Clínica y Salud (2021) 32(3) 95-102

Clínica y Salud

https://journals.copmadrid.org/clysa

Evaluación de la Calidad de Estudios de Metaanálisis sobre la Eficacia de las Intervenciones en Revistas Españolas de Psicología

Hilda Gambara[a], Juan I. Durán[b] y Álvaro Santana[a]

[a]Universidad Autónoma de Madrid (UAM), España; [b]Universidad a Distancia de Madrid (UDIMA), España

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 6 de mayo de 2020
Aceptado el 23 de septiembre de 2020
Online el 12 de abril de 2021

Palabras clave:
Calidad de los estudios primarios
Metaanálisis
Revistas españolas de psicología

RESUMEN

En este trabajo se revisa la evaluación de la calidad de los estudios primarios incluidos en los metaanálisis publicados en las principales revistas españolas de psicología. Concretamente se analiza la codificación y evaluación de la calidad de los estudios en metaanálisis sobre eficacia de intervenciones, así como el propósito de esta evaluación y la relación entre la calidad y tamaños del efecto. Se encuentra que el 79% de los metaanálisis analizados incluyeron una evaluación de la calidad. Se discute la relación entre la menor calidad de los estudios en los metaanálisis y los resultados con mayores tamaños del efecto. Finalmente, se enfatiza la necesidad de mejorar el informe de los metaanálisis aportando evidencias de gran calidad.

Evaluation of the quality of studies in meta-analysis of intervention effectiveness in Spanish psychology journals

Keywords:
Primary study quality
Meta-analysis
Spanish psychology journals

ABSTRACT

In this paper, we reviewed primary study quality assessments in meta-analyses published in the main Spanish psychology journals. Specifically, we analyzed whether the coding and evaluation of the quality of the primary studies in meta-analysis based on the efficacy of interventions is a common practice. The purpose of this evaluation is also to report the relationship between quality and the reported results (effect sizes). It is found that 79% of the meta-analyses analyzed included a quality assessment. The relationship between the lowest quality of studies included in meta-analyses and larger effect sizes is also discussed. Finally, we stress the need to improve the reporting of meta-analyses including high-quality evidence.

Una de las primeras críticas que recibió el metaanálisis (MA) es la conocida como "*garbage in, garbage out*", que implica la imposibilidad de realizar buenas síntesis a partir de trabajos de escasa calidad (Botella y Sánchez-Meca, 2015). Desde entonces es patente la preocupación por realizar MA donde se ponga de manifiesto de forma transparente el tipo de estudios que se integran, preocupación compartida por una práctica basada en la evidencia. Sin embargo, en Psicología la evaluación de la calidad metodológica de la investigación solo se ha incorporado recientemente (Protogerou y Hagger, 2019).

No existe un criterio único para evaluar la calidad de los estudios primarios (CEP).

Entre ellos encontramos el estatus de la publicación, la calidad del informe, el tipo de diseño según una jerarquía, las características de los mismos o el riesgo de sesgo (RS) (Wells y Littell, 2009). Aunque en las últimas décadas se ha desarrollado un número creciente de indicadores de dicha calidad, las revisiones sobre el tema han puesto de manifiesto limitaciones tanto en su construcción como en su uso (Chacón-Moscoso et al., 2016; Deeks et al., 2003; Wells y Little, 2002), constatándose resultados divergentes dependiendo de las herramientas utilizadas. En consecuencia, las conclusiones de algunos MA pueden depender del procedimiento de evaluación seguido (Herbison et al., 2006; Juni et al., 1999; Losilla et al., 2018).

Lo que estos problemas están reflejando es que no existe una concepción única de la CEP (Johnson et al., 2015). Nosotros nos adscribimos a la definición de Valentine (2009), para quien la calidad es: "*the fit between a study's goals and the study's design and implementation characteristics*" (p. 130), definición que enlaza con la propuesta sobre la validez como aproximación a la veracidad de las inferencias sobre la relación entre variables, no como algo dicotómico o unidimensional, de todo o nada, sino como algo gradual (e.g., Shadish et al., 2002). Características diversas del diseño de un estudio y su implementación conducirán a inferencias que serán más o menos válidas

Para citar este artículo: Gambara, H., Durán, J. I. y Santana, A. (2021). Evaluación de la calidad de estudios de metaanálisis sobre la eficacia de las intervenciones en revistas españolas de psicología. *Clínica y Salud, 32*(3), 95-102. https://doi.org/10.5093/clysa2021a4

Correspondencia: hilda.gambara@uam.es (H. Gambara).

en una o más dimensiones. En el contexto de este artículo, la validez más vinculada a la calidad es la validez interna. Ahora bien, esta es una propiedad más de las inferencias que del diseño en sí. Dado que no hay un único método que asegure la validez interna de las inferencias causales, en cada circunstancia habrá que valorar en qué medida cada método implica mayores o menores garantías en este sentido.

Desde la medicina basada en la evidencia (MBE) se considera que la mejor evidencia es aquella derivada de los diseños experimentales de grupos aleatorios (ensayos controlados aleatorizados). En esta línea, la colaboración Cochrane y Campbell aconseja realizar MA con la mejor evidencia (utilizando el diseño como criterio de inclusión en la revisión) y analizar la CEP a partir del riesgo de sesgo (RS). Sin embargo, en el ámbito de la Psicología no es tan claro que en un mismo MA no deban incluirse estudios no estrictamente experimentales por considerarlos de menor calidad. Como plantea Valentine (2009), al menos no hay evidencia de que esto tenga que ser así; habrá que considerar el contexto de investigación, el tipo de diseño que se haya utilizado para alcanzar los objetivos y lo cuidadoso que se haya sido en el control de las amenazas a la validez. Por otra parte, en Psicología es usual que los MA incluyan estudios con metodologías no tan homogéneas como en Medicina (e.g., únicamente de grupos aleatorios), siendo más frecuente el uso de diseños cuasiexperimentales para evaluar intervenciones.

A pesar del auge de esta línea de investigación, existen pocos estudios que hayan analizado la relación de la CEP con los MA en Psicología. Hohn et al (2018) han analizado dicha relación recientemente con una muestra de 386 MA, encontrando que en menos de un tercio se informaba explícitamente sobre la evaluación de la calidad, de los cuales aproximadamente la mitad (54.7%) describieron el procedimiento seguido explícitamente. Se constató también una gran heterogeneidad entre los instrumentos de evaluación. Estos resultados sugieren que informar sobre los estándares de calidad no es una práctica que haya penetrado en Psicología.

El objetivo general de esta revisión es evaluar el papel de la CEP en los MA publicados en las principales revistas españolas. Una foto que evalúe la calidad de las fuentes puede contribuir a mejorar la credibilidad sobre las evidencias detectando los posibles problemas a resolver. Nos centramos en los MA sobre eficacia de intervenciones. Hohn et al. (2018) encontraron que es en estos donde más se reporta la calidad (un 55.2% en MA de eficacia de intervenciones frente a un 20.1% en MA de no intervenciones).

Específicamente, en este trabajo se analiza si la evaluación de la CEP en MA sobre eficacia de intervenciones publicados en revistas españolas es una práctica común, la relación entre la calidad metodológica de los estudios y los TE estimados y, finalmente, los fines de dicha evaluación.

Método

Selección y Muestra de Estudios

En primer lugar, se seleccionaron las principales revistas españolas a través del índice de impacto (*Journal of Citations Reports*, JCR, 2017) y el índice H de Google. A saber: *Psicothema, International Journal of Clinical and Health Psychology, Anales de Psicología, Psicología Conductual, Revista de Psicología Social, Infancia y Aprendizaje, Spanish Journal of Psychology, Psicológica, Estudios de Psicología, Clínica y Salud y Psychosocial Intervention.*

Tabla 1. Metaanálisis publicados en revistas españolas

Revista	Número de MA	Número de MA sobre intervención
International Journal of Clinical and Health Psychology	10	3
Psychosocial Intervention	2	1
Psicothema	18	13
Psicología Conductual	13	10
Revista de Psicología Social Aplicada	1	0
Anales de Psicología	9	6
Spanish Journal of Psychology	5	1
Psicológica	0	0
Estudios de Psicología	2	0
Clínica y Salud	2	0
Total	62	34

En segundo lugar, se realizó una búsqueda (abril, 2019; sin límite temporal) de los MA utilizando las bases de datos PsycINFO y SCOPUS (a través del portal EBSCOhost), con las palabras clave en título y/o resumen: *Meta-analysis* o metaanálisis, *Systematic Review* o Revisión Sistemática, filtrando por el nombre de las revistas. Se obtuvieron 62 MA candidatos (Tabla 1). A través del título, resumen, y/o texto completo se depuraron las duplicaciones, los artículos teóricos, revisiones narrativas, resúmenes y los MA que no eran de eficacia de programas de intervención o tratamientos[1].

Tabla 2. Porcentaje de acuerdo interjueces e índice kappa de Cohen para las variables y categorías codificadas

Variable	Categorías	% Acuerdo	Kappa
¿Ha seguido el MA alguna guía para escribir el informe?		100	1
Número de estudios incluidos		100	1
Tipos de diseño incluidos		94[1]	-
	Experimentales	100	1
	Cuasiexperimentales con grupo control	100	1
	Cuasiexperimentales sin grupo control	94	.88
¿Se han incluido variables metodológicas como moderadoras?		100	1
¿Se ha analizado la CEP?		97	.95
¿Para qué se ha analizado la CEP?		76[1]	-
	Con criterios de inclusión	100	1
	Con fines descriptivos	94	.87
	Como variable moderadora	76	.53
¿Qué criterios se utilizaron para evaluar la calidad?		71[1]	-
	Características del diseño	94	.77
	Características de las medidas	97	.92
	Según el tamaño de la muestra	79	.59
Los autores discuten explícitamente la importancia de la evaluación de la calidad sobre los resultados del MA		82	.63

Nota. [1]Variables en las que era posible seleccionar más de una categoría. Los casos en los que hubo un acuerdo parcial (i.e., había acuerdo sobre algunas categorías, pero no en todas) se consideraron como desacuerdos en el cálculo del porcentaje de acuerdo global.

Tabla 3. Porcentaje de metaanálisis en los que se incluye cada tipo de diseño (se incluyan o no otros también)

Tipo de diseños incluidos	Frecuencia (% casos)	Media (mediana) de número de estudios incluidos
Solo experimental	5 (15%)	37.8 (11)
Experimentos y cuasiexperimentos con grupo de control	14 (42%)	21.6 (19)
Experimentos y cuasiexperimentos con y sin grupo de control	11 (33%)	22.8 (23)
Solo cuasiexperimentos	3 (9%)	20.3 (19)
$n = 1$	0 (0%)	-

Se incluyeron todos los MA de eficacia de intervenciones independientemente del tipo de diseño utilizado en los estudios primarios. Se realizó, por último, una revisión manual de los números de las revistas que por antigüedad no constaban en SCOPUS host. La muestra final la constituyeron 34 MA.

Codificación de Estudios

Se elaboró un libro de codificación con categorías relativas a la CEP (procedimiento de evaluación, finalidad y repercusiones sobre los resultados). Se codificó revista, año de publicación, área de intervención, uso de guías estandarizada (MARS, PRISMA, etc.), tipo de diseño, variables metodológicas consideradas como moderadoras, explicitación del análisis de la CEP y finalidad del análisis de la calidad sobre el MA. Se codificaron también las relaciones explicitadas entre distintos aspectos de la calidad y los TE.

Todos los estudios fueron codificados independientemente por el segundo y tercer autor. Los desacuerdos se discutieron y aclararon entre los tres autores. El acuerdo interjueces varió entre el 71% y el 100% (detalles de los acuerdos en Tabla 2).

Análisis de Datos

Tras la codificación se obtuvieron los porcentajes de acuerdo y el índice kappa de Cohen para cada una de las variables codificadas. En el caso de las variables en las que era posible seleccionar más de una categoría, los casos en los que hubo un acuerdo parcial (i.e., había acuerdo sobre algunas categorías, pero no en todas) se consideraron como desacuerdos. Una vez aclarados los desacuerdos, se obtuvieron estadísticos descriptivos a partir de la base de datos definitiva con el software de análisis estadístico SPSS v23.0. En concreto, los estadísticos fueron frecuencias absolutas y relativas correspondientes a las distintas categorías de cada variable y tablas de contingencia con frecuencias conjuntas absolutas y relativas para las distintas combinaciones de pares de variables.

Resultados

Es de destacar que de los 34 MA un 59% pertenecen a un mismo grupo de investigación, por lo que estas revisiones guardan características comunes.

El período en el que más MA se publicaron es el comprendido entre 1998 y 2002 (Figura 1). Esto se debe principalmente a que la revista *Psicología Conductual* publicó un monográfico en 2002 dirigido por el grupo de investigación mencionado que representa un 18% de la muestra.

Sobre el ámbito de estudio de los MA revisados, la mayoría van dirigidos a valorar la eficacia de programas de intervención o tratamientos de problemas de salud o psicológicos: ansiedad, fobia (7), enuresis, depresión, depresión infantil, TOC, insomnio, consumo de drogas, adicción al tabaco (2), maltrato (2), maltratadores. Una minoría se centra en programas de prevención, mientras que otra minoría se centra en la eficacia de intervenciones en lectura, escritura y comprensión.

En el 91% de las revisiones no se explicitó si se seguía alguna guía estandarizada. Tres MA, todos publicados después del 2016, utilizaron la guía de informe estandarizada PRISMA. Ninguno hizo mención a las guías de la APA ni a ninguna otra guía.

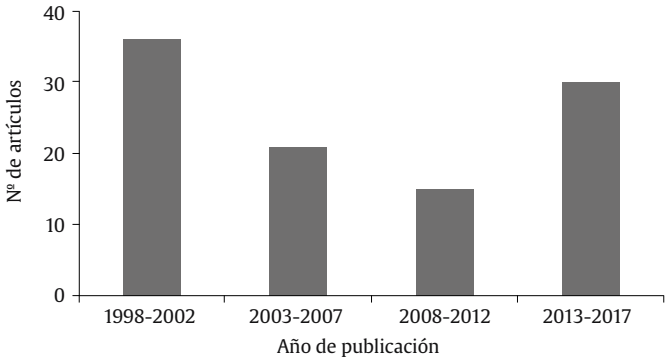

Figura 1. Número de metaanálisis publicados por períodos de 5 años desde 1998.

La mediana de estudios incluidos en los MA es de 19. En cuanto a los diseños de los estudios, solo cinco MA (15%) incluyeron exclusivamente diseños experimentales. Lo más frecuente (42%) fue combinar diseños de grupos aleatorios con cuasiexperimentales con grupo de control, aunque también encontramos casos en los que se incluyeron, además de los anteriores, diseños sin grupo de control (33%). No se incluye ningún MA de $n = 1$ (Tabla 3).[2] Respecto a estos resultados, es importante tener en cuenta que en ocasiones no queda claro cuál es el tipo de diseño incluido al que se refieren los autores del MA. Por ejemplo, en ocasiones se hace referencia a estudios pre-experimentales, que fueron valorados en este trabajo como equivalentes a un diseño pre-post con un solo grupo a falta de una referencia más clara.

El 71% de los MA incluyen variables metodológicas como moderadoras.

Tabla 4a. Tabla de contingencias con frecuencias y porcentajes de fila con los metaanálisis que realizaron análisis de la CEP y/o incluyeron variables metodológicas como moderadoras

	Inclusión de variables metodológicas como moderadoras	
Análisis de la CEP	Sí	No
Sí	21 (78%[1])	6 (22%[1])
No	3 (43%[1])	4 (57%[1])

[1]Porcentajes respecto al total por filas.

Prácticamente todos los MA vinculados al mismo grupo de investigación siguen a Lipsey y Wilson (2001), quienes aconsejan recoger información relativa a variables extrínsecas, externas y metodológicas. Dentro de las variables metodológicas se codifican variables relativas a aspectos relacionados con el diseño, como puede ser la aplicación de asignación aleatoria, la realización de medidas de seguimiento o el tipo de grupo de control empleado, siendo también común valorar aspectos como el porcentaje de mortandad experimental o el tamaño muestral. De entre los estudios que valoran la calidad de los estudios primarios, el 78% incluye también variables metodológicas como moderadoras (Tabla 4a).[3]

Tabla 4b. Frecuencia y porcentajes de fila de metaanálisis que incluyeron variables metodológicas como moderadoras que realizaron análisis de la CEP y propósito de la codificación de la CEP por tipo de diseños incluidos

Diseño incluido	Propósito de la codificación de la CEP			
	Análisis de la CEP	Inclusión de v. moderadoras metodológicas	Criterios de inclusión	Fines descriptivos
Solo experimental	4 (80%)	2 (40%)	4 (80%)	2 (40%)
Experimentos y cuasiexperimentos (con grupo control)	12 (86%)	11 (79%)	8 (67%)	9 (75%)
Experimentos y cuasiexperimentos (con y sin grupo control)	9 (82%)	8 (73%)	2 (22%)	7 (78%)
Solo cuasiexperimentos	2 (67%)	2 (67%)	0 (0%)	2 (100%)

Centrándonos en la CEP, del total de MA analizados, 27 MA (79%) informaron llevar a cabo una evaluación de la calidad; de estos MA, 18 (67%) indicaron explícitamente como se llevó a cabo. Los criterios utilizados para evaluar la calidad fueron las características del diseño (88% de los casos), el tamaño muestral (56%) y las características de las medidas (26%). En un 9% de las ocasiones no se indicó. Por otra parte, tal y como aparece en la Tabla 4b, es más habitual realizar una evaluación de la CEP cuanto más restrictivos son los criterios de inclusión empleados sobre los diseños de esos estudios. Así, el 80% de los MA que incluyeron solo experimentos realizaron una evaluación sobre la calidad metodológica de los estudios. Este porcentaje es similar para los MA que combinaron estudios experimentales y cuasiexperimentales (86% y 82%, incluyendo estudios con y sin grupo de control, respectivamente). En los casos en los que solo se incluyen cuasiexperimentos, este porcentaje baja hasta un 67%.

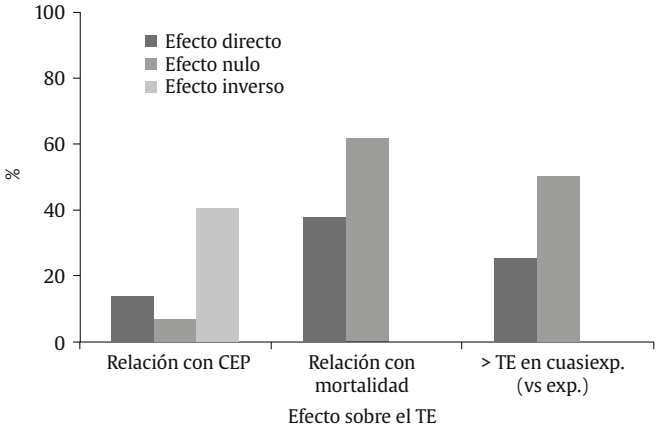

Figura 2. Porcentaje de casos en los que aparecieron los efectos más frecuentemente reportados entre las moderadoras metodológicas, de entre los estudios que las analizar.

La mayoría de los MA que comparten equipo utilizaron una escala *ad hoc* de 9 ítems. Estos ítems no se especifican siempre, de manera que no podemos asegurar que se trate de la misma escala o se haya adaptado al contexto del MA[4]. Ocho MA utilizan siete herramientas estandarizadas para evaluar la CEP (la escala Jadad, Maryland Scale of Scientific Rigor (2), escala de Estimación de la Calidad, herramienta del NHS Center for Reviews and Dissemination, Collaborative Outcome Data, Quality Assessment Tool for Quantitative Studies y Scottish Intercollegiate Guidelines Network).

Sobre la relación entre CEP y resultados del MA, como se muestra en la Tabla 5 y Figura 2, de los 15 MA que analizaron la calidad como variable moderadora, en un 61% de ellos se encontró una relación inversa entre calidad y tamaño del efecto (TE) a nivel descriptivo, que en el 40% de los casos fue estadísticamente significativa. Sólo un MA (6.7%) reporta una relación directa entre la calidad y el TE; en el resto de casos no se encontró o no se reporta ningún tipo de relación.

Atendiendo a otras variables metodológicas, se encontró una relación directa entre mortandad experimental y TE a nivel descriptivo en un 53% de los casos en los que se estudió y estadísticamente significativa en un 38%. Ningún MA reportó una relación inversa entre estas dos variables. Por último, también es frecuente encontrar un mayor TE en diseños cuasiexperimentales que en experimentos cuando se utilizó el diseño como variable moderadora (57% a nivel descriptivo y 25% con diferencia estadísticamente significativa).

Sin embargo, un MA indica mayores TE en diseños experimentales que en cuasiexperimentales a nivel descriptivo y otro MA reporta este mismo efecto estadísticamente significativo.

Discusión

Contrasta el muy escaso uso de guías de informe estandarizadas en comparación con lo reportado en la literatura dentro del ámbito médico; este dato es consistente con el encontrado por Hohn et al. (2018) en el ámbito de la Psicología. Trabajos anteriores han examinado esta cuestión, concluyendo que en la mayoría de los MA los informes carecen de la información suficiente para ser replicados (e.g., Dieckmann et al., 2009; Harwell y Maeda, 2008). Seguir estas guías facilita el reporte de síntesis más transparentes, consistentes y replicables (Atkinson et al., 2015). En este trabajo hemos constatado una falta de adherencia a dichas guías; por ejemplo, en algunos casos no se incluyó el término metaanálisis en el propio título, aunque puede ser justificable por la antigüedad de los trabajos, algo que se debería tener en cuenta a la hora de realizar revisiones.

La falta de utilización de estas guías puede deberse a un desconocimiento de las mismas, a que la publicación del MA sea previa a la creación de dichas guías y/o al seguimiento de otros manuales o recomendaciones publicadas por expertos en este tema en España. Si esto último es cierto, sugerimos incluir en la publicación las recomendaciones seguidas; de esta manera, el lector del MA (y los revisores) podrá evaluar mejor el MA. En algunas revistas internacionales se está recomendando cumplimentar una lista de chequeo estandarizada que se envía junto al informe MA

Tabla 5. Resultados del estudio de la repercusión de variables relacionadas con la CEP sobre la magnitud del TE

Moderadora	Número de MA en los que se analiza	Relación a nivel descriptivo (% casos)	Relación estadísticamente significativa (% casos)
CEP	15	Inversa (61)	Inversa (40)
Mortandad experimental	13	Directa (53)	Directa (38)
Cuasiexperimentos vs. experimentos	12	Mayor en cuasiexperimentos (57)	Mayor en cuasiexperimentos (25)

para su publicación. Recientemente, Rubio-Aparicio et al. (2018) han publicado recomendaciones para el reporte del MA. La APA en su séptima edición abunda en este aspecto. También en el siguiente enlace del grupo Equator puede encontrarse una amplia guía de informes (más del ámbito médico y de la salud) para diferentes estudios primarios y MA (https://www.equator-network.org/reporting-guidelines/) que no solo pueden mejorar el informe, sino que pueden ayudar a la realización de la propia investigación.

En relación al ámbito de intervención, los MA analizados se relacionan mayoritariamente con la eficacia de tratamientos clínicos y solo una minoría evalúa la eficacia de otro tipo de programas (e.g., de prevención).

Sobre el tipo de diseños incluidos en los MA, aunque tanto la colaboración Cochrane como la Campbell aconsejan realizar síntesis homogeneizando el tipo de diseño, hay autores que abren la posibilidad a incluir diseños heterogéneos (e.g., experimentales y cuasiexperimentales) pero analizando el RS (la validez interna) y agregando solo aquellos con mayor calidad (Sterne et. al, 2016; Wilson y Lipsey, 2006). Esta situación está más cercana a las investigaciones sobre eficacia de programas o tratamientos psicológicos, donde se constata que habitualmente se aplican diseños experimentales y cuasiexperimentales. Restringir el tipo de diseño en el contexto de la evaluación de la eficacia de programas o de tratamientos psicológicos limitaría en gran medida las evidencias que nos pueden aportar diferentes estudios realizados con rigor metodológico (Codray y Murphy, 2009).

El resultado encontrado sobre las principales características metodológicas de los estudios primarios es consistente con Hohn et al. (2018). Estas son el diseño y la pérdida de participantes. Por otro lado, más de la mitad (67%) de los MA que valoran la calidad explicitan cómo se analiza, porcentaje superior al 55% reportado por Hohn et al. (2018). Habría que analizar qué ocurre con respecto a los MA de no intervención en un futuro trabajo.

Centrándonos en los problemas de las escalas de calidad, en los artículos revisados se reflejan los problemas documentados en la literatura. Por ejemplo, en algunas de las escalas utilizadas se puntúa si hay asignación aleatoria para, a continuación, codificar si el diseño es experimental o cuasiexperimental. Está claro que existe un solapamiento de información, puesto que si el diseño es experimental necesariamente habrá asignación aleatoria o se evalúa la pérdida de participantes, pero no se diferencia de la pérdida no aleatoria de participantes, amenaza a la validez interna que puede ser mayor.

Así, reflejar en un solo número aspectos que no tienen por qué estar relacionados, como medidas de validez de los resultados y procedimiento de asignación de los participantes, es una fuente de inconsistencia (Valentine y Cooper, 2008). Nos podríamos encontrar dos estudios con la misma puntación en calidad, uno con una validez interna débil y otro con una fuerte validez de medida.

En definitiva, revisadas las escalas utilizadas en los MA de este trabajo, en línea con la colaboración Cochrane y Cambpell, no parece recomendable su utilización, al menos sin explicitar los ítems que se incluyen en dicha escala.

Hohn et al. (2018) plantearon las siguientes explicaciones sobre la escasa evaluación de la CEP en Psicología: a) no es un hábito en este ámbito, b) sí se considera para la inclusión de los estudios, pero no se explicitan las herramientas utilizadas, c) no se sabe qué procedimiento o herramienta escoger de entre todas las existentes (estos autores reportaron 34 herramientas en los 69 MA que aplicaron alguna). En nuestro caso, esta variabilidad ha sido bastante menor.

Sobre la finalidad de esta evaluación, en ningún caso se utilizó como criterio de ponderación o para un análisis de sensibilidad (Lipsey y Wilson, 2001). En la mayoría de los MA se analizó empíricamente la calidad del estudio como potencial variable moderadora de las estimaciones de los TE.

El término riesgo de sesgo solo aparece en un MA. Hay que tener en cuenta que este término apareció en 2006 y no se incluyó en el manual de la APA del 2010; sí es recogido por Appelbaum et al. (2018) e incorporado en la séptima edición APA (2020), por lo que seguramente este término se irá extendiendo. Por último, planteamos la siguiente cuestión que habría que constatar empíricamente y que sería coherente con algunos resultados que han suscitado la crisis de replicabilidad.

Puesto que con el trascurso del tiempo se ha ido incidiendo cada vez más en la importancia de incluir evidencias de calidad en los MA, cabe pensar que los MA más cercanos en el tiempo incorporan estudios cualitativamente mejores, lo que podría repercutir sobre las conclusiones de las revisiones; podría ocurrir que los TE promedio fuesen menores ahora que en MA anteriores, que podrían estar sobreestimando TE.

En cualquier caso, siguiendo a Botella y Durán (2019), creemos que el progreso de la ciencia implica cambiar desde una perspectiva competitiva a otra cooperativa. La idea de que cada estudio de investigación contribuye a la ciencia, no por su aportación singular o revolucionaria, algo que ocurre excepcionalmente, sino porque junto con otros colabora para alcanzar resultados consistentes, nos impulsa no solo a desarrollar investigaciones de calidad sino a informar de las investigaciones pensando en su aportación. El cuidado que debe ponerse en los informes no se convierte en una cuestión formal o estética, sino que puede ser imprescindible para que un trabajo concreto pueda ser incorporado o no en una futura revisión sistemática. En consecuencia, la CEP y la calidad del informe deberían ir de la mano.

Extended summary

The interest concerning the assessment of primary study quality (PSQ) comes from areas such as meta-analysis (MA) and Evidence-Based Medicine (EBM). Since reports need to include high quality studies in order to make good syntheses, the concern about doing MAs which explicitly report the type of studies included in them is increasing.

Nevertheless, there is no unique criterion for the mentioned quality assessment yet. Some of these criteria are, for example, publication status, quality of the report or type of design, amongst others. Research has shown limitations in the construction of scales for evaluating PSQ. As a consequence, results might differ depending on the scale used, as well as the conclusions drawn from MAs using them.

We follow Valentine (2009), who defines quality as "the fit between a study's goals and the study's design and implementation characteristics" (p. 130). Design characteristics lead to inferences with different validity in one or more dimensions. But there is no single method to ensure internal validity (i.e., the kind of validity which is more closely related to quality) and we should evaluate the extent to which we can guarantee it in every circumstance.

From the EBM perspective, randomized controlled trials (RCTs) provide the best evidence, and both the Cochrane and Campbell collaborations recommend to make syntheses using this type of design. However, in Psychology and other related areas it would not be adequate to include solely RCTs in the reviews since the use of other designs (e.g., quasi-experiments) is widespread and plenty of evidence would be disregarded.

Recently, Hohn et al. (2018) evaluated the report of the primary study quality (PSQ) in a sample of 386 MAs in Psychology. They found that less than a third of the syntheses reported PSQ, and only half of the studies which reported it explicitly described the procedure. Therefore, that study shows the low use of quality reports in MAs performed in Psychology.

As a result, the general objective of this work is the evaluation of PSQ reports published in the main Spanish journals in a sample of MAs whose aim is to summarize evidence about the efficacy of

interventions or treatments. Thus, we pretend to detect whether PSQ assessments in Spanish MAs is a common practice, if it is explicitly reported or not, what its purposes are, and whether there is a relationship between the PSQ and the effect sizes (ES). We evaluated only intervention studies because Hohn et al. (2018) found that the highest percentage of PSQ assessments were performed in that area.

Method

Eleven journals were selected according to their Journal of Citation Reports (JCR, 2017) impact factor, namely: *Psicothema, International Journal of Clinical and Health Psychology, Anales de Psicología, Psicología Conductual, Revista de Psicología Social, Infancia y Aprendizaje, Spanish Journal of Psychology, Psicológica*, *Estudios de Psicología, Clínica y Salud,* and *Psychosocial Intervention*.

A search of the MA was carried out (April, 2019) in PsycINFO and Scopus, using meta-análisis [and metaanálisis], meta-analysis, *revisión sistemática,* and systematic review as keywords. Sixty two MAs were obtained at first (Table 1). Through title, abstract and/or full text, we dropped out duplications, theoretical articles, and other MAs which did not evaluate the efficacy of interventions or treatments. The final sample was integrated by 34 MAs, which included primary studies with different types of designs.

A coding manual was elaborated to register: journal, year of publication, area of intervention, use of standardized guidelines (MARS, PRISMA, etc.), type of design of the primary studies included, methodological variables considered as moderators variables, explanation of the PSQ analysis and purpose of the quality analysis in MAs. Finally, the relationships found between quality and ES were coded. The second and third author of this work independently coded all these aspects, with an interrater agreement ranging from 71% to 100%.

Results

Most MAs in the sample evaluated health or psychological intervention programs concerning issues such as anxiety, depression, enuresis, OCD, and others. To a lesser extent, they summarized the effect of prevention programs or efficacy of interventions in reading, writing, and comprehension.

Ninety one percent of MAs did not explicitly follow any reporting guide. Only three of them used PRISMA, but none of them referenced APA or other guidelines. Regarding the types of design included, 42% of the reviews combined randomized controlled trials and quasi-experiments, 33% included designs with no control group and 15% included only experimental designs.

Seventy one percent of MAs included methodological variables, such as type of design, randomization, follow-up measures, type of control group used, experimental mortality, or sample size as moderator variables. Seventy eight percent of MAs that assessed PSQ included these methodological variables as moderators.

Seventy nine percent of the reviews reported having carried out a PSQ assessment. Eighteen (67%) of these indicated explicitly how it was performed. The most employed criteria to assess quality were: design characteristics (88%), sample size (56%), and characteristics of measures (26%). Table 4b shows that MAs that employ more restrictive inclusion criteria carried out a PSQ evaluation in a greater proportion. Eighty percent of MAs which incorporated solely experimental designs carried out this evaluation, while 67% of the syntheses which included entirely quasi-experimental designs did.

Eight MAs use up to 7 scales to assess PSQ, including the Jadad Scale, Maryland Scale of Scientific Rigor, and others. Table 5 illustrates that 61% of MAs which incorporated PSQ as a moderating variable found an inverse relationship between PSQ and ES at a descriptive level, and 40% found a statistically significant relationship. Most of the remaining reviews did not show or report any relationship between these two variables, and only one of them mentioned a direct relationship regarding PSQ and ES.

Furthermore, a direct relationship between experimental mortality (or attrition) and ES was found in 53% of MAs in a descriptive way, and in 38% of the entire sample of MAs the relationship was found statistically significant. Another variable that influences effect sizes is the type of design, since 57% of reviews registered a greater effect in quasi- experiments rather than experiments at a descriptive level, and 25% found this difference to be statistically significant.

Discussion

In accordance with the findings of Hohn et al. (2018), there is a low use of standardized reporting guidelines in the field of Psychology with respect to other areas. Previous studies have pointed out the lack of information that many reports have, a fact that makes it difficult to replicate these syntheses (Dieckmann et al., 2009; Harwell & Maeda, 2008). This replication is necessary for more transparent, consistent, and replicable progress in science.

The scarce use of these guidelines may be due to a lack of knowledge, an adherence to other references or to the guides being subsequently created with respect to the oldest investigations. Following Rubio-Aparicio et al. (2018), we recommend citing the guidelines followed in future MAs, since they allow a faster and more effective evaluation of the reports. The following link contains different reporting guides for meta-analysts: https://www.equator-network.org/reporting-guidelines/.

As some authors pointed out, the inclusion of heterogeneous designs in the same MA might be desirable in Psychology, especially in areas related to psychosocial, educational or psychological treatment programs, where experimental designs are not frequently performed. By including only a certain type of design, the amount of evidence provided by other methodologically rigorous studies would be restricted (Codray & Murphy, 2009).

According to the findings reported in Hohn et al. (2018), the type design and the loss of participants are the methodological characteristics most taken into account. Furthermore, in our sample, 67% of MAs which analyzed the quality explained how it was done. This percentage is higher than the 55% reported by Hohn et. al. (2018).

The problems related to quality scales are not only referred to their scarce use, but also to their construction. For example, there is an overlap of information when including items about random assignment and also about experimental or quasi-experimental design, since experimental designs necessarily have a random assignment of the participants to the different conditions of the study. Therefore, there may be two studies with differing internal validity and measurement validity with the same score on a quality scale. Some of these evaluations cannot distinguish between issues which have little in common, like validity measures of the results and the method used to assign the participants to the conditions. Therefore, since a unidimensional approach of quality seems simplistic, it does not seem appropriate to use these scales without clearly explaining the items that constitute them.

According to Hohn et. al. (2018), the scarce evaluation of PSQ in Psychology can be explained by a) a lack of habit, b) quality is considered as an inclusion criterion but the assessment tools used are not reported, c) lack of agreement amongst all the options available. In our review, we did not find as much variability in the use of different tools as these authors reported, probably because our sample was smaller.

The majority of MAs used PSQ as a moderating variable in the estimation of ES, but none of them used it as a weighting criterion or for performing a sensitivity analysis.

It is conceivable that primary studies with higher quality are increasingly incorporated in the closest MAs in time. This may impact the conclusions of the reviews. To this end, a research question, which is beyond the aim of this study, is whether ES could be lower now compared to older meta-analytical studies, according to the inverse nature of the relationship between methodological quality and ES found in some reviews.

Finally, we follow Botella and Durán (2019) in emphasizing the cooperative perspective necessary for the progress of science. Furthermore, we recommend reporting research in a consistent and systematic way, following standardized guidelines for performing MA, as this may impact the incorporation of such investigations in a future review and it can make them more accessible and replicable for other colleagues. In other words, measuring PSQ could be useless if it is not followed by an appropriate report, and vice versa.

Conflicto de Intereses

Los autores de este artículo declaran que no tienen ningún conflicto de intereses.

Notas

[1]No se puso ningún tipo de restricción en cuanto a la temática del tipo de intervención o programa prevención (salud, educativo o social).

[2]La suma de 33 estudios aquí en lugar de los 34 seleccionados se debe a que en uno de ellos (Méndez et al., 2002) menciona que los estudios incluidos fueron "diseños de grupo que comparen un grupo de tratamiento frente a uno de control o tratamiento alternativo" y aunque probablemente se trate de cuasiexperimentos, al no saberse con seguridad no fue incluido en ninguna de las categorías que se muestran en la tabla, por lo que no aparece representado en ella.

[3]Los tres estudios que no analizan la CEP fueron los siguientes: Beelman y Lösel (2006), que analizaron como moderadora el tamaño muestral, el tipo de medida del resultado (post y/o seguimiento) y el tiempo entre la intervención y última medida de la VD; el trabajo de Mendez et al. (2002), que tiene como variables moderadoras la vía de reclutamiento y tipo de grupo control; Redondo et al. (2002), que tuvieron en cuenta el diseño aleatorio (si/no), las bajas y el período hasta medida de seguimiento.

[4]Como ejemplo de estas escalas, una ellas es una escala de 9 ítems [0-9] en la que se codifican los siguientes ítems: 1) asignación aleatoria de los participantes (sí = 1; no, con control de variables confundentes = 0.5 y no, sin control de variables confundentes = 0); 2) tipo de diseño (cuasiexperimental = 0.5; experimental = 1); 3) tamaño muestral del grupo tratado en el post-test ([n < 6] = 0, [6 < n < 9] = 0.5, [n > 10] = 1); 4) pérdida de sujetos del grupo tratado en el postest (igual o mayor del 30% = 0, menor del 30% = 0.5, no hubo mortandad = 1); 5) seguimiento (menor de 6 meses = 1, entre 6 y 11 meses = 0.5, 12 o más meses = 1); 6) mediciones de la misma variable dependiente en el pretest y postest (0, 0.5, 1); 7) Calidad de los instrumentos de evaluación; 8) Uniformidad en el tratamiento (no = 0, sí = 1); 9) Ceguera de los evaluadores y de los participantes (0 = no existe ciego, 0.5 = simple ciego, 1 = doble ciego).

Referencias

Las referencias señaladas con asterisco pertenecen a estudios de metaanálisis.

American Psychological Association. (2020). *Publication Manual of the American Psychological Association* (7th ed.). American Psychological Association.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M. y Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist, 73*(1), 3-25. https://doi.org/10.1037/amp0000191

*Arias, E., Arce, R. y Vilariño, M. (2013). Better intervention programmes: A meta-analytic review of effectiveness. *Psychosocial Intervention, 22*(2), 153-160. https://doi.org/10.5093/in2013a18

Atkinson, K. M., Koenka, A. C., Sanchez, C. E., Moshontz, H. y Cooper, H. (2015). Reporting standards for literature searches and report inclusion criteria: Making research syntheses more transparent and easy to replicate. *Research Synthesis Methods, 6*(1), 87-95. https://doi.org/10.1002/jrsm.1127

*Beelmann, A. y Lösel, F. (2006). Child social skills training in developmental crime prevention: Effects on antisocial behavior and social competence. *Psicothema, 18*(3), 603-610.

Botella, J. y Duran, J. I. (2019). A meta-analytical answer to the crisis of confidence of Psychology. *Anales De Psicología, 35*(2), 350-356. https://doi.org/10.6018/analesps.35.2.345291

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

Chacón-Moscoso, S., Sanduvete-Chaves, S. y Sánchez-Martín, M. (2016). The development of a checklist to enhance methodological quality in intervention programs. *Frontiers in Psychology, 7*, 1811. https://doi.org/10.3389/fpsyg.2016.01811

Codray, D. S. y Murphy, P. (2009). Research synthesis and public policy. En H. Cooper, L. V., Hedges y J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed.). Russell Sage Foundation.

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., Petticrew, M. y Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment,7*(27), 1-179. https://doi.org/10.3310/hta7270

Dieckmann, N. F., Malle, B. F. y Bodner, T. E. (2009). An empirical assessment of meta- analytic practice. *Review of General Psychology, 13*(2), 101-115. https://doi.org/10.1037/a0015107

*Espada, J. P., Gonzálvez, M. T., Orgilés, M., Lloret, D. y Guillén-Riquelme, A. (2015). Meta-analysis of the effectiveness of school substance abuse prevention programs in Spain. *Psicothema, 27*(1), 5-12.

*Espada, J. P. Méndez, X, Botvin, G. J., Gilbert, J. Botvin, Griffin, K. W., Orgilés, M. y Rosa-Alcázar, A. I. (2002). ¿Éxito o fracaso de la prevención del abuso de drogas en el contexto escolar? Un meta-análisis de los programas en España. *Psicología Conductual, 10* (3), 581-604.

*Garrido, V., Morales, L. A. y Sánchez-Meca, J. (2006). What works for serious juvenile offenders? A systematic review [¿Qué funciona con los delincuentes juveniles graves? Una revisión sistemática]. *Psicothema, 18*(3), 611-619.

Harwell, M. y Maeda, Y. (2008). Deficiencies of reporting in meta-analyses and some remedies. *The Journal of Experimental Education, 76*(4), 403-430. https://doi.org/10.3200/JEXE.76.4.403-430

Herbison, P., Hay-Smith, J. y Gillespie, W. J. (2006). Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology, 59*(12), 1249-1256. https://doi.org/10.1016/j.jclinepi.2006.03.008

Hohn, R. E., Slaney, K. L. y Tafreshi, D. (2018). Primary study quality in psychological meta-analyses: An empirical assessment of recent practice. *Frontiers in Psychology, 9*, 2667. https://doi.org/10.3389/fpsyg.2018.02667

*Holloway, K. R., Bennett, T. H. y Farrington, D. P. (2006). The effectiveness of drug treatment programs in reducing criminal behavior: A meta-analysis [La efectividad de los programas de tratamiento de la drogadicción en la reducción de la delincuencia: Un metaanálisis]. *Psicothema, 18*(3), 620-629.

Johnson, B. T., Low, R. E. y MacDonald, H. V. (2015) Panning for the gold in health research: Incorporating studies' methodological quality in meta-analysis. *Psychology & Health, 30*(1), 135-152. https://doi.org/10.1080/08870446.2014.953533

*Killias, M. y Villetaz, P. (2008). The effects of custodial vs non-custodial sanctions on reoffending: Lessons from a systematic review. *Psicothema, 20*(1), 29-34.

Juni, P., Witschi, A., Bloch, R. y Egger, M. (1999). The hazards of scoring the quality of clinical trials for meta-analysis. *Journal of the American Medical Association, 282*(11), 1054-1060. https://doi.org/10.1001/jama.282.11.1054

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

Losilla, J. M., Oliveras, I., Marín-García, J. A. y Vives, J. (2018). Three risk of bias tools lead to opposite conclusions in observational research synthesis. *Journal of Clinical Epidemiology, 101,* 61-72. https://doi.org/10.1016/j.jclinepi.2018.05.021

*Martínez, M. P., Miró, E. y Sánchez, A. I. (2016). Beneficios clínicos globales de la terapia cognitivo conductual para el insomnio y de la terapia basada en conciencia plena aplicadas a la fibromialgia: revisión sistemática y metaanálisis. *Psicología Conductual, 24*(3), 459-480.

*Méndez-Carrillo, X., Orgilés-Amorós, M. y Rosa-Alcázar, A. I. (2005). Los tratamientos psicológicos en la fobia a la oscuridad: una revisión cuantitativa [The psychological treatments for the phobia of darkness: A quantitative review]. *Anales de Psicología, 21*(1), 73-82.

*Méndez, X., Rosa-Alcázar, A. I., Montoya, M. Espada, J. P., Olivares, J. y Sánchez-Meca, J. (2002) Tratamiento psicológico de la depresión infantil y adolescente: ¿evidencia o promesa? *Psicología Conductual, 10*(3), 563-580.

*Méndez, X., Rosa-Alcázar, A. I. y Orgilés, M. (2005). Eficacia diferencial de los tratamientos psicológicos en la fobia a los animales: un estudio meta-analítico. *Psicothema, 17*(2), 219-226.

*Navarro-Bravo, B., Párraga-Martínez, I., Hidalgo, J. L. T., Andrés-Pretel, F. y Rabanales- Sotos, J. (2015). Group cognitive-behavioral therapy for insomnia: A meta-analysis. *Anales de Psicología, 31*(1), 8-18. https://doi.org/10.6018/analesps.31.1.168641

*Olivares, J. O., Rosa-Alcazar, A. I., Caballo, V. E., García-López, L. J., Amorós, M. O. y López-Gollonet, C. (2003). El tratamiento de la fobia social en niños y adolescentes: una revisión meta-analítica. *Psicología Conductual, 11*(3), 599-622

*Olivares, J., Rosa-Alcazar, A. I., Piqueras, J. A., Sánchez-Meca, J., Méndez, X. y García- López, L. (2002). Timidez y fobia social en niños y adolescentes: un campo emergente. *Psicología Conductual, 10*(3), 523-542.

*Olivares, J., Sánchez-Meca, J. y Rosa-Alcazar, A. I. (1999). Eficacia de las intervenciones conductuales en problemas de ansiedad en España. *Psicología Conductual, 7*(2), 283- 300.

*Orgilés, M., Rosa-Alcazar A. I., Santacruz, I., Méndez, X., Olivares, J. y Sánchez Meca, J. (2002) Tratamiento psicológico en la infancia y adolescencia: una revisión de su eficacia desde el meta-análisis. *Psicología Conductual, 10*(3), 481-502.

*Perestelo-Perez, L., Barraca, J., Peñate, W., Rivero-Santana, A. y Alvarez-Perez, Y. (2017). Mindfulness-based interventions for the treatment of depressive rumination: Systematic review and meta-analysis. *International Journal of Clinical and Health Psychology, 17*(3), 282-295. https://doi.org/10.1016/j.ijchp.2017.07.004

Protogerou, C. y Hagger M. S. (2019) A Case for a study quality appraisal in survey studies in psychology. *Frontiers in Psychology, 9,* 2788. https://doi.org/10.3389/fpsyg.2018.02788

*Redondo-Illescas, S., Sánchez-Meca, J. y Genovés, V. G. (2002). Los programas psicológicos con delincuentes y su efectividad: la situación europea. *Psicothema, 14*(supl.), 164-173.

*Rosa-Alcázar, A. I., Ingles, C. I., Olivares, J., Espada, J. P., Sánchez-meca, J. y Méndez, X. (2002). Eficacia del entrenamiento en habilidades sociales con adolescentes: de menos a más. *Psicología Conductual, 10*(3), 543-561.

*Rosa-Alcázar, A. I. R., Rodríguez, J. O. y Sánchez Meca, J. (1999). Meta-análisis de las intervenciones conductuales de la enuresis en España [Meta-analysis of behavioural interventions of enuresis in Spain]. *Anales de Psicología, 15*(2), 157-167.

*Rosa-Alcázar, A. I., Sánchez-Meca, J. y López-Soler, C. (2010). Tratamiento psicológico del maltrato físico y la negligencia en niños y adolescentes: un meta-análisis [Psychological treatment of physical maltreatment and negligence in children and adolescents: A meta-analysis]. *Psicothema, 22*(4), 627-633.

*Rosa-Alcázar, A., Sánchez-Meca, J., Rosa-Alcázar, Á, Iniesta-Sepúlveda, M., Olivares-Rodríguez, J. y Parada-Navas, J. (2015). Psychological treatment of obsessive-compulsive disorder in children and adolescents: A meta-analysis. *The Spanish Journal of Psychology, 18*, e20. https://doi.org/10.1017/sjp.2015.22

Rubio-Aparicio, M., Sánchez-Meca, J., Marín-Martínez, F. y López-López, J. A. (2018). Recomendaciones para el reporte de revisiones sistemáticas y meta-análisis. *Anales de Psicología, 34*(2), 412-420. https://doi.org/10.6018/analesps.34.2.320131

*Rueda-Sánchez, M. I. y López-Bastida, P. (2016). Efectos de la intervención en conciencia morfológica sobre la lectura, escritura y comprensión: Meta-análisis [Effects of morphological awareness training on reading, writing and comprehension: Meta-analysis]. *Anales de Psicología, 32*(1), 60-71. https://doi.org/10.6018/analesps.32.1.196261

*Sánchez-Meca, J. S., Martínez, F. M., Rodríguez, J. O. y Alcázar, A. I. R. (1999). Variables influyentes en el tratamiento de la adicción al tabaco. Un estudio de las tasas de abstinencia en España. *Psicología Conductual, 7*(2), 301-321.

*Sánchez Meca, J., Olivares Rodríguez, J. y Rosa-Alcázar, A. I. (1998). El problema de la adicción al tabaco: meta-análisis de las intervenciones conductuales en España [The problem of tobacco addiction: Meta-analysis of behavioural interventions in Spain]. *Psicothema, 10*(3), 535-549.

*Sánchez-Meca, J., Rosa-Alcázar, A. I. y López-Soler, C. (2011). The psychological treatment of sexual abuse in children and adolescents: A meta-analysis. *International Journal of Clinical and Health Psychology, 11*(1), 67-93.

*Sánchez-Meca, J., Rosa-Alcázar, A. I. y Olivares Rodríguez, J. (1999). Las técnicas cognitivo-conductuales en problemas clínicos y de salud: meta-análisis de la literatura española [Cognitive-behavioral techniques in clinical and health problems: Meta-analysis of Spanish literature]. *Psicothema, 11*(3), 641-654.

*Sánchez-Meca, J. S., Rosa-Alcázar, A. I. y Rodríguez, J. O. (2004). El tratamiento de la fobia social específica y generalizada en Europa: Un estudio meta-analítico. *Anales de Psicología, 20*(1), 55-68.

*Santacruz, I., Orgilés, M., Rosa-Alcázar, A. I., Sánchez-Meca, J., Méndez, X. y Olivares, J. (2002). Ansiedad generalizada, ansiedad por separación y fobia escolar: el predominio de la terapia cognitivo-conductual. *Psicología Conductual, 10*(3), 503-522.

*Schmucker, M. y Lösel, F. (2008). Does sexual offender treatment work? A systematic review of outcome evaluations. *Psicothema, 20*(1), 10-19.

Shadish, W. R., Cook, T. D. y Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

*Soldino, V. y Carbonell-Vayá, E. J. (2017). Effect of treatment on sex offenders' recidivism: A meta-analysis. *Anales de Psicología, 33*(3), 578–588. https://doi.org/10.6018/analesps.33.3.267961

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savovic, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chen, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. T. (2016). ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ, 355*, i4919. https://doi.org/10.1136/bmj.i4919

*Tong, L. S. y Farrington, D. P. (2008). Effectiveness of «reasoning and rehabilitation» in reducing reoffending. *Psicothema, 20*(1), 20-28.

*Torre-Luque, A., Gambara, H., López, E. y Cruzado, J. A. (2016). Psychological treatments to improve quality of life in cancer contexts: A meta-analysis. *International Journal of Clinical and Health Psychology, 16*(2), 211-219. https://doi.org/10.1016/j.ijchp.2015.07.005

Valentine, J. C. (2009). Judging the quality of primary research. En H. Cooper, L. V. Hedges y J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis.* Russell Sage Foundation.

Valentine, J. C. y Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The study design and implementation assessment device (Study DIAD). *Psychological Methods, 13*(2), 130-149. https://doi.org/10.1037/1082-693 989X.13.2.130

Vicente, M. V., Brage, L. B., del Carmen, O., Socías, M. y Fernández, J. A. (2017). Meta-analysis of family-based selective prevention programs for drug consumption in adolescence. *Psicothema, 29*(3), 299-305.

Wells, K. y Little, J. (2009) Study quality assessment. in systematic reviews of research on intervention effects. *Research on Social Work Practice, 19*(1), 52-62. https://doi.org/10.1177/1049731508317278

Wilson, S. J. y Lipsey, M. W. (2006). The effects of school-based social information processing interventions on aggressive behavior. *Campbell Systematic Reviews, 2*(1), 1-42. https://doi.org/10.4073/csr.2006.5