recomienda utilizar el checklist REGEMA (REliability GEneralization Meta-Analysis; Sánchez-Meca et al., 2021; véase el sitio web en el apartado 'Enlaces'). Para el correcto reporte de una RS que no es un MA, es decir, que aplica otros métodos de síntesis diferentes al de la síntesis meta-analítica, puede utilizarse el checklist SWIM (Synthesis WIthout Meta-analysis; Campbell et al., 2020; véase el sitio web en el apartado 'Enlaces'), si bien éste debe aplicarse en combinación con el checklist PRISMA 2020.

Por último, cabe mencionar la Colaboración Campbell ('Campbell Collaboration'), una organización colaborativa sin ánimo de lucro que aglutina a los principales expertos en RSs/MAs en los ámbitos de la Educación, el Trabajo Social y la

Criminología (véase el sitio web en el apartado 'Enlaces'). Esta Colaboración tiene como propósito fomentar la realización de RSs y MAs de calidad en estos tres ámbitos científicos (Petrosino et al., 2001; Sánchez-Meca et al., 2002). El Grupo de Trabajo de Educación es especialmente interesante para obtener orientaciones, consejos, guías para la correcta realización de este tipo de investigaciones en Educación. Además, incorpora un repositorio con todas las RSs/MAs que se han realizado bajo los auspicios de esta Colaboración en el ámbito educativo.

4. DISCUSIÓN Y CONCLUSIONES

Las RSs y los MAs constituyen actualmente un método de investigación sólido y reconocido en las ciencias sociales y de la salud. Se pueden mencionar varias ventajas que los MAs aportan al proceso de acumulación del conocimiento científico en cualquier disciplina. En primer lugar, el MA aporta una mayor eficiencia, es decir, una mayor capacidad para tratar grandes cantidades de información. En

segundo lugar, las RSs y los MAs dotan de rigor científico al proceso de revisión de la literatura científica, facilitando su replicabilidad o reproducibilidad por otros investigadores. En tercer lugar, los MAs, al basarse en los tamaños del efecto, son más sensibles para detectar efectos que, aun siendo pequeños, pueden tener gran relevancia práctica. Esta capacidad para detectar efectos pequeños, pero relevantes, se debe a que un MA acumula los tamaños muestrales de los estudios individuales, logrando una mayor potencia estadística. Otra ventaja de los MAs es el énfasis en el tamaño del efecto como el mejor modo de cuantificar el resultado de una investigación, ya que estos índices ayudan a hacer una valoración de la significación práctica o real del resultado de una investigación. Otra gran ventaja de los MAs es su capacidad para explicar resultados heterogéneos o contradictorios entre los estudios mediante el análisis de variables moderadoras. Por último, y como resultado de todas estas ventajas, las RSs y los MAs

permiten alcanzar interpretaciones más fiables al basarse en una metodología cuantitativa, rigurosa y objetiva (Cooper, 2016; Cooper et al., 2019).

Las RSs y los MAs, como cualquier metodología de investigación, no están exentos de críticas ni de limitaciones. Una crítica que ha recibido el MA desde sus inicios es la conocida como 'el problema de las manzanas y las naranjas' ('the problem of mixing apples and oranges'), según el cual, estudios diferentes no deberían ser combinados. Sin embargo, como ha quedado demostrado a lo largo de la historia del MA, la inclusión de estudios parecidos pero diferentes permite investigar las razones de la heterogeneidad de sus resultados. Otra crítica que han recibido los MAs es la conocida como 'el problema de la basura dentro – basura fuera' ('garbage in – garbage out'), según la cual, si la calidad metodológica de los estudios empíricos es deficiente, entonces la validez de los resultados del meta-análisis se verá comprometida. Es por ello que

toda RS/MA debe incorporar alguna escala o checklist de valoración de la calidad metodológica de los estudios empíricos incluidos. De esta forma, es posible excluir estudios que presenten una calidad deficiente (ej., por debajo de un determinado umbral de la escala de calidad, o aquéllos que no cumplan con determinados ítems del checklist), o bien se puede analizar la posible relación entre calidad metodológica y los tamaños del efecto de los estudios, ya sea mediante análisis de subgrupos o por meta-regresión.

Otra crítica que se suele hacer a los MAs es la presencia de sesgo de publicación ('publication bias'). Dado que la mayoría (si no todos) de los estudios que se suelen incluir en un MA son estudios publicados, si existe sesgo de publicación en ese ámbito de investigación, entonces los resultados del MA estarán sesgados al alza, es decir, el tamaño del efecto medio sobreestimará el verdadero efecto poblacional. En primer lugar, es preciso señalar que el sesgo de publicación no es un problema generado por las RSs y los MAs. De hecho, aunque no existieran las RSs ni los MAs el sesgo de publicación

seguiría existiendo. En segundo lugar, dado que el sesgo de publicación puede ser una seria amenaza contra la validez de los resultados de un MA, se han desarrollado numerosos métodos estadísticos para detectar el problema y, en su caso, corregirlo mediante estimaciones ajustadas del verdadero efecto poblacional (Rothstein et al., 2005).

Los MAs también pueden sufrir sesgo de selección ('selection bias'), según el cual los criterios de inclusión de los estudios pueden verse influidos por el conocimiento previo que el meta-analista puede tener de los resultados del conjunto de estudios potenciales. Para contrarrestar este problema, es recomendable publicar el Protocolo de la RS/MA previo al inicio de la investigación y, con objeto de ofrecer la mayor transparencia, deben especificarse con total claridad los criterios de inclusión y exclusión de los estudios, de forma que cualquier lector pueda valorar la posible existencia de este sesgo.

Otro problema que pueden sufrir los MAs es el sesgo de reporte ('reporting bias') en los estudios empíricos. Este sesgo ocurre cuando los estudios primarios

tienden a reportar sistemáticamente sólo aquellas variables de respuesta que alcanzaron resultados favorables a la hipótesis de interés. Si no se tiene en cuenta esta circunstancia, entonces el MA puede alcanzar resultados sesgados. Es por ello que uno de los ítems de calidad metodológica que debe incluirse en los checklists es precisamente la existencia de sesgo de reporte en los estudios primarios, con objeto de valorar su potencial influjo sobre los resultados meta-analíticos.

En conclusión, las RSs y los MAs constituyen un tipo de revisión idóneo para desvelar el estado del arte sobre un determinado problema o fenómeno (educativo, psicoeducativo, psicológico,

práctico, etc.). No obstante, es preciso analizar críticamente las RSs/MAs, ya que pueden estar sujetos a diversos sesgos. Una correcta definición de los criterios de selección de los estudios, una

búsqueda comprehensiva de los estudios y una valoración de la calidad metodológica de los mismos, son aspectos centrales para obtener conclusiones razonadas de los MAs. Así desarrollados, los MAs, y las RSs en general, ofrecen evidencias y pruebas menos sesgadas que las revisiones narrativas tradicionales y contribuyen a una acumulación adecuada del conocimiento científico en cualquier ciencia empírica.

5. ENLACES

Sitio web de la organización Open Science Framework:
https://osf.io/

Sitio web del registrador de Protocolos PROSPERO:
https://www.crd.york.ac.uk/prospero/

Sitio web en el que se encuentra accesible el modelo de diagrama de flujo PRISMA 2020:
https://prisma-statement.org/prismastatement/flowdiagram.aspx

Sitio web del 'Equator Network' para localizar escalas y checklists de calidad metodológica según el tipo de estudios empíricos de un meta-análisis:

https://www.equator-network.org/

Sitio web de la escala PEDro para la valoración de la calidad metodológica de estudios primarios sobre la eficacia de programas, tratamientos e intervenciones:
https://pedro.org.au/english/resources/pedro-scale/

Sitio web de la escala metodológica Newcastle-Ottawa (NOS) de estudios no experimentales:
https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

Sitio web de la organización PRISMA:
https://www.prisma-statement.org/

Sitio web del checklist AMSTAR-2:
https://amstar.ca/

Sitio web del checklist REGEMA:
https://www.um.es/metaanalysis/REGEMA.php

Sitio web del checklist SWIM:
https://swim.sphsu.gla.ac.uk/

Sitio web de la Colaboración Campbell:
https://www.campbellcollaboration.org/

6. RECONOCIMIENTOS O FINANCIACIÓN

Estudio financiado por la Agencia Estatal de Investigación (Gobierno de España). PID2019-104080GB-I00/MCIN/AEI/10.13039/50110 00110 33. IP: Julio Sánchez Meca.

7. REFERENCIAS BIBLIOGRÁFICAS

Badenes-Ribera, L., Rubio-Aparicio, M. y Sánchez-Meca, J. (2020). Meta-análisis de generalización de la fiabilidad. *Informació Psicològica, 119*, 17-32. https://doi.org/dx.medra.org/10.14635/IPSIC.2020.119.6

Bahadivand, S., Doosti-Irani, A., Karami, M., Qorbani, M. y Mohammadi, Y. (2021). Prevalence of high-risk behaviors among Iranian adolescents: A comprehensive systematic review

and meta-analysis. *Journal of Education and Community Health, 8*(2), 135-142. https://doi.org/10.29252/jech.8.2.135

Becker, B.J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*, 257-278. https://doi.org/10.1111/j.2044-8317.1988.tb00901.x

Borenstein, M. y Hedges, L.V. (2019). Effect sizes for meta-analysis. En Cooper, H., Hedges, L.V. y Valentine, J.C. (Eds.), *The handbook of research synthesis and meta-analysis* (3ª ed.) (pp. 207-243). Russell Sage Foundation.

Borenstein, M., Hedges, L.V., Higgins, J.P.T. y Rothstein, H.R. (2019). *Introduction to meta-analysis* (2ª ed.). Wiley.

Botella, J. y Sánchez Meca, J. (2015). *Meta-análisis en ciencias sociales y de la salud*. Síntesis.

Breidbord, J. y Croudace, T.J. (2013). Reliability generalization for Childhood Autism Rating Scale. *Journal of Autism and*

Developmental Disorders, 43(12), 2855-2865. https://doi.org/10.1007/s10803-013-1832-9

Campbell, M., Katikireddi, S.V., Sowden, A. y Thomson, H. (2019). Lack of transparency in reporting narrative synthesis of quantitative data: A methodological assessment of systematic reviews. *Journal of Clinical Epidemiology, 105*, 1-9. https://doi.org/10.1016/j.jclinepi.2018.08.019

Campbell, M., McKenzie, J.E., Sowden, A., Katikireddi, S.V., Brennan, S.E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V. y Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *British Medical Journal, 368*(l6890). http://dx.doi.org/10.1136/bmj.l6890

Card, N.A. (2012). *Applied meta-analysis for social science research*. Guilford Press.

Cheng, L., Ritzhaupti, A.D. y Antonenko, P. (2019). Effects of the flipped classroom instructional strategy on students' learning outcomes: A meta-analysis. *Educational Technology Research and Development, 67*, 793-824. https://doi.org/10.1007/s11423-018-9633-7

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2ª ed.). Erlbaum.

Conn, V.S. y Rantz, M.J. (2003). Research methods: Managing primary study quality in meta-analyses. *Research in Nursing and Health, 26,* 322-333. https://psycnet.apa.org/doi/10.1002/nur.10092

Cooper, H. (2016). *Research synthesis: A step-by-step approach* (5ª Ed.). Sage.

Cooper, H., Hedges, L.V. y Valentine, J.F. (Eds)

(2019). *The handbook of research synthesis and meta-analysis* (3ª Ed.). Rusell Sage Foundation.

Cortina, J.M. y Nouri, H. (2000). *Effect size for ANOVA designs*. Sage.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies, 47*(2), 108- 12 1 . https://doi.org/10.1111/1467-8527.00106

Dekker, I. y Meeter, M. (2022). Evidence-based education: Objections and future directions. *Frontiers in Education, 7*:941410. https://doi.org/10.3389/feduc.2022.941410

Downes, M.J., Brennan, M.L., Williams, H.C. y Dean, R.S. (2016). Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). *British Medical Journal Open, 6*, e011458. https://doi.org/10.1136/bmjopen-2016-011458

Duval, S. y Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455-463. https://doi.org/10.1111/j.0006-341x.2000.00455.x

Egger, M., Higgins, J.P.T. y Smith, G.D. (2022). *Systematic reviews in health research: Meta-analysis in context* (2ª ed.). Wiley.

Egger, M., Smith, G.D., Schneider, M. y Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical Journal, 315*, 629-634.

Ellis, P.D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis and the interpretation of research results*. Cambridge University Press.

Erion, J. (2006). Parent tutoring: A meta-analysis. *Education and Treatment of Children, 29*, 79-106.

Giustini, D. (2019). Retrieving grey literature, information, and data in the digital age. En Cooper. H., Hedges, L.V. y Valentine, J.C. (Eds.), *The*

handbook of research synthesis and meta-analysis (3ª ed.) (pp. 101-126). Russell Sage Foundation.

Glanville, J. (2019). Searching bibliographic databases. En Cooper. H., Hedges, L.V. y Valentine, J.C. (Eds.), *The handbook of research synthesis and meta-analysis* (3ª ed.) (pp. 73-99). Russell Sage Foundation.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3-8. https://doi.org/10.3102/0013189X005010003

Glass, G.V. y Smith, M.L. (1978). *Meta-analysis of research on the relationship of class-size and achievement*. Far West Laboratory for Educational Research and Development, San Francisco (CA).

Glass, G.V., McGaw, B. y Smith, M.L. (1981).

Meta-analysis in social research. Sage.

Grissom, R.J. y Kim, J.J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2ª ed.). Routledge.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128. https://doi.org/10.3102/10769986006002107

Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J. y Welch, V.A. (Eds.) (2022). *Cochrane handbook for systematic reviews of interventions* (2ª ed.). Wiley. Disponible en: https://training.cochrane.org/handbook/current

Hutton, B., Salanti, G., Caldwell, D.M., Chaimani, A., Schmid, C.H. et al. (2015). The PRISMA Extension Statement for Reporting of Systematic Reviews Incorporating Network Meta-analyses of Health Care Interventions: Checklist and Explanations. *Annals of Internal Medicine, 162*, 777-784. https://doi.org/10.7326/M14-2385

IntHout, J., Ioannidis, J.P.A., Rovers, M.M. y

Goeman, J.J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *British Medical Journal Open, 6*, e010247. https://doi.org/10.1136/bmjopen-2015-010247

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. Russell Sage Foundation.

Kline, R.B. (2019). *Becoming a behavioral science researcher: A guide to producing research that matters* (2ª Ed.). Guilford Press.

Light, R.J. y Pillemer, D.B. (1984). *Summing up: The science of reviewing research*. Harvard University Press.

Lindberg, S.M., Hyde, J.S., Petersen, J.L. y Linn, M.C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136*(6), 1123-1135. https://doi.org/10.1037/a0021276

Lipsey, M.W. (2019). Identifying potentially interesting variables and analysis opportunities. En Cooper. H., Hedges, L.V. y Valentine, J.C. (Eds.), *The handbook of research synthesis and meta-analysis* (3ª ed.) (pp. 141-151). Russell Sage Foundation.

Lipsey, M.W. y Wilson, D.B. (2001). *Practical meta-analysis*. Sage.

López-López, J.A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W. y Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology, 67*, 30-48. https://doi.org/10.1111/bmsp.12002

McKenzie, J.E. y Brennan, S.E. (2022). Chapter 12: Synthesizing and presenting findings using other methods. En Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J. y Welch, V.A. (Eds.), *Cochrane handbook for systematic reviews of interventions* vers. 6.3. Cochrane. Disponible en: www.training.cochrane.org/handbook.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P. y Stewart, L.A. (2015). Preferred Reporting Items for

(PRISMA-P) 2015 statement. *Systematic Reviews, 4*(1):1. https://doi.org/ 10.1186/2046-4053-4-1

Morris, S.B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11,* 364-386. https://doi.org/10.1177/1094428106291059

Morris, S.B. y DeShon, R.P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-group designs. *Psychological Methods, 7*, 105-125. https://doi.org/10.1037/1082-989x.7.1.105

Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D. et al. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency,

openness, and reproducibility. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Page, M.J., Cumpston, M., Chandler, J. y Lasserson, T. (2021). Reporting the review. En Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J. y Welch, V.A. (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.2* (actualizada en Febrero 2021). Cochrane Collaboration. Disponible en www.training.cochrane.org/handbook

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews, 10*(89). https://doi.org/10.1186/s13643-021-01626-4

Patros, H.G., Alderson, R.M., Kasper, L.J., Tarle, S.J., Lea, S.E. y Hudec, K.L. (2016). Choice-impulsivity in children and adolescents with attention deficit/hyperactivity disorder (ADHD): A meta-analytic review.

Clinical Psychology Review, 43, 162-174. https://doi.org/10.1016/j.cpr.2015.11.001

Petrosino, A., Boruch, R.F., Soydan, H., Duggan, L. y Sánchez-Meca, J. (2001). Meeting the challenges of evidence-based policy: The Campbell Collaboration. *Annals of the American Academy of Political and Social Science, 578*, 14-34. https://doi.org/10.1177/0002716201578001002

Petticrew, M. y Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Blackwell.

Pigott, T. y Polanin, J.R. (2019). Methodological guidance paper: High-quality meta-analysis in a systematic review. Review of Educational Research, 90(1), 24-46. https://doi.org/10.3102/0034654319877153

Piñeiro-López, S., Martí-Vilar, M. y González-Sala, F. (2022). Intervenciones educativas en conducta prosocial y empatía en alumnado con altas capacidades: Una revisión sistemática. *Bordón, 74*(1), 141-157. https://doi.org/10.13042/Bordon.2022.90586

Rosa-Alcázar, A.I., Sánchez-Meca, J., Gómez-Conesa, A. y Marín-Martínez, F. (2008). The psychological treatment of obsessive-compulsive disorder: A meta-analysis. *Clinical Psychology Review, 28*, 1310-1325. https://doi.org/10.1016/j.cpr.2008.07.001

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (ed. rev.). Sage.

Rosenthal, R., Rosnow, R.L. y Rubin, D.B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

Rothstein, H.R., Sutton, A.J. y Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis:*

Prevention, assessment, and adjustments. Wiley.

Rubio-Aparicio, M., López-López, J. A., Viechtbauer, W., Marín-Martínez, F., Botella, J. y Sánchez-Meca, J. (2020). Testing categorical moderators in mixed-effects meta-analysis in presence of heteroscedasticity. *Journal of Experimental Education, 88*(2), 288-310. https://doi.org/10.1080/00220973.2018.1561404

Rubio-Aparicio, M., Sánchez-Meca, J., Marín-Martínez, F. y López-López, J.A. (2018). Guidelines for reporting systematic reviews and meta-analyses. *Annals of Psychology, 34*(2), 412-420 . http://dx.doi.org/10.6018/analesps.34.2.320131

Salameh, J.-P., Bossuyt, P.M., McGrath, T.A., Thombs, B.D., Hyde, C.J., Macaskill, P. et al. (2020). Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. *British Medical Journal, 370*(m2632).

Sánchez Martín, M., Navarro Mateu, F. y Sánchez-Meca, J. (2022). Las revisiones sistemáticas y la educación basada en evidencias. *Espiral. Cuadernos del Profesorado, 15*(30), 108-120.

Sánchez-Meca, J. (2008). Meta-análisis de la investigación. En Verdugo, M.A., Crespo, M., Badía, M. y Arias, B. (Coords.), *Metodología en la investigación sobre discapacidad: Introducción al uso de las ecuaciones estructurales* (pp. 121-139). Publicaciones del INICO (Colección ACTAS, 5/2008).

Sánchez-Meca, J. (2010). Cómo realizar una revisión sistemática y un meta-análisis. *Aula Abierta, 38*, 53-64.

Sánchez-Meca, J., Boruch, R.F., Petrosino, A. y Rosa-Alcázar, A.I. (2002). La Colaboración Campbell y la práctica basada en la evidencia. *Papeles del Psicólogo, 22*(83), 44-48.

Sánchez-Meca, J., López-López, J.A. y López-Pina, J.A. (2013). Some recommended statistical analytic practices when reliability generalization (RG) studies are conducted. *British*

Journal of Mathematical and Statistical Psychology, 66, 402-425. https://doi.org/10.1111/j.2044-8317.2012.02057.x

Sánchez-Meca, J. y López-Pina, J.A. (2008). El enfoque meta-analítico de generalización de la fiabilidad. *Acción Psicológica, 5*, 37-64.

Sánchez-Meca, J. y Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods, 13*, 31-48. https://doi.org/10.1037/1082-989X.13.1.31

Sánchez-Meca, J. y Marín-Martínez, F. (2010). Meta-analysis. En P. Peterson, E. Baker y B. McGaw (Eds.), *International Encyclopedia of Education* (3ª ed.), Vol. 7 (pp. 274-282). Elsevier.

Sánchez-Meca, J., Marín-Martínez, F., López-López, J. A., Núñez, Núñez, R. M., Rubio-Aparicio, M., López-García, J. J., López-Pina, J. A., Blázquez-Rincón, D. M., López-Ibáñez, C. y López-Nicolás, R. (2021). Improving the reporting quality of reliability generalization meta-

analyses: The REGEMA checklist. *Research Synthesis Methods, 12*(4), 516-536. https://doi.org/10.1002/jrsm.1487

Sánchez-Serrano, S., Pedraza-Navarro, I. y Donoso-González, M. (2022). ¿Cómo hacer una revisión sistemática siguiendo el protocolo PRISMA? Usos y estrategias fundamentales para su aplicación en el ámbito educativo a través de un caso práctico. *Bordón, Revista de Pedagogía, 74*(3), 51-66. https://doi.org/10.13042/Bordon.2022.95090

Sapp, M. (2017). *Primer on effect sizes, simple research designs, and confidence intervals*. Charles C. Thomas Pub., Ltd.

Saunders, L.D., Soomro, G.M., Buckingham, J., Jamtvedt, G. y Raina, P. (2003). Assessing the methodological quality of nonrandomized

intervention studies. *Western Journal of Nursing Research, 25*, 223-237. https://doi.org/10.1177/019394590225003 9

Scherer, R. y Shiddiq, F. (2019). The relation between students' socioeconomic status and ICT literacy: Findings from a meta-analysis. *Computers in Education, 138*, 13-32. https://doi.org/10.1016/j.compedu.2019.04.011

Schmidt, F.L. y Hunter, J.E. (2015). *Methods of meta-analysis: Correcting error and bias in research synthesis* (3ª Ed.). Sage.

Shea, B.J., Reeves, B.C., Wells, G., Thuku, M., Hamel, C. et al. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *British Medical Journal, 358*(j4008). http://dx.doi.org/10.1136/bmj.j4008

Stroup, D.F., Berlin, J.A., Morton, S.C., Olkin, I., Williamson, G.D., et al. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. Meta-analysis Of

Observational Studies in Epidemiology (MOOSE) group. Journal of the American Medical Association, 283, 2008-2012. https://doi.org/10.1001/jama.283.15.2008

Valentine, J.C., Aloe, A.M. y Wilson, S.J. (2019). Interpreting effect sizes. En Cooper, H., Hedges, L.V. y Valentine, J.C. (Eds.), *The handbook of research synthesis and meta-analysis* (3ª ed.) (pp. 433-452). Russell Sage Foundation.

Verhagen, A.P., de Vet, H.C., de Bie, R.A., Kessels, A.G., Boers, M., Bouter, L.M., Knipschild, P.G. et al. (1998). The Delphi list: A criteria list for quality assessment of randomised clinical trials for conducting systematic reviews developed by Delphi consensus. *Journal of Clinical Epidemiology, 51*(12), 1235-1241. https://doi.org/10.1016/s0895-4356(98)00131-0

Vevea, J.L., Coburn, K. y Sutton, A. (2019). Publication bias. En Cooper. H., Hedges, L.V. y Valentine, J.C. (Eds.), *The handbook of research synthesis and meta-analysis* (3ª ed.) (pp.

383-429). Russell Sage Foundation.
Vevea, J.L., Zelinsky, N.A.M. y Orwin, R.G. (2019). Evaluating coding decisions. En Cooper. H., Hedges, L.V. y Valentine, J.C. (Eds.), *The handbook of research synthesis and meta-analysis* (3ª ed.) (pp. 173-204). Russell Sage Foundation.

Wells, G.A., Shea, B., O'Connell, D., Peterson, J., Welch, V., Losos, M. y Tugwell, P. (2000). *The Newcastle–Ottawa Scale (NOS) for assessing the quality of non-randomized studies in meta-analysis*. Manuscrito no publicado, Universidad de Ottawa (Canadá).
White, I.R., Schmid, C.H. y Stijnen, T. (2021). Choice of effect measure and issues in extracting outcome data. En Schmid, C.H., Stijnen, T. y White, I.R. (Eds.), *Handbook of meta-analysis* (pp. 27-39). CRC Press.

Wilson, D.B. (2019). Systematic coding for research synthesis. En Cooper. H., Hedges, L.V. y Valentine, J.C. (Eds.), *The handbook of research synthesis and meta-analysis* (3ª ed.) (pp. 153-172). Russell Sage Foundation.

INFORMACIÓN SOBRE EL AUTOR

Julio Sánchez Meca
Universidad de Murcia

Catedrático de Metodología de las Ciencias del Comportamiento en la Facultad de Psicología. Responsable de la Unidad de Meta-análisis de dicha Facultad. Su línea de investigación prioritaria desde hace 40 años es la mejora de la metodología meta-analítica. Ha publicado más de 180 artículos sobre la metodología de los meta-análisis y estudios meta-analíticos aplicados en diversas Ciencias Sociales y de la Salud, en revistas tales como Psychological Methods, Research Synthesis Methods, Behavior Research Methods, Journal of Experimental Education, Journal of Educational and Behavioral Statistics. Cabe destacar su libro titulado 'Meta-análisis en Ciencias Sociales y de la Salud', publicado en 2015 junto con el profesor Juan Botella (Editorial Síntesis). Ha sido IP de ocho proyectos sobre meta-análisis financiados por el Plan Nacional I+D+i y

uno regional. Miembro fundador de la Society for Research Synthesis Methodology, la European Association of Methodology y la Asociación Española de Metodología de las Ciencias del Comportamiento.
