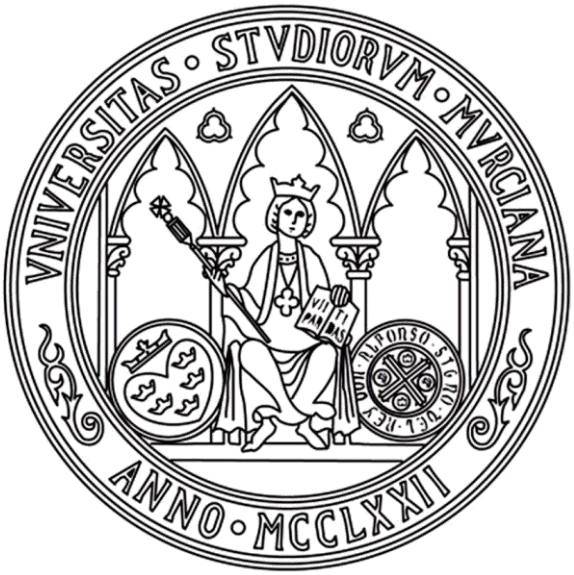


UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Optimizing Statistical Methods in Meta-Analysis

Optimizando Métodos Estadísticos en Meta-análisis

Dña. María Rubio Aparicio

2018

UNIVERSITY OF MURCIA

INTERNATIONAL SCHOOL OF DOCTORAL STUDIES

Optimizing Statistical Methods in Meta-analysis

Optimizando Métodos Estadísticos en Meta-análisis

Doctoral Thesis

Author: Dª. María Rubio Aparicio
Supervisors: Dr. Julio Sánchez-Meca (University of Murcia)
Dr. Fulgencio Marín-Martínez (University of Murcia)
Dr. José Antonio López-López (University of Bristol)

Murcia, 2018

Trabajo financiado con un Contrato Predoctoral
del Ministerio de Economía y Competitividad

Trabajamos por la eficiente acumulación del conocimiento
(Rosenthal, 1991)

Agradecimientos

A mis directores de tesis, **Julio Sánchez Meca**, **Fulgencio Marín Martínez** y **José Antonio López López**, por la confianza depositada en mí y por animarme a superarme constantemente. Sin sus correcciones, experiencia y lecciones no habría sido posible la elaboración de este trabajo. Gracias por vuestro cariño y por haberme impregnado con vuestro entusiasmo en las tareas de investigación y docencia. Grandes profesionales y mejores personas.

A mis compañeros, **Aurora Orenes**, **Juan José Madrid**, **Noelia Sánchez**, **Violeta Pina** y **Violeta Provencio**. Gracias por vuestros consejos, por los momentos compartidos, y por llenar de alegría mi día a día en la facultad.

A mis compañeros de área, **Manolo Ato**, **Ana Benavente**, **Mª Dolores Hidalgo**, **Antonio Velandrino** y **Rafael Rabadán**. Gracias en especial al equipo humano que forma la Unidad de Meta-análisis, **José Antonio López Pina**, **Rosa Nuñez** y **Juan José López,** por vuestro apoyo y cariño.

A **mi familia**, en especial a mi madre, por haberme proporcionado la mejor educación y lecciones de vida, y a mi hermana, por su paciencia y compresión.

A **Javi**, por hacerme ver la vida de una forma diferente y por confiar en mis decisiones.

A **mis amigas**, por sus mensajes de apoyo y por sacarme una sonrisa en los momentos en los que más lo he necesitado.

En definitiva, gracias a todos los que me habéis acompañado durante estos cuatro años.

Index

Resumen……... 1

1. Introduction…… 7

1.1. Meta-analysis……..7

1.2. Phases of a meta-analysis……... 8

1.2.1. Defining the research question…….. 8

1.2.2. Literature research…… 8

1.2.3. Coding of studies……...9

1.2.4. Calculating an effect-size index…….. 10

1.2.5. Statistical analysis and interpretation…….. 11

1.2.6. Publication……... 13

1.3. Monte Carlo studies in meta-analysis……...14

1.4. Optimizing statistical methods in meta-analysis…….. 15

2. Study 1: "A methodological review of meta-analyses of the effectiveness of clinical psychology treatments"……17

2.1. Introduction…….. 17

2.1.1. Types of standardized mean differences…… 19

2.2. Methodology…….24

2.2.1. Search procedure and selection criteria of the meta-analyses…… 24

2.2.2. Data extraction…….25

2.2.3. Meta-analytic calculations……...25

2.2.4. Data analysis……28

2.3. Results…….. 29

2.3.1. Characteristics of the meta-analyses……29

2.3.2. Number of studies……32

2.3.3. Effect sizes distribution……... 35

2.3.4. Sample size distribution……...36

2.3.5. Correlation between effect sizes and sample sizes……...37

2.3.6. Heterogeneity……...37

2.4. Discussion……38

2.4.1. Limitations of the study……...42

2.4.2. Recommendations overview……42

2.5. Conclusions…….. 44

3. Study 2: "Estimating an overall effect size in random-effects meta-analysis when the distribution of random effects departs from normal"…….45

3.1. Introduction…….. 45

3.1.1. The random-effects model……...46

3.2. Methods to estimate an overall effect size……49

3.2.1. The fixed-effect model…… 49

3.2.2. The random-effects model……...50

3.2.3. Heterogeneity variance estimators……...53

3.3. Method of the simulation study…… 54

3.4. Results…….. 61

3.4.1. Bias of the average effect estimators……...62

3.4.2. Mean squared error of the average effect estimators……...63

3.4.3. Coverage probability of the CIs……...65

3.4.4. Width of the CIs…….. 67

3.4.5. Bias of the standard error…… 68

3.5. Discussion…….70

4. Study 3: "Analysis of categorical moderators in mixed-effects meta-analysis: Consequences of using pooled versus separate estimates of the residual between-studies variances"……..74

4.1. Introduction…….. 74

4.1.1. Mixed-effects model……75

4.1.2. Omnibus test of between-groups differences…….. 76

4.1.3. Estimating the residual between-studies variance……... 77

4.1.4. An example……..80

4.1.5. Purpose of the study…… 82

4.2. Method of the simulation study…… 85

4.3. Results…….. 88

4.3.1. Type I error rate……... 88

4.3.2. Statistical power…….. 90

4.4. Discussion…….93

4.4.1. Limitations and future research……... 96

5. Study 4: "A comparison of hypothesis tests for categorical moderators in meta-analysis using mixed-effects models"……...97

5.1. Introduction…….. 97

5.1.1. Tests of between-groups differences……... 99

5.1.2. Estimating the residual between-studies variance……. 100

5.2. Method of the simulation study…….. 102

5.3. Results…… 104

5.3.1. Type I error rate……105

5.3.2. Statistical power…… 108

5.4. Discussion……...112

6. Conclusions……...116

References……120

Appendices……... 134

Appendices in Chapter 2:

Appendix 2A……..134

Appendix 2B……...140

Appendices in Chapter 3:

Appendix 3A……..143

Appendix 3B……...149

Appendices in Chapter 4:

Appendix 4A……..151

Appendix 4B……...153

Appendices in Chapter 5:

Appendix 5A……..157

Appendix 5B……...159

Appendix 5C……...164

Resumen

La producción científica ha crecido exponencialmente a lo largo de las últimas décadas en la mayoría de campos de investigación. Como consecuencia de este notable aumento de conocimiento, las tareas de síntesis y revisión han ganado importancia para entender el estado de la cuestión sobre un determinado tema. En este contexto, el meta-análisis ha emergido como una metodología que permite a los investigadores integrar cuantitativamente los resultados de un conjunto de estudios primarios sobre un tópico común. Desde que Glass (1976) acuñó el término meta-análisis para referirse a este tipo de investigación metodológica hasta la actualidad, los meta-análisis han ganado popularidad en diferentes áreas de investigación como la educación, la psicología y las ciencias de la salud. Los tres principales objetivos en meta-análisis son estimar un tamaño del efecto medio a partir de los estudios primarios, estudiar la heterogeneidad de los tamaños del efecto en torno a ese tamaño del efecto medio, y buscar características de los estudios que pudieran explicar al menos parte de la variabilidad exhibida en los tamaños del efecto individuales (Botella y Gambara, 2002, 2006; Botella y Sánchez-Meca, 2015; Hedges y Olkin, 1985, Sánchez-Meca y Marín-Martínez, 2010).

Los meta-análisis deben ser llevados a cabo con el mismo rigor científico que los estudios empíricos, es decir, objetividad, sistematización y replicabilidad. Debido a que el objetivo del meta-análisis es integrar estudios individuales, su unidad de análisis es el estudio, mientras que en la investigación primaria la unidad de análisis es el sujeto. Por tanto, el tamaño muestral en meta-análisis es el número de estudios.

Actualmente, la mayor parte de las conclusiones sobre la acumulación del conocimiento en psicología están basadas en meta-análisis. La psicología basada en la evidencia es una herramienta metodológica que enfatiza la importancia de la evidencia científica en la práctica psicológica. El objetivo de esta estrategia consiste en modificar la forma de trabajo de los psicólogos, pues estos profesionales aplicados deben tener en cuenta la mejor evidencia científica para tomar sus decisiones sobre un cierto problema

(Sánchez-Meca y Botella, 2010). En este sentido, el meta-análisis es una metodología esencial que ayuda a los psicólogos aplicados a tomar decisiones bien informadas basadas en información científica.

Un meta-análisis es una investigación científica y por tanto, su estructura es muy similar a la de cualquier estudio empírico, aunque con ciertas especificaciones. Básicamente, un meta-análisis se lleva a cabo siguiendo las siguientes seis fases: (1) Definición la pregunta de investigación, (2) búsqueda de estudios, (3) codificación de los estudios, (4) cálculo del índice del tamaño del efecto, (5) análisis estadístico, y (6) publicación (Botella y Gambara, 2002; Lipsey y Wilson, 2001; Sánchez-Meca y Marín-Martínez, 2010). Esta tesis se centra en la fase relacionada con el análisis estadístico, concretamente en los métodos estadísticos aplicados en meta-análisis.

Durante los últimos 30 años, se ha desarrollado una intensa actividad para mejorar y extender la aplicabilidad de la metodología meta-analítica. Uno de los campos que ha generado más investigación es el de los métodos estadísticos aplicados en meta-análisis. Se han realizado numerosos estudios de simulación Monte Carlo para investigar qué técnicas y procedimientos son los más adecuados dadas las características de una base meta-analítica. En el contexto del meta-análisis, los estudios de simulación Monte Carlo son especialmente necesarios cuando la teoría axiomática no es capaz de dar respuesta a los problemas relativos al funcionamiento de los procedimientos meta-analíticos. Además, el meta-análisis es una metodología relativamente joven y por ello, es también necesario el desarrollo de estudios de este tipo para perfeccionar las técnicas usualmente aplicadas.

En la literatura meta-analítica, existe consenso en considerar el modelo de efectos aleatorios y el modelo de efectos mixtos como los que mejor se ajustan a las características de los meta-análisis aplicados en las ciencias empíricas, en general, y en la psicología, en particular. Los métodos inferenciales aplicados bajo estos modelos muestran un funcionamiento deficiente bajo algunas condiciones. En este sentido, se han propuesto nuevos métodos para intentar mejorar los habituales métodos estándar, aunque aún es necesaria más investigación para estudiar su funcionamiento y así, poder responder a interrogantes todavía presentes.

Esta tesis doctoral está compuesta por tres estudios de simulación Monte Carlo (Capítulos 3, 4 y 5) que comparan procedimientos y técnicas aplicados en meta-análisis

bajo los modelos de efectos aleatorios y efectos mixtos, con el objetivo último de proporcionar recomendaciones a los meta-analistas aplicados sobre qué métodos son los más óptimos bajo ciertas condiciones.

Como paso previo, en el Capítulo 2 presentamos una revisión metodológica de 54 meta-análisis sobre la efectividad de tratamientos psicológicos en el ámbito de la psicología clínica que emplearon como índice del tamaño del efecto la diferencia de medias estandarizada. Uno de los objetivos de esta revisión fue ayudarnos, tanto a nosotros como a otros investigadores, a diseñar nuestros futuros estudios de simulación Monte Carlo manipulando condiciones de la forma más realista posible. En este estudio analizamos la distribución de los tamaños muestrales en los estudios de cada meta-análisis, la distribución de los tamaños del efecto en cada meta-análisis, la distribución de los valores de la varianza inter-estudios, y las correlaciones de Pearson entre el tamaño del efecto y el tamaño muestral de cada meta-análisis. Los resultados se presentan en función del tipo de diferencia media estandarizada: diferencia media tipificada calculada con las puntuaciones del postest, cambio medio tipificado del pretest al postest, y diferencia de cambio medio tipificado entre grupos. Uno de los hallazgos más interesantes encontrados en esta revisión fue que la mayoría de los meta-análisis usaron la diferencia media estandarizada a partir de las puntuaciones del postest para comparar dos grupos (por ejemplo, grupo experimental y grupo control), y aunque la mejor opción para comparar dos grupos es la diferencia de cambios medios tipificados, este índice rara vez se utilizó. Por otro lado, los resultados sugieren la existencia de un número de estudios relativamente bajo, gran cantidad de heterogeneidad en los tamaños del efecto, violación del supuesto de normalidad en la distribución de los tamaños del efecto, y correlaciones positivas y negativas entre los tamaños del efecto y los tamaños muestrales. Por último, también encontramos que los tres cuartiles de la distribución de los tamaños del efecto medios para los meta-análisis que usaron tanto la diferencia media tipificada con las puntuaciones del postest como la diferencia de cambios medios tipificados eran similares al criterio propuesto por Cohen (1988), mientras que los tres cuartiles de la distribución de los tamaños el efecto medios en los meta-análisis que usaron el cambio medio tipificado del pretest al postest fueron más grandes que los valores propuestos por Cohen (1988). Por tanto, de este resultado concluimos que el análisis de la distribución de los tamaños del efecto de los meta-análisis proporciona una guía específica y contextualizada para la interpretación de la significación clínica de los diferentes tipos de diferencias

medias tipificadas dentro del campo de la psicología clínica. El primer estudio de simulación, descrito en el Capítulo 3, tiene como objetivo estudiar la influencia del incumplimiento del supuesto de normalidad de la distribución de efectos paramétricos cuando se calcula el tamaño del efecto medio y su intervalo de confianza bajo el modelo de efectos aleatorios. En este trabajo se compara el funcionamiento de varios métodos meta-analíticos de efectos aleatorios (método estándar, método de Hartung, método de verosimilitud perfil, y el bootstrapping no paramétrico), aplicando a su vez tres estimadores de la varianza inter-estudios (DerSimonian y Laird, máxima verosimilitud restringida y el estimador empírico de Bayes). Los métodos se evaluaron en términos de sesgo y media cuadrática de error de las estimaciones del efecto medio, cobertura empírica y amplitud confidencial de los intervalos de confianza, y sesgo del error estándar. Los resultados sugieren que los métodos de efectos aleatorios son robustos a las desviaciones del supuesto de normalidad, siendo el método de Hartung y el método de verosimilitud perfil los que alcanzaron un mejor funcionamiento bajo condiciones subóptimas.

Los estudios de simulación presentados en los Capítulos 4 y 5 de esta tesis se centran en el análisis de moderadores cualitativos - análisis de subgrupos - aplicando el modelo de efectos mixtos. Concretamente, el segundo estudio de simulación en el Capítulo 4 compara el impacto de dos procedimientos de estimación de la varianza inter-estudios residual, estimaciones separadas en cada categoría del moderador versus estimación conjunta a partir de todas las categorías del moderador, sobre el funcionamiento estadístico del estadístico estándar de heterogeneidad intergrupos, la prueba Q_B. El estimador de la varianza inter-estudios residual aplicado fue el método de DerSimonian y Laird. El funcionamiento de los métodos estudiados se evaluó en términos de error Tipo I y potencia estadística. Los resultados de este estudio sugieren un funcionamiento similar de ambos procedimientos de estimación de la varianza inter-estudios residual siempre que el número de estudios del meta-análisis sea de al menos 20 estudios y que además la distribución del número de estudios sea equilibrada en las categorías del moderador. Por el contrario, cuando el número de estudios se distribuye de forma desequilibrada, las consecuencias prácticas de tener varianzas inter-estudios residuales heterogéneas en las categorías del moderador son más evidentes. Bajo estas condiciones el procedimiento de estimación más adecuado es la estimación conjunta, a menos que las varianzas inter-estudios residuales de cada categoría sean claramente

diferentes y que haya un suficiente número de estudios en cada categoría para lograr estimaciones separadas concretas.

El tercer y último estudio de simulación explicado en el capítulo 5 también se centra en el análisis de subgrupos bajo el modelo de efectos mixtos. Sin embargo, en este estudio damos un paso más y evaluamos el funcionamiento en análisis de subgrupos del conocido y mejorado método de Knapp y Hartung (2003) para moderadores continuos. Por tanto, en este estudio examinamos ambos procedimientos de estimación de la varianza inter-estudios residual (estimación conjunta versus estimación separada) en combinación con dos métodos para estudiar la significancia estadística de un moderador, el estadístico estándar de heterogeneidad intergrupos (la prueba Q_B), y el estadístico mejorado de Hartung (la prueba *F*). Además, se aplicaron tres métodos diferentes para estimar la varianza inter-estudios residual (DerSimonian y Laird, máxima verosimilitud restringida, y Paule y Mandel). El funcionamiento de los diferentes métodos se evaluó también en términos de error Tipo I y potencia estadística. Los resultados de este estudio sugieren que el estadístico mejorado, la prueba *F*, calculado estimando la varianza inter-estudios residual de forma conjunta a través de las categorías es la mejor opción en la mayoría de las condiciones estudiadas, aunque la prueba *F* calculada estimando la varianza inter-estudios residual de forma separada en cada categoría es preferible si las varianzas residuales de cada categoría son heterogéneas y el número de estudios de cada categoría se distribuye desequilibradamente. Por otro lado, los resultados mostraron el mismo patrón para todos los tres estimadores utilizados de la varianza inter-estudios residual. La principal conclusión de este estudio es que el método mejorado de Hartung, la prueba *F*, manifiesta ventajas sobre el método estándar, la prueba Q_B, y que la elección del procedimiento de estimación de la varianza inter-estudios residual (separada versus conjunta) debería hacerse tras examinar las características de la base meta-analítica.

La interpretación conjunta de los hallazgos encontrados en los cuatro estudios que forman la tesis permite ofrecer a los investigadores y/o meta-analistas aplicados una serie de recomendaciones. Por ejemplo, una de las principales aportaciones de este trabajo es la presentación a la comunidad científica de una guía contextualizada para la interpretación de los tamaños del efecto de la familia *d* en el ámbito de la psicología clínica, además de ayudar en el diseño de futuros estudios de simulación Monte Carlo utilizando como condiciones o parámetros manipulados las características metodológicas de los 54 meta-análisis. Por otro lado, otro hallazgo importante es la robustez de la

mayoría de los métodos meta-analíticos ante el incumplimiento del supuesto de normalidad en la distribución de los tamaños del efecto paramétricos. Para terminar, los resultados permiten aconsejar el uso del método mejorado de Hartung en análisis de subgrupos, estimando la varianza inter-estudios residual de forma conjunta a partir de las categorías del moderador, pero siempre tomando esta decisión tras examinar las características de los estudios que van a conformar tu meta-análisis.

Chapter 1

Introduction

1.1. Meta-analysis

Research production has exponentially grown along the last decades in most scientific fields. As a consequence, the tasks of synthesis and revision are increasingly important in order to figure out the state of the art in a specific phenomenon. In this context, meta-analysis has emerged as a methodology that allows researchers to integrate quantitatively the results from a set of primary studies on a same topic. Since Glass (1976) coined the term meta-analysis to refer to this research methodology, meta-analyses have been gaining popularity in many different research areas such as education, psychology, and health care. The three main statistical objectives in a meta-analysis are to estimate the mean effect size through the primary studies, to assess the heterogeneity of the effect size estimates around the mean effect size, and to search for moderators that can explain part of the heterogeneity among the individual effect size estimates. In the behavioral, social, educational, and healthcare sciences, these moderators include the differential characteristics of the studies, such as the type of design, characteristics of the participant samples, or types of interventions (Botella & Gambara, 2006; Hedges & Olkin, 1985; Rosenthal, 1991; Sánchez-Meca & Marín- Martínez, 2010).

Meta-analyses must be carried out with the same scientific rigor as that demanded for empirical studies, that is to say, objectivity, systematization and replicability. As meta-analysis aims to integrate studies, the analysis unit is the study, whereas in primary

research the analysis unit is the subject. Thus, the sample size in meta-analysis is the number of studies.

Nowadays, most conclusions about cumulative knowledge in psychology are based on meta-analyses. Evidence-based psychology is a methodological tool which emphasizes the importance of scientific evidence to inform psychological practice. This approach aims to modify the way psychologists work so that professionals take into consideration the best scientific evidence to make their decisions (Sánchez-Meca & Botella, 2010). In this vein, meta-analyses are an essential methodology to synthesize the scientific evidence available on a given research question at a give time point.

1.2. Phases of a meta-analysis

A meta-analysis is a scientific investigation and, consequently, it involves carrying out the same outline as in an empirical study. However, a few specificities need to be mentioned. Basically, a meta-analysis can be conducted following six phases: (1) Defining the research question, (2) literature search, (3) coding of studies, (4) calculating an effect-size index, (5) statistical analysis, and (6) publication (Cooper, Hedges, & Valentine, 2009; Lipsey & Wilson, 2001; Sánchez-Meca & Marín-Martínez, 2010).

1.2.1. Defining the research question

The background in a meta-analysis consists of defining clearly and objectively the research question. Thus, the constructs whose relationships are intended to be studied must be specified, as well as all variables implied in these relationships, including not only dependent and independent variables, but also some potential moderator variables.

1.2.2. Literature search

Once the research question has been formulated, the next step consists of defining the selection criteria that the primary studies must fulfil in order to be included in the meta-analysis. Although the selection criteria will depend on the question addressed in the meta-analysis, there are several criteria that should be present in any meta-analysis

such as range of years considered, design type in the empirical studies, language restrictions, and a minimum sample size.

In order to search for the studies that fulfil the selection criteria, a combination of several formal and informal searching strategies should be the best option. Electronic bibliographic databases (e.g., PsycINFO, MedLine, ERIC, Google Scholar) should be consulted including the keywords used and how they were combined. To warrant the maximum comprehensiveness in this process, the formal search strategy is usually complemented by carrying out manual searches in specific journals and books for the topic of interest, by checking the references listed in the selected studies, and by contacting recognized researchers in the field.

1.2.3. Coding of studies

Once we have retrieved the primary studies to be included in the meta-analysis, the next step is to record all relevant study characteristics. To this end, a codebook and a protocol for registering the characteristics of the studies must be produced. Furthermore, the authors must make available the codebook and the protocol for the scientific community in order to warrant the transparency and replicability of the coding process. The relevant information to be extracted from each primary study includes numerical variables that will be used in the main analyses (see next section), but also potential moderators of the association of interest. Although that list of potential moderators will vary from one meta-analysis to another, three broad categories of moderator variables can be distinguished: methodological, substantive, and extrinsic variables. Substantive characteristics are those related to the research question of the meta-analysis, including sociodemographic characteristics of the sample. Methodological variables are characteristics of the designs and methods of the studies. It is advisable to register methodological variables that allow assessing the methodological quality of the studies, such as random versus non-random assignment of participants to the groups, experimental mortality and the use of blinded evaluators in assessing the outcomes. The meta-analyst can then make the decision to eliminate studies that do not have a minimum of methodological quality, or to include/exclude them in sensitivity analyses. Finally, extrinsic variables are those characteristics that have nothing to do with the research enterprise so that, in principle, they should not be related at all with the study results.

These include publication year of the study, publication status (published or unpublished), and the educational profile of the main author.

In practice, the process of coding studies is often complex because the information reported in some primary studies may be incomplete or ambiguous. Therefore, the reliability of the coding process should be analysed. For that purpose, all or a random sample of the primary studies should be coded independently by two (or more) previously trained coders. The reliability can be assessed using indices such as intra-class correlations and kappa coefficients for continuous and categorical moderators, respectively.

1.2.4. Calculating an effect-size index

In addition to coding and recording the moderator variables of the studies, an essential issue in meta-analysis is to calculate a quantitative index that summarizes the results of each study in a common metric. It is very common that the studies included in the meta-analysis have measured the effects of treatment with different instruments (e.g., different psychological tests), so that their results are not directly comparable. Homogenization of results can be achieved by applying some effect size index. Depending on the study design and the type of dependent variables (e.g. continuous, dichotomous), different effect-sizes indices can be applied. The effect-size indices most frequently used in meta-analysis are grouped into: *d* family, *r* family, and risk indices.

In psychology and related areas, one of the most frequently study design involves a comparison of two groups in a continuous variable. In a two-group design (usually experimental vs. control), the effect size most usually applied from the *d* family is the standardized mean difference, which enables to transform results using different scales into a common scale. The standardized mean difference is defined as the difference between two means divided by a pooled within-group standard deviation: (Hedges & Olkin, 1985; Rubio-Aparicio, Marín-Martínez, Sánchez-Meca, & López-López, in press). In a repeated measures design, in which continuous pretest and posttest measures are registered for a sample of subjects (e.g., before and after the intervention, or before and at follow-up), the standardized mean change is a more appropriateeffect size, calculated as the difference between the pretest and posttest means divided by the pretest standard deviation: (Hedges & Olkin, 1985; Rubio-Aparicio et al., in press). Finally, if

the studies include experimental and control groups with pretest and posttest measures, it is recommendable to use the standardized mean change difference index (see e.g., Morris, 2008; Rubio-Aparicio et al., in press). The family *d* indices will be described with more detail in Chapter 2.

On the other hand, when the primary study applied a correlational design to analyse the degree of association among two variables, a correlation coefficient can be used as the effect-size index (e.g., the Pearson correlation coefficient, its Fisher´s Z transformation, the phi coefficient, the point-biserial correlation coefficient, etc.).

Lastly, when the dependent variables are dichotomous risk indices must be applied: the risk differences defined as the difference between the failure (or success) proportions for two groups, the risk ratio defined as the ratio between two proportions, and the odds ratio, defined as the ratio between the odds of the two groups, are some examples (Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003).

Once the effect-size index most appropriate to the characteristics of the studies has been selected, it is applied to each individual study and its sampling variance is also calculated with the corresponding formulas (e.g., Sánchez-Meca & Marín-Martínez, 2010). In addition, as in process of coding the characteristics of the studies, the computation of the effect sizes must be subjected to an analysis of intercoder reliability.

1.2.5. Statistical analysis and interpretation

Once the information from the studies has been summarized, statistical analyses can be conducted. A preliminary step in the statistical analysis consists of describing the characteristics of the primary studies with the aim of portraiting the "typical" study in the data set, its composition and size.

After descriptive analyses, the first inferential purpose in meta-analysis is to calculate an average effect size and its interval confidence. When computing this average, it iscustomary to apply weighting procedures to give more weight to the effect sizes obtained from the studies with larger sample sizes. The most appropriate weighting method involves using the inverse variance of each effect size estimate as the weighting factor (Cooper et al., 2009; Hedges & Olkin, 1985; Marín-Martínez & Sánchez-Meca, 2010). To accomplish this first objective, two statistical models can typically considered:

the fixed-effect and the random-effects models. Under the fixed-effect model, it is assumed that all studies in the meta-analysis estimate a common population effect size and the only source of variability among the effect sizes is sampling error due to the random selection of participants in each study (Konstantopoulos & Hedges, 2009). Conversely, in the random-effects model it is assumed that each study in the meta-analysis estimates a different population effect size, and that studies are randomly selected from a population of studies. It is also assumed that the corresponding population effect sizes are normally distributed. As a consequence, in the random-effects model, the effect sizes present two sources of variability: between-studies and within-study variability. Nowadays, there is broad consensus that the random effects model is more realistic than the fixed-effect model in most situations, due to the methodological and substantive differences that are typically found among the studies combined in a meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2010; Hedges & Vevea, 1998; Raudenbush, 1994, 2009). The model choice will have an influence on the statistical procedures implemented for integrating the information and on the generalizability of the results (Hedges & Vevea, 1998). Furthermore, there is currently wide consensus on the convenience of applying the improved method proposed by Hartung (1999; IntHout, Ioannidis, & Borm, 2014; Sánchez-Meca & Marín-Martínez, 2008) to compute the confidence interval around the overall effect size estimate.

Secondly, the meta-analyst must assess the heterogeneity of the individual effect sizes around the average effect size. To that aim, the *Q* statistic (Hedges & Olkin, 1985) is often employed to test the null hypothesis that variability among the effect sizes is only due to random sampling error (e.g. there is no true heterogeneity among effect sizes). However, the *Q* test has poor statistical power to detect true heterogeneity among effect sizes when meta-analyses include a small number of studies (Sánchez-Meca & Marín-Martínez, 1997). Thus, it is recommendable to complement the statistical conclusion of the *Q* test with the I^2 index (Higgins & Thompson, 2002), which quantifies the heterogeneity exhibited by effect sizes as a percentage (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006).

If substantial heterogeneity among the effect sizes is found in a meta-analysis (as is often the case), the third step consists of searching for moderator variables that can account for that variability. A general recommendation when conducting such moderator analyses is to adopt a mixed-effects model, in which the effect sizes are taken as a

random-effects variable, and study-level moderator variables – or individual-level moderators, should individual participant data be available – are taken as fixed-effects variables. Moderator analyses can be conducted through meta-regression analyses for continuous variables and weighted ANOVA for qualitative variables, with the improved method proposed by Knapp and Hartung (Knapp & Hartung, 2003; López-López, Botella, Sánchez-Meca, & Marín-Martínez, 2013; Viechtbauer, López-López, Sánchez-Meca, & Marín-Martínez, 2015). It is also recommendable to estimate the proportion of variance accounted for by the moderator variables calculated by means of R^2 (López-López, Marín-Martínez, Sánchez-Meca, Van den Noortgate, & Viechtbauer, 2014).

This dissertation is focused on this phase of a meta-analysis, concretely on the statistical methods applied in meta-analysis.

1.2.6. Publication

Finally, the results of a meta-analysis can be summarised in a report for further publication. Many guidelines have been developed with the aim of helping authors improve the reporting of meta-analyses. Of all of them, the PRISMA Statement (*P*referred *R*eporting *I*tems for *S*ystematic reviews and *M*eta-*A*nalyses; Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009) and the AMSTAR statement (*A*ssessment of *M*ultiple *Sys**T**em**A*tic *R*eviews; Shea et al., 2007) are specially designed for their application in meta-analyses on effectiveness of interventions.

The structure of a meta-analytic report is similar to that of any other scientific paper: Introduction, method, results, discussion, and conclusions. In the introduction the need to carry out a meta-analysis must be justified, together with the definitions of the constructs and variables implied in the research question. Furthermore, the purpose of the meta-analysis must be stated, specifying objectives and hypotheses, if applicable. In the method section, the selection criteria of the studies, the search strategies, the coding process of the study characteristics, the computation of the effect-size index, and the statistical analyses must be outlined. The results section must present some characteristics of the included studies, the effect-size distribution, the mean effect size, the assessment of the heterogeneity, the analyses of moderator variables, additional analyses (e.g., sensitivity analyses and assessment of publication bias), and as a further step, it is advisable to fit an explanatory model including the most relevant moderator variables (if

the number of studies allows it). In addition, including tables and charts (e.g., forest plots, funnel plots) in this section can be helpful to the reader. Finally, in the discussion and conclusions section the main results are presented and discussed in the light of previous meta-analyses. The implications for future research and limitations of the meta-analysis must be outlined.

1.3. Monte Carlo studies in meta-analysis

A simulation study entails drawing (usually large) random samples from a theoretical distribution function with known parameters Simulation studies allow researchers to examine the sampling distribution of one or more parameters of interest that could not be obtained with studies of real data alone. The label 'Monte Carlo method' is used for any empirical study in which randomly generated variables are present. Thus, in a Monte Carlo simulation study, several data sets are independently created by random number generation, using functions based on probability distributions (Burton, Altman, Royston, & Holder, 2006; Schulze, 2004).

In a Monte Carlo simulation study, several steps can be found. Firstly,, the statistical model, the parameters to be estimated and the experimental factors must be defined, based on substantive knowledge. Then, the combinations of the experimental factors and the number of iterations are defined. In a simulation study, these aspects depend on the objectives and available resources (e.g, computational time). In a third step, in each combination of factors (or scenario) random data are generated, and the statistical methods under examination are implemented. This stage is repeated as many times as replicas have been previously defined. Finally, the results obtained must be analysed.

Recent technological advances have resulted in the development of more powerful devices and more efficient algorithms, which has led to a substantial decrease in computational time. This explains the increasingly frequent use of computers to perform statistical analysis. In the context of meta-analysis, Monte Carlo simulation studies are designed to investigate the properties of statistical procedures and techniques usually applied when carrying out a meta-analysis. Monte Carlo simulation studies are particularly necessary when the axiomatic theory is not able to give answer to the issues relative to the performance of different meta-analytic procedures. Further, meta-analysis is a relatively young methodology, and hence it is also necessary to develop Monte Carlo simulation studies in order to investigate which among some newly developed procedures

are most adequate given the characteristics of a meta-analytic database. Nowadays, the scientific community recognizes meta-analysis as one of the methodologies able to offer the highest quality scientific evidence. In this vein, a growing community of researchers are finding out methodological alternatives to help applied meta-analysts conduct meta-analytic reviews. The Meta-analysis Unit of the University of Murcia (http://www.um.es/metaanalysis/), headed by Dr. Julio Sánchez Meca, has been doing research on meta-analysis for more than 25 years. During this time, this team has developed a wide scientific production on the assessment and application of new statistical procedures in meta-analysis, including numerous Monte Carlo simulation studies.

1.4. Optimizing statistical methods in meta-analysis

The last 30 years have seen an intense activity aimed at improving the statistical methods applied in meta-analysis. This chapter provides an overview of the key analysis goals in a meta-analysis. As previously pointed out, nowadays there is consensus in considering random effects and mixed effects models as those that best fit the characteristics of most meta-analyses that are applied in the empirical sciences, in general, and in psychology in particular (Borenstein, Hedges, Higgins & Rothstein, 2009; Cooper et al., 2009; Hedges & Olkin, 1985). Nonetheless, the statistical inference methods usually applied in random and mixed-effects meta-analysis show deficient performance under some conditions. New methods have been proposed that try to outperform that the standards ones, but research is needed to assess their performance. In this vein, the purpose of this dissertation is to carry out Monte Carlo simulation studies to optimize the inferential statistical methods under random and mixed-effects models for their future application.

First, in order to ensure that the manipulated conditions in our Monte Carlo simulation studies are as realistic as possible, a methodological review of meta-analyses of the effectiveness of clinical psychology treatments is carried out in Chapter 2. One of the purposes of this study is to offer a guide for the design of future research studies on the performance of meta-analytic procedures, based on the manipulation of realistic assumptions and parameters. Furthermore, the results of this review allow contextualized interpretation of the effect sizes in the specific area of the evaluation of the effectiveness of clinical psychological treatments.

Two of the main objectives in meta-analysis are to estimate the average effect size and to perform an analysis of moderators to identify sources of heterogeneity among the individual effect size estimates. Chapters 3, 4 and 5 present three Monte Carlo simulation studies examining the performance of several statistical methods available to address those goals. The first simulation study is conducted for the first objective and the second and third simulation study for the second objective.

In Chapter 3, the influence of the departure from the normality assumption in the population effects distribution, when computing an average effect size and a confidence interval (CI) in random-effects meta-analysis, is assessed.

Chapters 4 and 5 are focused on categorical moderator analyses under mixed-effects models. Concretely, the main purpose in Chapter 4 was to compare the impact of two procedures for estimating the residual between-studies variance, separate estimates and pooled estimate in each category of the moderator, on the statistical performance of the standard between-groups heterogeneity statistic. Furthermore, Chapter 5 incorporates the Knapp-Hartung improved method in addition to examining both approaches for estimating the residual between-studies variance on the standard between-groups heterogeneity test.

Finally, some general conclusions and recommendations for future research are provided in Chapter 6.

Chapter 2

Study 1:

"A methodological review of meta-analyses of the effectiveness of clinical psychology treatments"

2.1. Introduction

Meta-analysis is a form of quantitative systematic review in which the results of a series of empirical studies on the same research topic are statistically summarized. When the individual studies report results in different scales (e.g., depression symptoms measured with different instruments), standardized effect size indices are often used to express the results across studies in a common metric. The standardized mean difference is one of the most used effect size indices in studies in which two or more groups are compared on a continuous outcome (Borenstein et al., 2009; Cooper et al., 2009).

The empirical analysis of the methodological characteristics of real meta-analyses in a specific area of study is useful, as it helps to portrait the "typical" meta-analytic review that is conducted in a research field (e.g., number of studies, sample size distribution in the primary studies, and effect size distribution). Furthermore, a methodological review of meta-analyses allows assessing the degree of compliance with model assumptions, such as normal distribution of the effect sizes and independence between the sample sizes and effect sizes.

The aim of the present study was to explore the methodological characteristics of 54 meta-analyses published in high standard journals, which examined the effectiveness of clinical psychological interventions using standardized mean differences as the effect size index. This enabled us to provide a guide for the interpretation and characterization of the meta-analyses in the context of clinical psychology.

As in our study, Levine, Asada, and Carpenter (2009) explored the characteristics of 51 published meta-analyses on topics relevant to communication researchers (e.g., persuasion and interpersonal communication, language intensity effects, or viewing presidential debates). Interestingly, this study revealed a negative correlation between effect size and sample size for most of the meta-analyses reviewed, which may have been caused by publication bias.

Another review of meta-analyses was conducted by Engels, Schmid, Terrin, Olkin, and Lau (2000). These authors revised 125 published meta-analyses in the field of clinical medicine. They compared the performance of two effect size indices, the odds ratio and risk difference, usually applied in studies with binary outcomes. Both indices yielded the same conclusion when testing the statistical significance of the mean effect size within the same meta-analysis. However, risk differences led to greater heterogeneity than did odds ratios.

Schmidt, Oh, and Hayes (2009) selected 68 meta-analyses in which a fixed-effect model was assumed, and they reanalyzed the findings while applying the more realistic random-effects model. These meta-analyses focused on gender differences and the relations between personality and aggressive behaviour. The fixed-effect confidence intervals around mean effect sizes showed an overstated and unrealistic precision, as compared to the wider random-effects confidence intervals.

Finally, Lipsey and Wilson (1993) reported an extensive review of meta-analyses of the efficacy of psychological and educational treatments. Some of the analyzed characteristics were the magnitude of the effects, the sample sizes of the primary studies, and the methodological quality of the meta- analyses. The main purpose of this study was to show the ability of meta-analysis to rigorously assess the degree of effectiveness of the treatments.

The present study focused on the methodological characteristics of meta-analyses of the effectiveness of treatments in the field of clinical psychology, with the standardized mean difference as the effect size index. Some of these methodological characteristics were the type of standardized mean difference (between groups or within groups), the distribution of the numbers of studies of the meta-analyses, the distribution of the sample sizes in the studies of each meta-analysis, the distribution of the effect sizes in each of the meta-analyses, the distribution of the between-studies variance values, and the Pearson correlations between the effect size and sample size in each meta-analysis.

With this methodological review of meta-analyses, we intend to offer a guide for the design of future research studies on the performance of meta-analytic procedures (e.g., Monte Carlo or theoretical studies), based on the manipulation of realistic assumptions and parameters in the meta-analyses. Furthermore, the analysis of the distribution of the average effect sizes through the meta-analyses will provide a guide for the interpretation of the clinical significance of the different types of standardized mean differences, in the field of the effectiveness of the clinical psychological treatments. In addition, our results will offer realistic estimates of effect size in this context, which is valuable information for researchers aiming to determine the optimal sample size when planning their investigations.

2.1.1. Types of standardized mean differences

If all studies included in the meta-analysis reported a continuous outcome in the same metric, raw mean differences could be used as the effect size index. However, this is seldom the case in the behavioral and social sciences, where different instruments to measure the same construct are usually considered across studies. This is why standardized mean differences are widely used in meta-analyses conducted in these fields.

Different types of standardized mean differences suit different study designs. In a two-group design (usually experimental vs. control) with a continuous outcome, the most usual formula to estimate the population effect size is (Hedges & Olkin, 1985):

$$d = \left[1 - \frac{3}{4(n_1 + n_2) - 9}\right] \frac{\bar{y}_1 - \bar{y}_2}{\hat{S}} \tag{2.1}$$

where d is an approximately unbiased estimator of the corresponding parameter, $\bar{y}_1$ and $\bar{y}_2$ are the means of the two groups in the outcome, n_1 and n_2 are the sample sizes, and $\hat{S}$ is an estimator of the pooled within-group standard deviation given by:

$$\hat{S} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}, \tag{2.2}$$

S_1^2 and S_2^2 being the unbiased variances of the two groups.

Hedges and OIkin (1985) also derived the formula of the variance of the d index, $\hat{\sigma}_d^2$:

$$\hat{\sigma}_d^2 = \frac{n_1 + n_2}{n_1 . n_2} + \frac{d^2}{2(n_1 + n_2)}. \tag{2.3}$$

In a repeated measures design, where continuous pretest and posttest measures are registered for a sample of subjects (e.g. before and after the intervention), Becker (1988) proposed the standardized mean change, based on the difference between the pretest and posttest means divided by a standard deviation. Depending on the value of the estimated standard deviation in the denominator, there are two proposed d indices that we will denote by d_{c1} and d_{c2}, respectively.

In a sample with *n* subjects, $\bar{y}_{pre}$ and $\bar{y}_{pos}$ being the means in the pretest and posttest, respectively, d_{c1} is defined by (Gibbons, Hedeker, & Davis, 1993):

$$d_{c1} = \left[1 - \frac{3}{4(n-1)-1}\right] \frac{\bar{y}_{pre} - \bar{y}_{pos}}{S_c}, \tag{2.4}$$

where d_{c1} is an approximately unbiased estimator of the corresponding parameter, and S_c is the standard deviation of the change scores from pretest to posttest. The variance of d_{c1} is given by (Morris & DeShon, 2002):

$$\hat{\sigma}^2_{d_{c1}} = \left[1 - \frac{3}{4(n-1)-1}\right]^2 \left(\frac{1}{n}\right)\left(\frac{n-1}{n-3}\right)(1 + n d_{c1}^2) - d_{c1}^2 . \tag{2.5}$$

The d_{c2} index is given by (Becker, 1988; Morris, 2000; Morris & DeShon, 2002):

$$d_{c2} = \left[1 - \frac{3}{4(n-1)-1}\right] \frac{\bar{y}_{pre} - \bar{y}_{pos}}{S_{pre}}, \tag{2.6}$$

where S_{pre} is the standard deviation of the pretest scores, that is not influenced by the effects of the intervention. Morris (2000) derived the formula for estimating the variance of d_{c2}:

$$\hat{\sigma}^2_{d_{c2}} = \left[1 - \frac{3}{4(n-1)-1}\right]^2 \left[\frac{2(1-r)}{n}\right]\left(\frac{n-1}{n-3}\right)\left[1 + \frac{n\, d_{c2}^2}{2(1-r)}\right] - d_{c2}^2 , \tag{2.7}$$

where *r* is the Pearson correlation between the pretest and posttest scores.

Note that, in studies with a two-independent-group design with continuous pretest and posttest measures, the most widely used effect size index is the standardized mean difference, *d*, as defined in Eq. 1, computed on the posttest scores. However, this index is only appropriate when there is random assignment of the subjects to the groups and equivalent pretest scores in both groups can be assumed. Furthermore, a disadvantage of computing the *d* index only on the posttest scores is that the valuable information of the pretest scores is ignored.

Becker (1988), Morris and DeSohn (2002), and Morris (2008) proposed three effect size indices based on the difference between the standardized mean change in the experimental and control groups, that we will denominate d_{g1}, d_{g2}, and d_{g3}. These indices, unlike the standardized mean difference computed only on the posttest scores, take into account the information in both the pretest and posttest scores of the experimental and control groups.

The d_{g1} index is given by:

$$d_{g1} = d_{c1,E} - d_{c1,C}, \tag{2.8}$$

where $d_{c1,E}$ and $d_{c1,C}$ are the standardized mean change defined in Eq. 2.4, for the experimental and control groups, respectively. The variance of the d_{g1} can be estimated by:

$$\hat{\sigma}^2_{d_{g1}} = \hat{\sigma}^2_{d_{c1,E}} + \hat{\sigma}^2_{d_{c1,C}}, \tag{2.9}$$

$\hat{\sigma}^2_{d_{c1,E}}$ and $\hat{\sigma}^2_{d_{c1,C}}$ being the estimated variances of the d_{c1} indices computed by Eq. 2.5 applied on the experimental and control groups, respectively.

An alternative index to d_{g1} is d_{g2}, computed as the difference between the standardized mean change defined in Eq. 2.6 for the experimental and control groups:

$$d_{g2} = d_{c2,E} - d_{c2,C}. \tag{2.10}$$

The estimated variance of the d_{g2} index is given by:

$$\hat{\sigma}^2_{d_{g2}} = \hat{\sigma}^2_{d_{c2,E}} + \hat{\sigma}^2_{d_{c2,C}}, \tag{2.11}$$

where $\hat{\sigma}^2_{d_{c2,E}}$ and $\hat{\sigma}^2_{d_{c2,C}}$ are the estimated variances of the d_{c2} indices computed by Eq. 2.7 for the experimental and control groups, respectively.

Assuming the homogeneity of the pretest standard deviations in the experimental and control groups, the d_{g3} index is given by:

$$d_{g3} = \left[1 - \frac{3}{4(n_E + n_c - 2) - 1}\right]\left[\frac{(\bar{y}_{pre,E} - \bar{y}_{pos,E}) - (\bar{y}_{pre,C} - \bar{y}_{pos,C})}{\bar{S}_{pre}}\right], \tag{2.12}$$

where n_E and n_C are the sample sizes of the experimental and control groups, $\bar{y}_{pre,E}$ and $\bar{y}_{pos,E}$ are the means of the experimental group in the pretest and posttest, $\bar{y}_{pre,C}$ and $\bar{y}_{pos,C}$ are the means of the control group in the pretest and posttest, and $\bar{S}_{pre}$ is given by:

$$\bar{S}_{pre} = \sqrt{\frac{(n_E - 1)S^2_{pre,E} + (n_C - 1)S^2_{pre,C}}{n_E + n_C - 2}}, \tag{2.13}$$

$S^2_{pre,E}$ and $S^2_{pre,C}$ being the variances of the experimental and control groups in the pretest.

Finally, the estimated variance of the d_{g3} index is given by:

$$\hat{\sigma}^2_{d_{g3}} = 2\left[1 - \frac{3}{4(n_E + n_C - 2) - 1}\right]^2 (1 - r)\left(\frac{n_E + n_C}{n_E n_C}\right)\left(\frac{n_E + n_C - 2}{n_E + n_C - 4}\right)\left[1 + \frac{n_E n_C d^2_{g3}}{2(1 - r)(n_E + n_C)}\right] - d^2_{g3} \tag{2.14}$$

where *r* is the mean of the Pearson correlations between the pretest and posttests scores in the experimental and control groups.

2.2. Methodology

2.2.1. Search procedure and selection criteria of the meta-analyses

The data for the present study were extracted from a sample of 50 published meta-analyses about the effectiveness of psychological treatments and interventions. The meta-analyses were obtained from journals with impact factor located in the first quartile of 2011 Journal Citation Reports in the clinical psychology field (*Clinical Psychology Review*, *Psychological Medicine, Journal of Consulting and Clinical Psychology*, *Depression and Anxiety*, *Health Psychology*, *Neuropsychology*, *Behaviour Research and Therapy*, and *Journal of Substance Abuse Treatment*). The search was conducted in Google Scholar and limited to meta-analyses published between 2000 and 2012 with the key words "meta-analysis" OR "systematic review" in the tittle.

First, reading the title and abstract of each reference allowed us to preselect the meta-analyses about the effectiveness of psychological programs, treatments and interventions about psychological, educational and psychosocial disorders. To be included in or study, meta-analyses had to comply with several selection criteria. First, we only included meta-analyses using an effect size index from the *d* family: the posttest standardized mean difference (Eq. 2.1), standardized mean change (Eqs. 2.4 or 2.6), and standardized mean change difference (Eqs. 2.8, 2.10, or 2.12). Furthermore, the meta-analyses should report the individual effect sizes and sample sizes for the primary studies. To ensure that the selected meta-analyses had sufficient data to provide valid results, they had to include seven or more studies, with sample sizes of at least five subjects per group.

A total of 206 published meta-analyses were revised of which 50 were finally included in the study. These included studies are marked with an asterisk in the references section. Some meta-analyses used two different effect sizes of the *d* family (Hesser, Weise, Westin, & Andersson, 2011; Nestoriuc, Rief, & Martin, 2008; Sockol, Epperson, & Barber, 2011; Virués-Ortega, 2010). In those cases, our decision was to consider them as independent meta-analyses. Thus, a total of 54 independent meta-analyses, or analysis

units, took part in the present study. These meta-analyses summarized the results of 1,285 individual studies.

2.2.2. Data extraction

A database was created in SPSS, in which the effects sizes and sample sizes of the individual studies were coded for each meta-analysis. For meta-analyses including several outcomes, we selected the most relevant clinical outcome taking into account the principal aim of the meta-analysis. The type of design in which the computation of the effect size was based, and the type of *d* index were also recorded. Designs were classified as between-groups and within-groups, and type of *d* was coded as posttest standardized mean difference (*d* in Eq. 2.1), standardized mean change (d_{c1} or d_{c2}, in Eqs. 2.4 or 2.6, respectively), and standardized mean change difference (d_{g1}, d_{g2}, or d_{g3}, in Eqs. 2.8, 2.10 or 2.12, respectively). For each *d* value, its variance was estimated with Eqs. 2.3, 2.5, 2.7, 2.9, 2.11, or 2.14, depending on the type of *d*.

The data from each meta-analysis were coded independently by two trained coders, with agreement percentages ranging between 94.44% and 100%. Inconsistencies between the coders were solved by consensus.

2.2.3. Meta-analytic calculations

Several computations were carried out using each meta-analytic database. The weighted average effect size was estimated using the following expression:

$$\overline{T} = \frac{\sum_i \hat{w}_i T_i}{\sum_i \hat{w}_i}, \qquad (2.15)$$

where T_i refers to any *d* family effect size index, and $\hat{w}_i$ is the estimated weighting factor computed through $\hat{w}_i = 1/(\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2)$. The within-study variance of each individual study, $\hat{\sigma}_i^2$, was estimated using the formula corresponding to the type of *d* index (see Eqs. 2.3, 2.5, 2.7, 2.9, 2.11, and 2.14). The between-studies variance, $\hat{\tau}_{DL}^2$, was calculated through

the procedure of DerSimonian and Laird (1986), the most commonly used in practice. In this procedure, the between-studies variance estimator is derived from the moment method

$$\hat{\tau}^2_{DL} = \frac{Q-(k-1)}{c}, \tag{2.16}$$

where k is the number of studies of the meta-analysis, and Q is a statistic to test the heterogeneity of the effect sizes, described by Cochran (1954), and obtained by

$$Q = \sum_i \hat{w}_i^* \left(T_i - \overline{T}^*\right), \tag{2.17}$$

with $\hat{w}_i^*$ being the estimated weights assuming a fixed-effect model, $\hat{w}_i^* = 1/\hat{\sigma}_i^2$; $\overline{T}^*$ being the mean effect size also assuming a fixed-effect model-that is, applying Eq. 2.15, but using $\hat{w}_i^*$ as weighting factor; and c being given by

$$c = \sum_i \hat{w}_i^* - \frac{\sum_i \left(\hat{w}_i^*\right)^2}{\sum_i \hat{w}_i^*}. \tag{2.18}$$

The mean effect size (Eq. 2.15) was always computed with DL estimator. Restricted Maximum Likelihood (REML) and Paule and Mandel (PM) estimators of $\hat{\tau}^2$ were also applied in order to know the distributions of the between-studies variances. Next we present formulas for these estimators.

The REML estimator is obtained iteratively from Sánchez-Meca and Marín-Martínez (2008) and Viechtbauer (2005):

$$\hat{\tau}^2_{REML} = \frac{\sum_i (\hat{w}_i)^2 \left[\left(T_i - \overline{T}\right)^2 - \hat{\sigma}_i^2\right]}{\sum_i (\hat{w}_i)^2} + \frac{1}{\sum_i \hat{w}_i}, \tag{2.19}$$

where $\hat{w}_i$ is the estimated weighting factor, T_i refers to any *d* family effect size index, $\hat{\sigma}_i^2$ is the within-study variance of each individual study, and $\overline{T}$ is defined in Eq. 2.15. When $\hat{\tau}^2_{REML} < 0$, it is truncated to zero.

The final estimator was also obtained through an iterative method, proposed by Paule and Mandel (1982). Applying this estimator, the between-studies variance is given by

$$\hat{\tau}^2_{PM} = \sum_i \hat{w}_i \left(T_i - \overline{T}\right) - (k-1) \qquad (2.20)$$

where $\hat{w}_i$ is the estimated weights, T_i is any of *d* family effect size, $\overline{T}$ is defined in Eq. 2.15, and *k* is the number of studies.

To test for true heterogeneity among the population effect sizes, we calculated the *Q*-statistic defined in Eq. 2.17, for each meta-analysis. Under the hypothesis of homogeneity among the effect sizes, the *Q* statistic follows a chi-square distribution with *k* - 1 degrees of freedom.

The *Q*-statistic does not inform researchers of the extent of true heterogeneity, only of its statistical significance. Furthermore, the *Q* test has poor power to detect true heterogeneity among the effect sizes when the meta-analysis includes a small number of studies ($k < 30$, Sánchez-Meca & Marín-Martínez, 1997). To overcome the shortcomings of the *Q* test, Higgins and Thompson (2002; Higgins, Thompson, Deeks, & Altman, 2003) proposed the I^2 index for assessing the magnitude of heterogeneity exhibited by the effect sizes. For each meta-analysis, the I^2 index was computed as

$$I^2 = \frac{Q-(k-1)}{Q} \times 100\% \qquad (2.21)$$

The I^2 index was interpreted as the percentage of the total variability in a set of effect sizes due to true heterogeneity-that is, to between-studies variability. Indicatively,

I^2 rates around 25%, 50% and 75% can be interpreted as reflecting low, medium and high heterogeneity, respectively (Huedo-Medina et al., 2006).

2.2.4. Data analysis

The statistical analyses were carried out in R. Specifically, the meta-analytic calculations were programmed with the *metafor* package (Viechtbauer, 2010), using the individual effect sizes and sample sizes coded for each meta-analysis as inputs. For repeated measures data, the correlation between pre- and post-assessment is required for computation of the variance of d_{c2} (Eq. 2.7), d_{g2} (Eq. 2.11), and d_{g3} (Eq. 2.14) indices. Following Rosenthal (1991), the criterion was set at $r = 0.7$, as a representative value of the expected correlation in this context.

The normality assumption for the effect size distribution in each meta-analysis was assessed with the Shapiro-Wilk test for small samples, and by computing the skewness and kurtosis of the distribution. Furthermore, the median, skewness and kurtosis were also computed for the sample size distribution in each meta-analysis. Descriptive analyses (minimum, maximum, mean, and quartiles) were carried out on the next indices across the meta-analyses: number of studies; mean effect size (Eq. 2.15); *p*-value of the Shapiro-Wilk test; skewness and kurtosis of the *d* values; median, skewness and kurtosis of the sample sizes distribution; Pearson correlation between effect sizes and sample sizes; and the *p*-value of the heterogeneity *Q* statistic (Eq. 17); I^2 index (Eq. 2.21); and $\hat{\tau}^2$ index (Eqs. 2.16, 2.19 and 2.20). These analyses were performed separately for meta-analyses using the posttest standardized mean difference, the standardized mean change, and the standardized mean change difference

The R code is provided in Appendix 2A. The 54 meta-analytic databases are available in the Open Science Framework (https://osf.io/yd52u/).

2.3. Results

2.3.1. Characteristics of the meta-analyses

A total of 54 meta-analyses were included in this study, of which 41 used the posttest standardized mean difference (between-groups design), 11 used the standardized mean change (within-groups design) and 2 used the standardized mean change difference (between-groups design). The database with the 54 meta-analyses is presented in the Appendix 2B. The type of *d* family effect size index, the equation applied to estimate the variance of each individual effect size, and some meta-analytic calculations were registered for each meta-analysis: number of studies, mean effect size, *p*-value associated with the *Q* statistic; I^2 ; and $\hat{\tau}^2_{DL}$, $\hat{\tau}^2_{REML}$ and $\hat{\tau}^2_{PM}$ values. We performed these calculations using the values of the effect sizes and sample sizes from each meta-analysis.

The values of the *d* indices reported by the authors of the meta-analyses were computed as in Eqs. 2.1, 2.4, 2.6, 2.8, 2.10, and 2.12; or with some slight variations of these Equations. Specifically, in some meta-analyses, the d_{c2} index (Eq. 2.6) was computed using pooled standard deviations from pretest and posttest data, instead of the standard deviations in the pretest (meta-analyses in Casement & Swanson, 2012; Driessen et al., 2010; Hansen, Höfling, Kröner-Borowik, Stangier, & Steil, 2013; and Williams, Hadjistavropoulos, & Sharpe, 2006). Also, in the meta-analysis of Aderka, Nickerson, Bøe, and Hofmann (2012), the d_{g3} index (Eqs. 2.12 and 2.13) was computed using the variances of the change scores, instead of the variances in the pretest.

Some meta-analyses included more than one type of *d* index and, consequently, the 50 published meta-analyses were disaggregated in 54 independent meta-analyses. For instance, the meta-analysis in Hesser et al. (2011) was disaggregated in two meta-analyses, since the standardized mean difference was used to compare the treatment and control groups at posttest, and the standardized mean change was used to evaluate the differences from pretest to posttest for some treatment groups (see Appendix 2B).

Next, the distributions of the number of studies, effect sizes, sample sizes in the primary studies, correlations between effect sizes and sample sizes, and heterogeneity indices of the meta-analyses, are presented as a function of the type of *d* index. Descriptive analyses of these distributions are shown for the meta-analyses using the posttest standardized mean difference (see Table 2.1), the standardized mean change (see

Table 2.2) and the standardized mean change difference (see Table 2.3). Figure 2.1 shows the corresponding box plots of the analysed distributions, for the meta-analyses using the posttest standardized mean differences (*d*) and the standardized mean changes (d_c), and Figure 2.2 presents histograms of mean effect sizes and between-studies variances distributions for the meta-analyses using the posttest standardized mean differences (*d*) and the standardized mean changes (d_c). Only two meta-analyses used the standardized mean change difference.

Table 2.1. Descriptive analyses of the meta-analytic calculations for posttest standardized mean difference.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
K	7	14	18	24.2	25	70
$\overline{d}$	0.068	0.249	0.409	0.472	0.695	1.075
*p*_norm	.000	.008	.138	.211	.312	.858
d_skewness	-1.947	0.179	0.571	0.503	0.994	2.354
d_kurtosis	-1.758	-0.839	-0.212	0.414	1.033	6.001
N_median	16	32	46.5	48.6	64	87.5
N_skewness	-1.085	0.914	1.357	1.350	1.762	3.487
N_kurtosis	-1.512	-0.477	0.722	1.749	2.684	14.170
$r_{d,N}$	-.612	-.329	-.212	-.119	.059	.734
*p*_*Q*	.000	.000	.000	.095	0.035	0.981
I^2	0	37.71	59.86	54	74.83	93.61
$\hat{\tau}^2_{DL}$	0.000	0.055	0.111	0.159	0.171	1.024
$\hat{\tau}^2_{REML}$	0.000	0.043	0.108	0.181	0.179	0.816
$\hat{\tau}^2_{PM}$	0.000	0.059	0.129	0.215	0.352	0.789

Note. Min. = minimum; 1st Qu. = First Quartile; 3rd Qu. = Third Quartile; Max. = maximum; *k* = number of studies; $\overline{d}$ = average effect sizes applying DL to estimate the between-studies variance; *p*_norm = *p*-value associated to the Shapiro-Wilk test; d_skewness = skewness of effect sizes; d_kurtosis = kurtosis of effect sizes; N_median = median of sample sizes ; N_skewness = skewness of sample sizes; N_kurtosis = kurtosis of sample sizes; $r_{d,N}$ = correlation between effect sizes and sample sizes; p_*Q* = *p*-value associated to the heterogeneity *Q* statistic; I^2 = index to quantify the amount of heterogeneity (in %); $\hat{\tau}^2_{DL}$ = between-studies variance estimated using

the DerSimonian and Laird (1986) method; $\hat{\tau}^2_{REML}$ = between-studies variance estimated using restricted maximum likelihood; $\hat{\tau}^2_{PM}$ = between-studies variance estimated using Paule and Mandel's (1982) method.

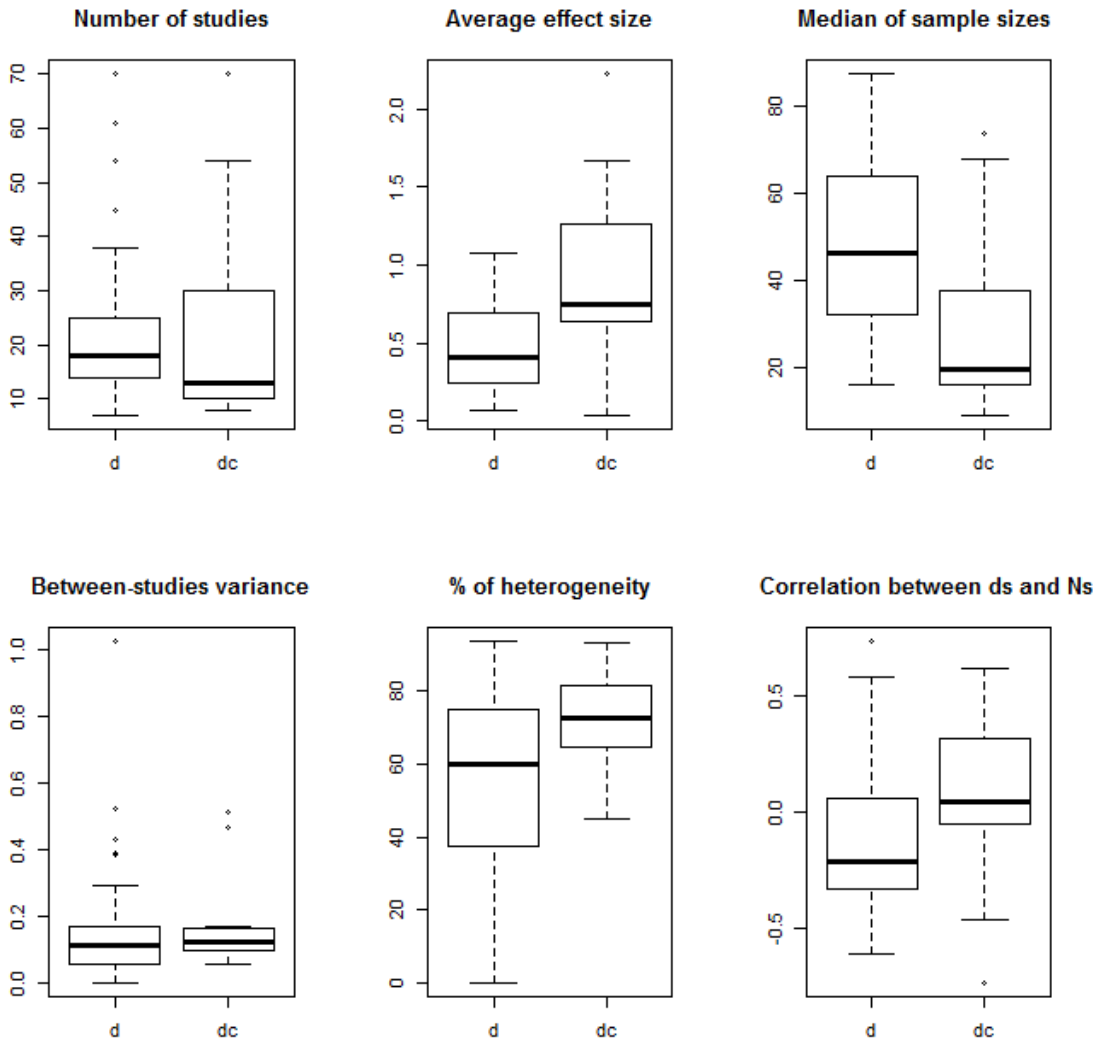


Fig. 2.1. Boxplots of some meta-analytic indices. Between-studies variance was estimated using the DerSimonian and Laird procedure. *d* = posttest standardized mean difference; d_c = standardized mean change

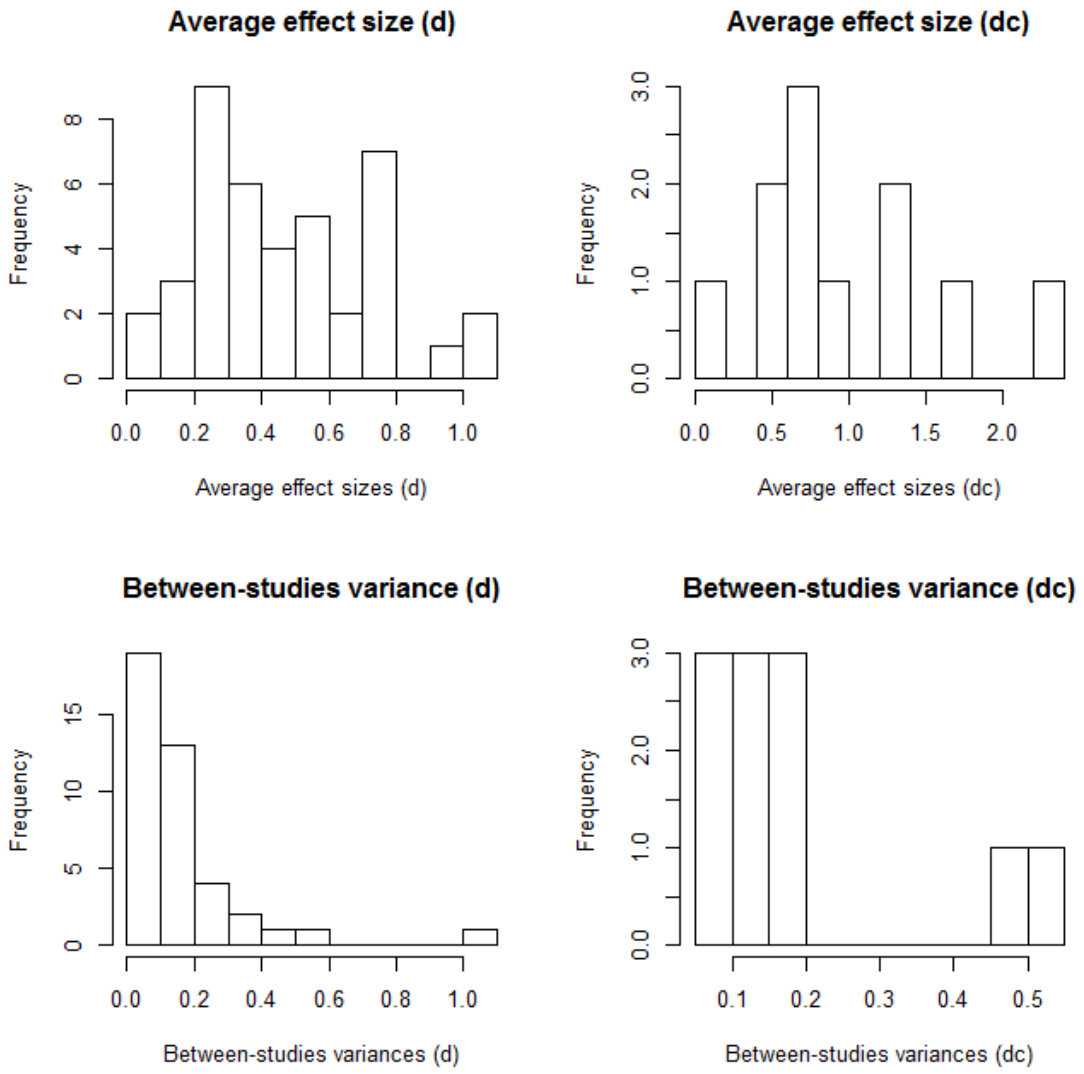


Fig. 2.2. Histograms of the distribution of mean effect sizes and between-studies variances for posttest standardized mean difference (*d*) and standardized mean change (d_c). The between-studies variance was estimated using the DerSimonian and Laird (1986) procedure.

2.3.2. Number of studies

In the 41 meta-analyses that used the posttest standardized mean difference as the effect size index, the number of primary studies ranged from $k = 7$, the minimum number of studies for a meta-analysis to be included in this review, to $k = 70$. The first quartile, median, mean, and third quartile were 14, 18, 24.2, and 25 studies, respectively (see Table

2.1). These results reflect a clear positive skewness, or the predominance of meta-analyses with a small number of studies. Furthermore, as can be seen in Fig 2.1, there were four outliers, namely 45, 54, 61, and 70 studies, resulting in the mean, 24.2, being larger than the median, 18.

The distribution of the number of studies in the standardized mean change meta-analyses was more variable and more skewed than that of the posttest standardized mean difference meta-analyses (see Fig. 2.1). The first quartile, median, mean, and third quartile were 10, 13, 24.09, and 30 studies, respectively (see Table 2.2). These results evidenced a more pronounced positive skewness than in the case of the posttest standardized mean difference meta-analyses. Once again, most meta-analyses included a small number of studies. The number of studies for the two meta-analyses using the standardized mean change difference were 9 and 19, respectively (see Table 2.3).

Table 2.2. Descriptive analyses of the meta-analytic calculations for standardized mean change.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
K	8	10	13	24.09	30	70
$\bar{d}$	0.038	0.640	0.747	0.976	1.258	2.219
*p*_norm	.000	.001	.334	.320	.586	.830
d_skewness	-1.179	-0.114	0.562	0.476	0.951	2.347
d_kurtosis	-1.418	-0.869	-0.483	0.755	1.009	8.559
N_median	9	16	19.5	30.86	37.5	74
N_skewness	0.153	0.683	1.284	1.208	1.695	2.234
N_kurtosis	-1.859	-1.055	-1.088	1.078	2.265	6.149
$r_{d,N}$	-.736	-.054	.045	.060	.318	.622
*p*_*Q*	.000	.000	.000	.002	.000	.013
I^2	44.99	64.86	72.67	72.74	81.61	93.46
$\hat{\tau}^2_{DL}$	0.056	0.099	0.124	0.185	0.163	0.512
$\hat{\tau}^2_{REML}$	0.064	0.105	0.136	0.211	0.219	0.588
$\hat{\tau}^2_{PM}$	0.062	0.088	0.161	0.299	0.341	0.588

Note. Min. = minimum; 1st Qu. = First Quartile; 3rd Qu. = Third Quartile; Max. = maximum; *k* = number of studies; $\overline{d}$ = average effect sizes applying DL to estimate the between-studies variance; *p*_norm = *p*-value associated to the Shapiro-Wilk test; d_skewness = skewness of effect sizes; d_kurtosis = kurtosis of effect sizes; N_median = median of sample sizes ; N_skewness = skewness of sample sizes; N_kurtosis = kurtosis of sample sizes; $r_{d,N}$ = correlation between effect sizes and sample sizes; p_*Q* = *p*-value associated to the heterogeneity *Q* statistic; I^2 = index to quantify the amount of heterogeneity (in %); $\hat{\tau}^2_{DL}$ = between-studies variance estimated using the DerSimonian and Laird (1986) method; $\hat{\tau}^2_{REML}$ = between-studies variance estimated using restricted maximum likelihood; $\hat{\tau}^2_{PM}$ = between-studies variance estimated through Paule and Mandel's (1982) method.

Table 2.3. Meta-analytic calculations for standardized mean change difference

	Meta-analysis 1	Meta-analysis 2
K	9	19
$\overline{d}$	1.307	0.629
*p*_norm	.173	.108
d_skewness	0.383	-0.514
d_kurtosis	-1.358	-1.076
N_median	28	38
N_skewness	1.026	1.745
N_kurtosis	0.006	2.296
$r_{d,N}$	.258	-.496
*p*_*Q*	.001	.000
I^2	69.49	68.45
$\hat{\tau}^2_{DL}$	0.242	0.109
$\hat{\tau}^2_{REML}$	0.213	0.083
$\hat{\tau}^2_{PM}$	0.190	0.066

Note. Min. = minimum; 1st Qu. = First Quartile; 3rd Qu. = Third Quartile; Max. = maximum; *k* = number of studies; $\overline{d}$ = average effect sizes applying DL to estimate the between-studies varian ce; *p*_norm = *p*-value associated to the Shapiro-Wilk test; d_skewness = skewness of effect sizes ; d_kurtosis = kurtosis of effect sizes; N_median = median of sample sizes ; N_skewness = skew ness of sample sizes; N_kurtosis = kurtosis of sample sizes; $r_{d,N}$ = correlation between effect si zes and sample sizes; p_*Q* = *p*-value associated to the heterogeneity *Q* statistic; I^2 = index to qua ntify the amount of heterogeneity (in %); $\hat{\tau}^2_{DL}$ = between-studies variance estimated using the De

rSimonian and Laird (1986) method; $\hat{\tau}^2_{REML}$ = between-studies variance estimated using restricte d maximum likelihood; $\hat{\tau}^2_{PM}$ = between-studies variance estimated through Paule and Mandel's (1982) method.

2.3.3. Effect sizes distribution

The mean effect size, the *p*-value of the Shapiro-Wilk test for normality, and the skewness and kurtosis of the effect sizes were computed for each meta-analysis. To analyse the distribution of the mean effect sizes, these means were taken in absolute value. Note that the sign of a *d* index is arbitrary, since it depends on the order in which the means of the two groups in each primary study are subtracted. Then, our interest was on the magnitude of the mean effect sizes.

In the posttest standardized mean difference meta-analyses, the first quartile, median, mean, and third quartile of the mean effect sizes distribution, were 0.249, 0.409, 0.472, and 0.695, respectively (see Table 2.1). These results are similar to the three values, 0.2, 0.5, and 0.8, reflecting a low, medium, and high magnitude, respectively, according to Cohen (1988).

The shape of the distribution of the posttest standardized mean differences in each meta-analysis was also examined. The Shapiro-Wilk test for normality was statistically significant in 39.02% of the meta-analyses, with .211 as the mean *p*-value associated to this normality test. The skewness of distributions ranged from -1.947 to 2.354, with 0.179 as the first quartile. Kurtosis ranged from -1.758 to 6.001 (see Table 2.1). This means that the effect size distribution was positively skewed in most meta-analyses, with a statistically significant departure from normality in almost 40% of the meta-analyses.

In the meta-analyses using the standardized mean change, the three quartiles, the mean, and the maximum values in the mean effect sizes distribution were larger than those in the standardized mean difference meta-analyses (see Table 2.2 and Figs. 2.1 and 2.2). Specifically, the three quartiles were 0.640, 0.747, and 1.258, the mean was 0.976, and the maximum 2.219, which was treated as an outlier. Note that these results remarkably exceed the 0.2, 0.5, and 0.8 values proposed by Cohen (1988).

Similar to posttest standardized mean differences, the shape of the standardized mean change distributions deviated normality. This deviation was statistically significant in 36.36% of the meta-analyses, according to the Shapiro-Wilk test. The skewness and

kurtosis distributions ranged from negative values in the first quartile to positive ones in the third quartile (see Table 2.2).

In the two meta-analyses using the standardized mean change difference, the mean effect sizes were 1.307 and 0.629, respectively (see Table 2.3). The skewness and kurtosis values were 0.383 and -1.358 for the first meta-analysis, and -0.514 and -1.076 for the second meta-analysis. However, in both meta-analyses, the Shapiro-Wilk test was not statistically significant, the *p*-values being .173 and .108, respectively.

2.3.4. Sample size distribution

We examined the sample size distribution through the *k* primary studies in each meta-analysis, by computing the median sample size, the skewness and kurtosis of the sample sizes. The distribution of these statistics was analyzed across the 41, 11, and 2 meta-analyses with different *d* effect size indices.

In the posttest standardized mean difference meta-analyses, the median sample size ranged from 16 to 87.5, with the mean being 48.6 (see Table 2.1). The first quartile of the skewness values was 0.914, which reflects a positive skewness of the sample size distributions in most meta-analyses (e.g., the primary studies predominantly had small sample sizes). The kurtosis values showed a large dispersion, ranging from -1.512 to 14.170, the first and third quartiles being -0.477 and 2.684, respectively.

The sample sizes in the primary studies of the standardized mean change meta-analyses were lower than those in the posttest standardized mean difference meta-analyses (see Table 2.2 and Fig. 2.1). The median sample size ranged from 9 to 74 (an outlier), with a positively skewed distribution, where the three quartiles and the mean (16, 19.5, 37.5, and 30.86, respectively) were remarkably lower than those in the meta-analyses using the posttest standardized mean difference (32, 46.5, 64, and 48.6, respectively). The skewness values of the sample size distributions were all positive, ranging from 0.153 to 2.234, again suggesting the predominance of small sample sizes. The kurtosis values, ranging from -1.859 to 6.149, once again showed a large variability.

In the two standardized mean change difference meta-analyses, the medians of the sample sizes were 28 and 38, respectively (see Table 2.3). The skewness of the sample

sizes were similar in the two meta-analyses, whereas kurtosis values showed a higher discrepancy.

2.3.5. Correlation between effect sizes and sample sizes

Regarding meta-analyses using the posttest standardized mean difference, the correlations between effect sizes and sample sizes ranged from -.612 to .734. Most correlations (70.73%) were negative, with -.119 as the mean value (see Table 2.1). Out of the total of correlations, 14.63% were statistically significant (three positive and three negative).

A wide range of correlations, from -.736 to .622, was also found in the standardized mean change meta-analyses (see Table 2.2 and Fig. 2.1). However, in this case most correlations (72.73%) were positive and the mean of the correlations was also positive (.060). Out of the total of correlations, 27.27% were statistically significant (2 positive and 1 negative).

As shown in Table 2.3, in the first meta-analysis the correlation between the standardized mean change differences and sample sizes was positive and not statistically significant (r = .258). In contrast, in the other meta-analysis the correlation was negative and statistically significant (r = -.496).

2.3.6. Heterogeneity

Three meta-analytic indices were used in order to study the heterogeneity of the effect sizes in the included meta-analyses: the Q statistic (Eq. 2.17), the I^2 index (Eq. 2.21), and the between-studies variance, $\hat{\tau}^2$, estimated using the DL, REML and PM procedures (Eqs. 2.16, 2.19 and 2.20, respectively). Because the results for the three estimators of the between-studies variance were very similar, we will only describe the findings relative to DL estimator, $\hat{\tau}^2_{DL}$.

In the meta-analyses using the posttest standardized mean difference, the third quartile of the distribution of p-values associated to the Q statistics was .035, below the .05 significance level (see Table 2.1). In particular, 75.6 % of the Q tests were statistically significant at the .05 level (see Appendix 2B).

As also shown in Table 2.1, the three quartiles of the I^2 distribution were 37.71%, 59.86%, and 74.83%. These results are relatively close to the three values, 25%, 50%, and 75% proposed by Higgins and Thompson (2002) as reflecting low, medium and high heterogeneity, respectively. Furthermore, 87.80% of the I^2 values were above 25%, that is to say, 36 out of the 41 meta-analyses showed a medium or high variability in the effect sizes (see Appendix 2B).

In the same vein, the three quartiles of the $\hat{\tau}^2_{DL}$ values were 0.055, 0.111, and 0.171, respectively, with four outliers in the distribution - namely 1.025, 0.433, 0.522, and 0.384 (see Figs. 2.1 and 2.2).

The heterogeneity *Q* test was statistically significant in all the standardized mean change meta-analyses (see Table 2.2). The I^2 values ranged from 44.99% to 93.46%, and the $\hat{\tau}^2_{DL}$ ones ranged from 0.056 to 0.512. Thus, all the meta-analyses exhibited a medium to large heterogeneity. As shown in Fig. 2.1, the I^2 and $\hat{\tau}^2_{DL}$ values were generally larger in these meta-analyses, as compared to the posttest standardized mean difference meta-analyses.

The two meta-analyses using the standardized mean change difference showed a statistically significant heterogeneity, with large I^2 and $\hat{\tau}^2_{DL}$ values, respectively above 68% and 0.10 (see Table 2.3).

2.4. Discussion

The aim of this study was to analyse the methodological characteristics of 54 meta-analyses of the effectiveness of psychological treatments in the clinical psychology area that used the standardized mean difference as the effect size index. These meta-analyses were extracted from the most high-impact journals in the field of clinical psychology, located in the first quartile of the ranking of the Journal Citation Reports.

The typical design in the primary studies evaluating the effectiveness of an intervention program was the pretest-postest-control group design. Most meta-analyses, 41 out of 54, used the standardized mean difference computed from the posttest scores to compare experimental and control groups (Eq. 2.1). Eleven meta-analyses used the standardized mean change from pretest to posttest only in the treated groups (Eqs. 2.4 and

2.6), usually because some of the primary studies compared different treatments without including control groups. Finally, only two meta-analyses (Aderka et al. 2012 and Virués-Ortega, 2010) used the standardized mean change difference proposed by Morris (2008) (Eq. 2.12), in which the gains from pretest to posttest are compared between the experimental and control groups.

The classic standardized mean difference computed from the posttest scores does not take into account the usual pre-differences between treatment and control groups, which also can occur in randomized studies. This poses a threat to the internal validity of the results. The standardized mean change from pretest to posttest in a treatment group can be affected by maturation, history or testing effects, which also represent a threat to the internal validity (Shadish, Cook, & Campbell, 2002). These limitations in both indices are partly overcome by using the standardized mean change difference between the experimental and control groups (Morris, 2008). Then, although in practice the standardized mean change difference is scarcely used (only in two out of our 54 meta-analyses), it should be considered in future meta-analyses.

The standardized mean change difference gives very similar results to those of the standardized mean difference computed from the change scores from pretest to posttest. This is especially true when the pretest scores are similar in the experimental and control groups. As a consequence, the methodological characteristics of our 41 meta-analyses using the posttest standardized mean difference can also serve to guide the design of the future research about the meta-analyses using the standardized mean change difference, as well as the correct interpretation of these meta-analyses.

In the global analysis of the 54 meta-analyses, many of them presented a relatively low number of studies (below 20) and a substantial heterogeneity in the effect sizes, with I^2 values generally larger than 50%, and $\hat{\tau}^2$ values larger than 0.10. Thus, the performance of the meta-analytic statistical methods under these conditions should be a research topic of interest. It is widely known that the Q statistic for heterogeneity is underpowered in meta-analyses with a low number of studies (Sánchez-Meca & Marín-Martínez, 1997). However, in our review, 44 out of the 54 meta-analyses showed a statistically significant heterogeneity in the effect sizes ($p < .05$). This is because of the large I^2 and $\hat{\tau}^2$ values found in most meta-analyses, with only two meta-analyses showing an $I^2 = 0\%$ and $\hat{\tau}^2 = 0$. These findings are in line with other studies supporting the random-

effects model as a more realistic option than the fixed-effect model, on the basis that there is substantial variability in the effect sizes of a meta-analysis (Hedges & Vevea, 1998; National Research Council, 1992; Raudenbush, 1994, 2009).

Cohen (1988) proposed a guide to interpret the magnitude of the standardized mean difference in the social sciences, where values around 0.2, 0.5, and 0.8, represent a low, medium and high magnitude, respectively. This guide should be adapted to the specific field of study, taking into account the typical distribution of effect sizes in the corresponding context (Ferguson, 2009; Hill, Bloom, Black, & Lipsey, 2008; Valentine & Cooper, 2003). In this vein, our study contributes to provide a tentative classification of the effect size magnitude of clinical psychology treatments, through the analysis of the distribution of mean effect sizes from our meta-analyses. Correct interpretation of the effect sizes in the empirical research makes it possible to determine the practical/clinical significance of the results, as a complement of the statistical significance (Kirk, 1996). Furthermore, the researcher can decide on the minimum effect size of interest to a priori determine the sample size of an empirical study, with the desired statistical power (Cohen, 1988).

The three quartiles of the mean effect size distribution were 0.249, 0.409, and 0.695, for the meta-analyses using the standardized mean difference computed from the posttest scores. These values, similar to those in Cohen (1988), could be interpreted as a low, medium, and high magnitude, respectively, in the clinical psychology context. To be more specific, for example, a value of $d = 0.80$ could be interpreted as a high magnitude above the 75^{th} perventile in the distribution of average effect sizes in the clinical psychology area. An important point is that this classification can only be applied to the posttest standardized mean differences, and the standardized mean change differences, but not to the standardized mean changes from pretest to posttest.

Meta-analyses using the standardized mean change as an effect size index are more common than researchers would expect to find. This is because of the absence of control groups in the empirical research for ethical reasons, or when the studies are confined to compare different active treatments without including a control group. In general terms, the values of the standardized mean change are larger than those of the posttest standardized mean difference. According to our review, the three quartiles 0.64, 0.747, and 1.258 could be interpreted as a low, medium, and high magnitude. This

classification should be used instead of Cohen's proposal, for the interpretation of the standardized mean change values in the clinical psychological context.

The distribution of the effect sizes in the reviewed meta-analyses deviated from the normality assumption in the random-effects model. The skewness and kurtosis values ranged from negative to positive values of a remarkable magnitude, and the Shapiro-Wilk test for normality was statistically significant in almost 40% of the meta-analyses, in spite of the low number of studies in most of them, which reduces the statistical power of the test. These findings suggest the need to examine the robustness of the meta-analytic procedures to the violation of the normality assumption in the distribution of the effect sizes (see, for example, Kontopantelis & Reeves, 2012a, 2012b), as well as the development of new robust meta-analytic procedures.

The Pearson correlation between effect sizes and sample sizes was statistically significant in 10 out of the 54 meta-analyses, with five positive correlations and five negative. Once again, the low number of studies in numerous meta-analyses reduces the statistical power of the *t*-test for the significance of a correlation, thus preventing the recognition of part of the true correlations. The distribution of the Pearson correlations in the meta-analyses, with values of a remarkable magnitude, could reflect publication selection bias or possibly some other moderator confounded with sample size (e.g., implementation quality; Levine et al., 2009). As a consequence, future research about the performance of different meta-analytic procedures should consider scenarios with positive and negative correlation values between effect sizes and sample sizes, similar to those found in our study.

The present review of meta-analyses provides the minimum, maximum, the mean, and three quartiles of the distribution of the different components in a meta-analysis: number of studies; mean effect size; skewness and kurtosis of the effect size distribution; median, skewness and kurtosis of the sample size distribution; Pearson correlation between effect sizes and sample sizes; and I^2 and $\hat{\tau}^2$ heterogeneity indices (see Tables 2.1, 2.2 and 2.3). These specific values are representative of the realistic conditions in a meta-analysis, which should be contemplated in the research about the performance of the meta-analytic procedures (see section on recommendations below).

2.4.1. Limitations of the study

A requirement of the current review was to include only meta-analyses that reported individual effect sizes and sample sizes for the primary studies. That inclusion criterion might lead to exclusion of meta-analyses with a large number of studies, due to journal space limitations. Nonetheless, our review included meta-analyses with a number of studies ranging from seven to 70, which can be regarded as a wide range that realistically covers the size of most meta-analyses conducted in social and behavioural sciences.

Only two meta-analyses out of the 54 in the review used the standardized mean change difference (Eq. 2.12), where the change scores from pretest to posttest between the experimental and control groups were compared. This index, although scarcely used in practice, overcomes some important limitations of the posttest standardized mean difference and the standardized mean change. As a consequence, it is suggested that future reviews include a larger number of meta-analyses using the standardized mean change difference.

This review is limited to meta-analyses about the effectiveness of clinical psychology treatments, using standardized mean differences as the effect index. Future reviews of meta-analyses in other research areas and with other effect size indices will shed light on the realistic meta-analytic conditions and the typical distribution of the effect sizes in those disciplines.

2.4.2. Recommendations overview

Several recommendations can be made for researchers carrying out a meta-analysis, a Monte Carlo or theoretical study about meta-analytic methods, or a primary study. For studies with a pretest-posttest control group design, the best option is to compute the standardized mean change difference in each study with Eq. 2.12. This index, although scarcely used in practice, has the advantage of controlling for pretest differences between groups, as well as maturation, history or testing effects from pretest to posttest. Our study presents three indices of the standardized mean change difference: d_{g1}, d_{g2}, and d_{g3}, (Eqs. 2.8, 2.10, and 2.12, respectively), and the latter has been found to outperform

the other indices in terms of bias, precision and robustness to heterogeneity of variance (Morris, 2008).

The posttest standardized mean difference, *d* (Eq. 2.1), although widely applied in numerous meta-analyses, does not control for baseline differences between groups, which can also occur in randomized studies. However, in meta-analyses including studies with and without pretest, the *d* index is the best option for all studies. This is because different standardized mean differences (e.g. posttest standardized mean differences, standardized mean changes from pretest to posttest, or standardized mean change differences) should not be combined in the same meta-analysis, since they are not directly comparable.

For studies with a pretest-posttest design without a control group, the usual approach is to compute a standardized mean change from pretest to posttest (d_{c1} and d_{c2} indices in Eqs. 2.4 and 2.6, respectively). These indices may be affected by maturation, history or testing effects. However, in meta-analyses in which a sizeable number of studies do not include a control group, due to ethical reasons or that only active treatments are compared, the d_c index could be computed in all studies. In this study we have presented two types of d_c indices that differ in the estimator of the standard deviation in the denominator of their formulas. The d_{c1} index (Eq. 2.4) uses the standard deviation of the change scores from pretest to posttest, whereas the d_{c2} index (Eq. 2.6) uses the standard deviation of the pretest scores. Most primary studies report the standard deviations of the pretest and posttest scores, whereas the standard deviation of the change scores is less frequently reported. Therefore, the computation of d_{c2} index - based on the standard deviation of the pretest scores - will be more feasible in practice and will provide an estimation of the effect size more similar to those in the intergroup designs.

Monte Carlo and theoretical studies with a scope including meta-analytic methods should consider scenarios found in real meta-analyses. Results in Tables 2.1, 2.2 and 2.3 of this study can inform the design of methodological studies in this context. For example, in a Monte Carlo study simulating data from meta-analyses using the posttest standardized mean difference or the standardized mean change difference, the number of studies, the sample size distribution in the primary studies, or the variance in the effect size distribution could be manipulated using the values in Table 2.1. For the number of studies, *k*, five values could be considered: 7, 14, 18, 25, and 70 (minimum, three

quartiles, and maximum). Similarly, the sample size distribution could be manipulated with average values 16, 32, 46, 64, and 87 (minimum, three quartiles, and maximum), skewness of 1.357 (median), and kurtosis of .722 (median). Finally, the variance of the effect size distribution, $\hat{\tau}^2$, could be set to values of 0, 0.055, 0.111, 0.171, and 1.024 (minimum, three quartiles, and maximum). These results may also be useful in a Bayesian framework, since they can define the construction of an empirical prior.

The distribution of average effect sizes throughout the reviewed meta-analyses can help researchers assess the practical significance (e.g. clinical significance) of an effect size in a primary empirical study or a meta-analysis in this context. For example, a value of d = 0.20 for the posttest standardized mean difference could be interpreted as a low magnitude below the 25th percentile (0.249) in the distribution of the average effect sizes in clinical psychology (see Table 2.1). Furthermore, the benchmarks (minimum, Quartiles 1-3, and maximum) can help the researcher decide on the minimum effect size to determine a priori the sample size of an empirical study with the desired statistical power.

2.5. Conclusions

The results of this review of meta-analyses allow proper interpretation of the magnitudes of the different types of standardized mean differences in the specific area of the evaluation of the effectiveness of the clinical psychological treatments. This is valuable information for interpreting the clinical significance of the results in both a primary research study and a meta-analysis, in terms of either the effect sizes of individual studies, and the average effect size, both overall and by subgroups of studies, in a meta-analysis.

Future research on the performance of the meta-analytic procedures should take into account the methodological characteristics of the real meta-analysis in different areas of research. Particularly, in this work we have analysed the number of studies, the sample size distribution in the studies, the effect size distribution, and the Pearson correlation between effect sizes and sample sizes of 54 real meta-analysis in the clinical psychology area. In this vein, Monte Carlo and theoretical studies could use the values reported in our study to simulate realistic scenarios.

Chapter 3

Study 2:

"Estimating an Overall Effect Size in Random-Effects Meta-analysis when the Distribution of Random Effects Departs from Normal"

3.1. Introduction

One of the main goals in a meta-analysis is to compute an overall effect estimate. This study is focused on various methods for computing an estimate of the mean effect size alongside its confidence interval (CI), when some assumptions of the underlying statistical model are not met.

There are two general statistical models for meta-analysis, fixed-effect and random-effects. The choice of model is crucial as it determines the statistical procedures employed to estimate the mean effect and its CI, as well as the generalizability of the meta-analytic results (Borenstein et al., 2009; Hedges & Vevea, 1998; Sánchez-Meca, López-López, & López-Pina 2013).

In this study we focused on the performance of the random-effects model, which allows for a broader generalization of results and conclusions and which is currently applied in most meta-analytic studies (Hedges & Vevea, 1998; Raudenbush, 2009).

3.1.1. The Random-Effects Model

Let *k* denote the number of studies included in a meta-analysis and $\hat{\theta}_i$ the effect size estimated in the *i*th study. The underlying statistical model can be defined as

$$\hat{\theta}_i = \theta_i + e_i, \tag{3.1}$$

where θ_i is the effect parameter for the *i*th study and e_i is the sampling error of $\hat{\theta}_i$. Usually e_i is assumed to be normally distributed, $e_i \sim N(0, \sigma_i^2)$, with σ_i^2 being the within-study variance for the *i*th study.

The random-effects model assumes that the effect parameters, θ_i, are randomly selected from a population of parameters. Thus, θ_i can be defined as

$$\theta_i = \mu_\theta + \varepsilon_i, \tag{3.2}$$

where μ_θ is a parameter representing the grand mean of the effect parameters, and ε_i denotes the difference between the effect parameter of the *i*th study, θ_i, and the grand mean μ_θ. It is assumed that $\varepsilon_i \sim N(0, \tau^2)$, with τ^2 being the between-studies variance. Therefore, combining Eqs. 3.1 and 3.2 enables us to formulate the random-effects model as

$$\hat{\theta}_i = \mu_\theta + e_i + \varepsilon_i, \tag{3.3}$$

where ε_i and e_i are assumed to be independent and, as a result, the effect size estimates $\hat{\theta}_i$ are assumed to be normally distributed with mean μ_θ and variance $\sigma_i^2 + \tau^2$, that is, $\hat{\theta}_i \sim N(\mu_\theta, \sigma_i^2 + \tau^2)$ (Borenstein et al., 2010; Raudenbush, 2009).

Figure 3.1 shows the double random sampling process underlying the standard (e.g. two-level) random-effects model. At the highest level, the *k* studies in the meta-analysis are considered to be a random (or at least representative) sample from a population of studies. Consequently, the effect parameters, θ_1, θ_2, ..., θ_i, ..., θ_k, are regarded as a random sample from a population of effect parameters. This population of parameters is usually assumed to be normally distributed, with mean μ_θ and variance τ^2 At the lowest level, the units (typically subjects) are also assumed to be randomly sampled from the target population in each study. Then, $\hat{\theta}_1$, $\hat{\theta}_2$, ..., $\hat{\theta}_i$, ..., $\hat{\theta}_k$ denote the effect sizes for each primary study, which provide estimates of the effect parameters θ_1, θ_2, ..., θ_i, ..., θ_k, respectively.

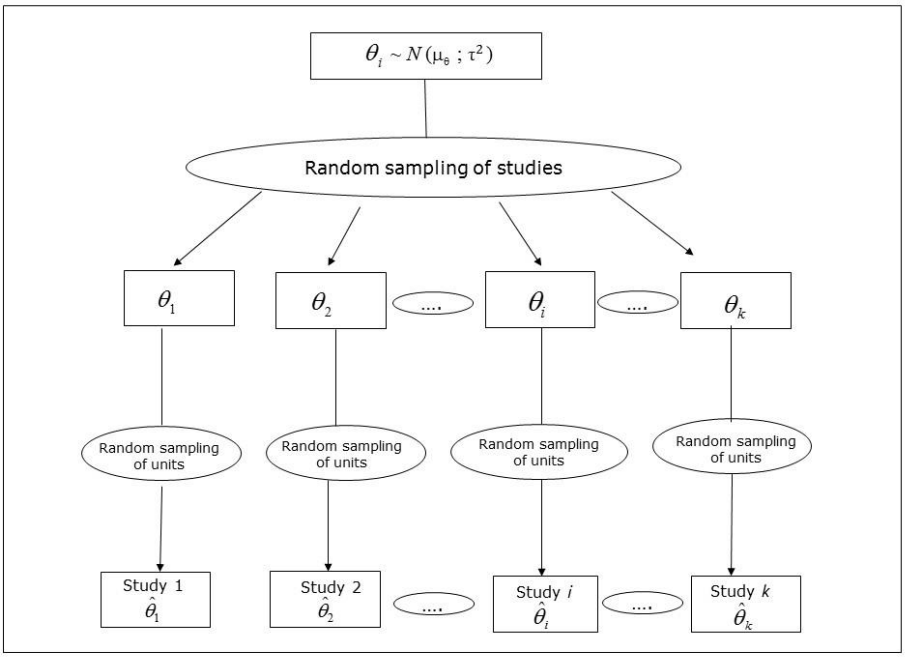


Fig. 3.1. Graphical representation of the random-effects model.

Although normality of the distribution of effect parameters is a common assumption in the random-effects model, it might not be realistic, even approximately, in a wide range of scenarios in practice (Borenstein et al., 2010; Brockwell & Gordon, 2001, 2007; Hardy & Thompson, 1996; Kontopantelis & Reeves, 2012a, 2012b; Schmidt et al., 2009). Departures from normality might affect the accuracy of results such as the

estimation of μ_θ and τ^2. This has important practical implications, since a substantial proportion of the meta-analyses conducted over the last two decades performed random-effects analyses on databases with a small to moderate number of studies. Therefore, assessing the consequences of the violation of the assumption of normality constitutes a relevant question in meta-analysis.

To the best of our knowledge, the works of Kontopantelis and Reeves (2012a, 2012b) are the only simulation studies focused on comparing the performance of several statistical methods for random-effects meta-analysis under non-normal scenarios. Eight statistical methods were examined and a wide range of scenarios were considered. In particular, these authors manipulated the distribution of the effect parameters (normal, skew-normal, and "extremely" non-normal), the number of studies in the meta-analysis and heterogeneity. Most methods were found to be highly robust against violations of the assumption of normality. It must be noted that these previous studies focused on the field of epidemiology, and that the set of simulated scenarios and outcome measures, as well as the effect size index (odds ratios), were selected accordingly following the results of a survey of meta-analyses published in the medical field (Engels et al., 2000).

Furthermore, Kontopantelis and Reeves (2012a, 2012b) generated the individual effect estimates following the method developed for log-odds ratios in Brockwell and Gordon (2001). That approach has two major limitations: the method of Brockwell and Gordon is not realistic, because it does not start from 2x2 tables (Hoaglin, 2015), and it is also not appropriate for other effect metrics.

In the present study, we aimed to assess the consequences of the violation of the assumption of normality in random-effects meta-analyses conducted in the psychological field and, in particular, in meta-analyses about the effectiveness of psychological treatments on various psychological or psychiatric disorders.

To sum up, the purpose of our study was to compare the performance of various random-effects meta-analytic methods for computing an average effect size and a CI around it when the normality assumption is not met. With that purpose, a wide range of scenarios were considered, including conditions with some degree of departure from normality. A Monte Carlo simulation was carried out using standardized mean differences as the effect size index. To avoid the problems in the Kontopantelis and Reeves (2012a, 2012b) studies, in our simulations the standardized mean differences were individually

generated by assuming non-central *t*-test distribution (Hedges & Olkin, 1985). Although our study focused on the random-effects model, the fixed-effect one was also included for comparison purposes.

In the next section, we outline the statistical methods considered in this study and describe the residual heterogeneity variance estimators. A simulation study comparing the performance of the methods is then detailed. Finally, a description of the results is provided and considerations arising from them are discussed.

3.2. Methods to Estimate an Overall Effect Size

3.2.1. The Fixed-Effect Model

The most efficient estimate of the mean effect size under a fixed-effect meta-analysis is given by the expression (Hedges & Olkin, 1985)

$$\hat{\mu}_{UMVU}^{FE} = \frac{\sum_i w_i^{FE} \hat{\theta}_i}{\sum_i w_i^{FE}}, \tag{3.4}$$

with w_i^{FE} being the optimal fixed-effect weights from the *i*th study defined as

$$w_i^{FE} = 1/\sigma_i^2, \tag{3.5}$$

and σ_i^2 the parametric within-study variance of $\hat{\theta}_i$. As the σ_i^2 parameters are unknown, the w_i^{FE} weights are usually estimated by

$$\hat{w}_i^{FE} = 1/\hat{\sigma}_i^2. \tag{3.6}$$

Thus, in practice the common parametric effect size is estimated by

$$\hat{\mu}_{FE} = \frac{\sum_i \hat{w}_i^{FE} \hat{\theta}_i}{\sum_i \hat{w}_i^{FE}}. \tag{3.7}$$

The sampling variance of $\hat{\mu}_{FE}$ is usually estimated by

$$\hat{V}_{FE} = \frac{1}{\sum_i \hat{w}_i^{FE}}. \tag{3.8}$$

Then, a 100(1-α)% CI around the mean effect can be calculated by

$$\hat{\mu}_{FE} \pm z_{1-\alpha/2}\sqrt{\hat{V}_{FE}}\,, \tag{3.9}$$

where $z_{1-\alpha/2}$ is the 100(1-$\alpha/2$) percentile of the standard normal distribution, α being the significance level.

3.2.2. The Random-Effects Model

In a random-effects model, the uniformly minimum variance unbiased estimator of μ_θ is given by (Sánchez-Meca & Marín-Martínez, 2008; Viechtbauer, 2005):

$$\hat{\mu}_{UMVU}^{RE} = \frac{\sum_i w_i^{RE}\hat{\theta}_i}{\sum_i w_i^{RE}} \tag{3.10}$$

with w_i^{RE} being the optimal weights, defined as $w_i^{RE} = 1/\left(\sigma_i^2 + \tau^2\right)$. The variance for $\hat{\mu}_{UMVU}$ is given by: $V_{UMVU} = 1/\sum_i w_i^{RE}$.

However, σ_i^2 and τ^2 are unknown in practice, hence they must be estimated from the studies. The grand mean, μ_θ, can be estimated with

$$\hat{\mu}_{RE} = \frac{\sum_i \hat{w}_i^{RE}\hat{\theta}_i}{\sum_i \hat{w}_i^{RE}} \tag{3.11}$$

where $\hat{w}_i^{RE}$ is an estimate of the random-effects weight for the *i*th study computed with

$$\hat{w}_i^{RE} = 1/(\hat{\sigma}_i^2 + \hat{\tau}^2)\,, \tag{3.12}$$

where $\hat{\sigma}_i^2$ is the estimated within-study variance of $\hat{\theta}_i$ and $\hat{\tau}^2$ is an estimate of the between-studies variance. Several estimators of the between-studies variance are described in a further section.

In the present study, we compare four alternative random-effects methods to construct a CI around the mean effect size: the standard method, Hartung's method, the profile likelihood (PL) method, and non-parametric bootstrapping.

Standard method. The most frequently used method to obtain a CI around the mean effect size estimate, $\hat{\mu}_{RE}$, in a random-effects meta-analysis assumes a normal distribution for $\hat{\mu}_{RE}$, and its sampling variance is usually estimated by

$$\hat{V}_{RE} = \frac{1}{\sum_i \hat{w}_i^{RE}}. \tag{3.13}$$

Therefore, a 100(1-α)% CI around the mean effect size can be computed as

$$\hat{\mu}_{RE} \pm z_{1-\alpha/2}\sqrt{\hat{V}_{RE}}, \tag{3.14}$$

with $z_{1-\alpha/2}$ being the 100(1-$\alpha/2$) percentile of standard normal distribution and 1 - α being the nominal confidence level.

Hartung´s method. Although the standard method is the usual procedure for calculating a CI around the mean effect size, this method assumes a normal distribution and does not take into account the uncertainty derived from the estimation process of the variance parameters. As a consequence, the *z* distribution-based CI has been shown to have empirical coverage below the nominal level, resulting in confidence intervals that are too narrow, especially as the between-studies variance increases and the number of studies decreases (Brockwell & Gordon, 2001). To solve that limitation, Hartung (1999) proposed assuming a *t* distribution, instead of the standard normal distribution, and using an improved variance estimator (see also Hartung & Knapp, 2001; Sidik & Jonkman, 2002). Thus, a 100(1-α)% CI is provided by the expression

$$\hat{\mu}_{RE} \pm t_{k-1;1-\alpha/2}\sqrt{\hat{V}_{HA}} \tag{3.15}$$

where $t_{k-1;1-\alpha/2}$ is the 100(1-$\alpha/2$) percentile of the *t* distribution with $k - 1$ degrees of freedom, $\hat{\mu}_{RE}$ is computed by Eq. 3.11 and $\hat{V}_{HA}$ is an estimate of the sampling variance of $\hat{\mu}_{RE}$ with a weighted extension of the usual formula

$$\hat{V}_{HA} = \frac{\sum_i \hat{w}_i^{RE}\left(\hat{\theta}_i - \hat{\mu}_{RE}\right)^2}{(k-1)\sum_i \hat{w}_i^{RE}}. \tag{3.16}$$

Compared to the standard random-effects method, Hartung's method has been found to yield wider CIs with better coverage probabilities, especially under suboptimal

scenarios (IntHout et al., 2014; Sánchez-Meca & Marín-Martínez, 2008), including scenarios with violation of the normality assumption (Kontopantelis & Reeves, 2012b).

Profile likelihood (PL) method. The profile likelihood (PL) is an iterative and computationally intensive method which can be used to obtain a likelihood-based CI around an overall estimate obtained with the random-effects model, taking into account the fact that μ_θ and τ^2 need to be estimated simultaneously (Hardy & Thompson, 1996). Conversely, the PL method provides two alternatives to calculate a CI around $\hat{\mu}_{RE}$: first-order likelihood method and higher-order Skovgaard´s method. In a simulation study, Guolo (2012) suggested that Skovgaard's method provides far more accurate results than first-order, especially with small sample sizes. More details for the coding scheme of this method are provided in Appendix 3A.

It is expected that likelihood approaches may improve the performance of standard random-effects methods in non-normal scenarios (Guolo, 2012; Hardy & Thompson, 1996; Henmi & Copas, 2010). While standard methods unrealistically assume that between-studies variance is known, the likelihood approach allows deriving likelihood-based confidence intervals for the between-studies variance and for the overall effect. The iterative and joint estimation of both parameters considers the fact that the other parameter is also unknown and must be estimated.

Non-parametric bootstrapping. Resampling methods may be appropiate when the data cannot be regarded as a random sample from a given population. In the context of a meta-analysis, bootstrapping methods are increasingly applied when the assumptions of the random-effects model are not met. This is due to the fact that they are free in theoretical distribution and therefore are expected to be more robust to violations of the normality assumption than standard meta-analytic techniques (Adams, Gurevitch, & Rosenberg, 1997; van den Noortgate & Onghena, 2005). In particular, a non-parametric bootstrapping approach consists of generating a distribution of the mean effect size estimate by resampling a large number of samples, for example, 1,000 samples (Efron, 1987; Efron & Hastie, 2016). Then, an estimate of the parametric mean effect size is obtained averaging the 1,000 effect estimates, and a 95% CI is calculated with the 2.5 and 97.5 percentiles of the effect estimates distribution. We examined two methods for the interval estimation of the mean effect size: the percentile method and the bias-corrected and accelerated (BCa) method. The percentile method involves calculating percentiles

from the bootstrap estimates. However, the BCa method is preferred in practice as it adjusts for both bias and skewness in bootstrap distribution (Efron, 1987, 1992). See Appendix 3A for computational details.

3.2.3. Heterogeneity Variance Estimators

An estimate of τ^2 is required to obtain the mean effect size estimate and its CI under a random-effects model, at least for the standard and Hartung's approaches. Several methods have been proposed to estimate the between-studies variance, τ^2, in random-effects meta-analysis (Sánchez-Meca & Marín-Martínez, 2008; Veroniki et al., 2016; Viechtbauer, 2005). Next, we present formulas for the three estimators considered in this study.

Dersimonian and Laird (DL) Estimator. The most commonly used estimator is that proposed by DerSimonian and Laird (1986), which is derived from the moments method and computed with the expression,

$$\hat{\tau}^2_{DL} = \frac{Q-(k-1)}{c} \tag{3.17}$$

where Q is a heterogeneity statistic computed with

$$Q = \sum_i \hat{w}_i^{FE}\left(\hat{\theta}_i - \hat{\mu}_{FE}\right)^2, \tag{3.18}$$

with $\hat{\mu}_{FE}$ and $\hat{w}_i^{FE}$ already defined in Eqs. 3.7 and 3.6, respectively; whereas c is given by

$$c = \sum_i \hat{w}_i^{FE} - \frac{\sum_i (\hat{w}_i^{FE})^2}{\sum_i \hat{w}_i^{FE}}. \tag{3.19}$$

When $Q < (k-1)$, then $\hat{\tau}^2_{DL}$ is negative and usually truncated to zero to avoid negative values. When the estimated weights $\hat{w}_i^{FE}$ are used instead of the optimal, the Q statistic no longer follows the chi-squared distribution usually assumed, and therefore will negatively affect the performance of $\hat{\tau}^2_{DL}$ as estimator of the heterogeneity variance (Hoaglin, 2016; Kulinskaya, Dollinger, & Bjørkestøl, 2011).

Restricted Maximum Likelihood (REML) Estimator. Another alternative for estimating the heterogeneity variance component is based on restricted maximium likelihood estimation. The REML estimator is obtained iteratively from (Sánchez-Meca & Marín-Martínez, 2008; Viechtbauer, 2005)

$$\hat{\tau}^2_{REML} = \frac{\sum_i \left(\hat{w}_i^{RE}\right)^2 \left[\left(\hat{\theta}_i - \hat{\mu}_{RE}\right)^2 - \hat{\sigma}_i^2\right]}{\sum_i \left(\hat{w}_i^{RE}\right)^2} + \frac{1}{\sum_i \hat{w}_i^{RE}}, \tag{3.20}$$

with $\hat{\mu}_{RE}$ and $\hat{w}_i^{RE}$ defined in Eqs. 3.11 and 3.12, respectively, whereas $\hat{\tau}^2$ is initially estimated by any of the non-iterative estimators of the heterogeneity variance.

When $\hat{\tau}^2_{REML} < 0$, it is truncated to zero.

Empirical Bayes (EB) Estimator. The final estimator of τ^2 that we included is the EB one. It is also an iterative method obtained by replacing $\left(\hat{w}_i^{RE}\right)^2$ with $\hat{w}_i^{RE}$ in Eq. 3.20 for $\hat{\tau}^2_{REML}$ (Berkey, Hoaglin, Mosteller, & Colditz, 1995; Morris, 1983). Thus, the EB estimator is obtained by

$$\hat{\tau}^2_{EB} = \frac{\sum_i \hat{w}_i^{RE} \left[\left(\hat{\theta}_i - \hat{\mu}_{RE}\right)^2 - \hat{\sigma}_i^2\right]}{\sum_i \hat{w}_i^{RE}} + \frac{1}{\sum_i \hat{w}_i^{RE}}. \tag{3.21}$$

Again negative values of $\hat{\tau}^2_{EB}$ are truncated to zero. The EB estimator is equivalent to the Paule-Mandel estimator (Veroniki et al., 2016; Viechtbauer et al., 2015).

3.3. Method of the Simulation Study

In the previous section, we presented three methods for estimating the mean effect size, μ_θ (i.e., fixed-effect model, standard random-effects model, and non-parametric bootstrapping), six methods for computing a CI around an estimate of μ_θ (i.e., fixed-effect model, standard random-effects model, Hartung´s method, profile likelihood method with

higher-order Skovgaard´s approach, and non-parametric bootstrapping with the BCa and the percentile methods), and three estimators of τ^2 (i.e., the DL, REML, and EB estimators) in the context of random-effects meta-analysis. The performance of combinations of these methods was compared using Monte Carlo simulation. However, not all of the methods were combined with each other. In particular, we combined the profile likelihood method with REML estimation and the non-parametric bootstrapping method with the DL estimator, whereas the standard and Hartung's methods were combined with the three τ^2 estimators, and no τ^2 estimators were needed for the fixed-effect model. This yielded five methods to estimate the mean effect size and 10 ways to calculate a CI around that estimate.

The simulation was programmed in R using the *metafor* (Viechtbauer, 2010), *metaLik* (Guolo & Varin, 2012), and *boot* (Canty & Ripley, 2012) packages. Appendix 3A contains the full R code of our simulation study. The standardized mean difference was used as the effect size measure. Designs comparing two groups (experimental and control) with respect to a continuous dependent variable were simulated, a scenario that is often found in psychology. Both populations were assumed to be normally distributed with common variance, $[N(\mu_E, \sigma^2), N(\mu_C, \sigma^2)]$. For each study, the population standardized mean difference, θ, was defined as (Hedges & Olkin, 1985)

$$\theta = \frac{\mu_E - \mu_C}{\sigma}. \qquad (3.22)$$

In a random effects model, a distribution of effect parameters, θ_i, is assumed, with a specific mean, μ_θ, heterogeneity variance, τ^2, and shape (details on how the distributions shapes were defined are provided below). To simulate a meta-analysis, k effect parameters are randomly selected from the distribution of effect parameters, so that there will be an individual parameter θ_i for each study.

The parametric effect size for the ith study, θ_i, is estimated using the nearly unbiased estimator proposed by Hedges and Olkin (1985, p. 81)

$$\hat{\theta} = c(m)g\,, \qquad (3.23)$$

g being a positively biased estimator computed from

$$g = \frac{\bar{y}_E - \bar{y}_C}{S}, \tag{3.24}$$

and c(*m*) a correction factor for small sample sizes, given by

$$c(m) = 1 - \frac{3}{4N - 9} \tag{3.25}$$

where $\bar{y}_E$ and $\bar{y}_C$ are the sample means of experimental and control groups, *S* is a pooled standard deviation computed through

$$S = \sqrt{\frac{(n_E - 1)S_E^2 + (n_C - 1)S_C^2}{n_E + n_C - 2}}, \tag{3.26}$$

n_E and n_C being the experimental and control sample sizes, respectively, S_E^2 and S_C^2 being the unbiased variances of the two groups, and $N = n_E + n_C$.

Eq. 3.23 applies to each study, so that $\hat{\theta}_i$ is an estimate of the effect parameter θ_i. Then, estimates of the sampling variance of $\hat{\theta}$ in each study were obtained by means of

$$\hat{\sigma}_{\hat{\theta}}^2 = \frac{n_E + n_C}{n_E n_C} + \frac{\hat{\theta}^2}{2(n_E + n_C)}. \tag{3.27}$$

Hedges and Olkin (1985, p. 79) showed that $\sqrt{n_E n_C / (n_E + n_C)}g$ follows a noncentral *t*-distribution with noncentrality parameter $\sqrt{n_E n_C / (n_E + n_C)}\theta$ and $n_E + n_C - 2$ degrees of freedom. Then for the *i*th study in a particular meta-analysis estimating the population effect size θ_i, the $\hat{\theta}_i$ value was directly simulated from $Z / \sqrt{X / m}$, where Z is a random normal variable with distributioin N(θ, $1/n_E + 1/n_C$) and X is a random chi-square variable with $m = n_E + n_C - 2$ degrees of freedom.

When calculating $\hat{\mu}_{FE}$ (Eq. 3.7) and $\hat{\mu}_{RE}$ (Eq. 3.11), a potential source of bias is the correlation between the standardized mean difference (Eq. 3.23) and its sampling variance (Eq. 3.27), in particular with small sample sizes.

In order to identify a range of the most realistic scenarios in this field, the manipulated conditions in the present study were set according to the results of a systematic review of 50 meta-analyses on the efficacy of psychological interventions, all using standardized mean differences as the effect-size index (Rubio-Aparicio et al., in press). For the number of studies, *k*, four values were considered, 10, 20, 40, and 60, corresponding to a small to large number of studies for the meta-analysis. The grand mean of the distribution of effect parameters, μ_θ, was set to 0, 0.2, 0.5, and 0.8, which reflect conditions of no effect and effects of low, medium, and large magnitude, respectively (Cohen, 1988). In addition, a wide range of values for the population between-studies variance, τ^2, was considered, 0, 0.03, 0.06, 0.11, 0.18, and 0.39. The simulated conditions for *k*, μ_θ, and τ^2 were within the range of values found in the 50 meta-analyses reported in Rubio-Aparicio et al. (in press).

The shape of the distribution of effect parameters, θ_i, was manipulated through six combinations of skewness and kurtosis values. First, a normal scenario (i.e., zero skewness and kurtosis) was set. To warrant realistic scenarios, five non-normal conditions were then considered based on the results found in Rubio-Aparicio et al. (in press). In that systematic review, the skewness distribution of the 50 meta-analyses presented a median value of 0.52, with 25 and 75 percentiles of 0.18 and 1.1, and minimum and maximum values of -2 and 3.67, respectively. Our purpose was to simulate a wide range of skewness values. Based on these results, skewness values of -2, -1, 0, 1, and 2 were selected to simulate the effect parameters distribution. Then, the nonlinear relationship exhibited by the 50 pairs of skewness and kurtosis found in the systematic review was used to predict kurtosis values. Figure 3.2 presents the scatterplot relating the skewness and kurtosis values of the 50 meta-analyses. A nonlinear predictive model was fitted to this dataset, leading to the predictive equation: Kurtosis = -0.581 + 0.023*Skewness + 1.069*Skewness^2. For the five skewness values previously defined, the five non-normal combinations between skewness and kurtosis values were: (-2, 3.65), (-1, 0.47), (0, -0.58), (1, 0.51), and (2, 3.74). With illustrative purposes Figure 3.3 presents histograms of effect parameters distributions for the six simulated combinations of skewness and kurtosis.

Appendix 3B presents five examples of real meta-analyses selected from Rubio-Aparicio et al. (in press) with similar skewness and kurtosis values to each of the five non-normal scenarios defined in our simulation study. The individual standardized mean differences and sampling variances of each of the five real meta-analyses are available in the Open Science Framework (https://osf.io/z4vsg/).

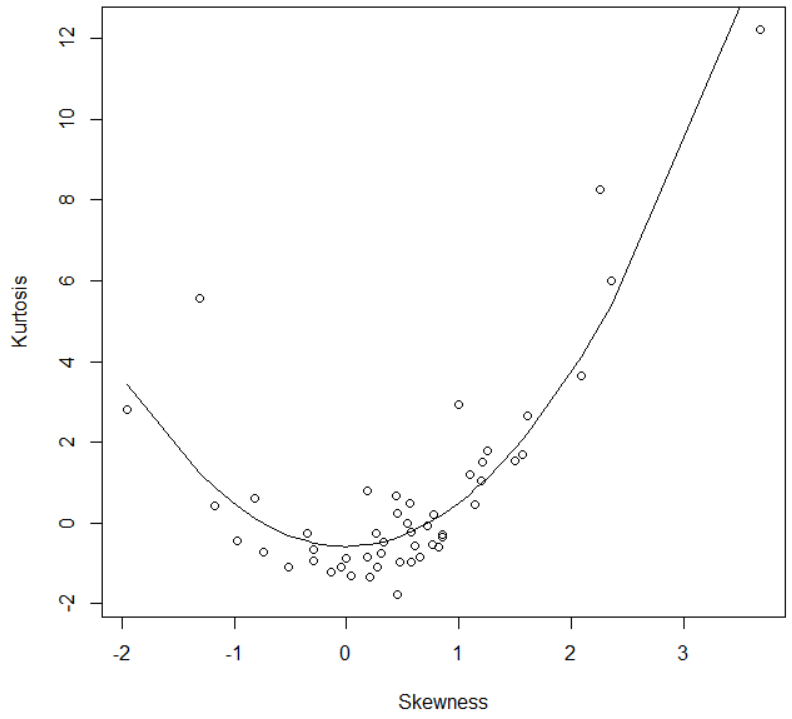


Fig. 3.2. Scatter plot of the skewness and kurtosis values found in a systematic review of 50 meta-analyses of on efficacy of psychological interventions (Rubio-Aparicio et al., in press).

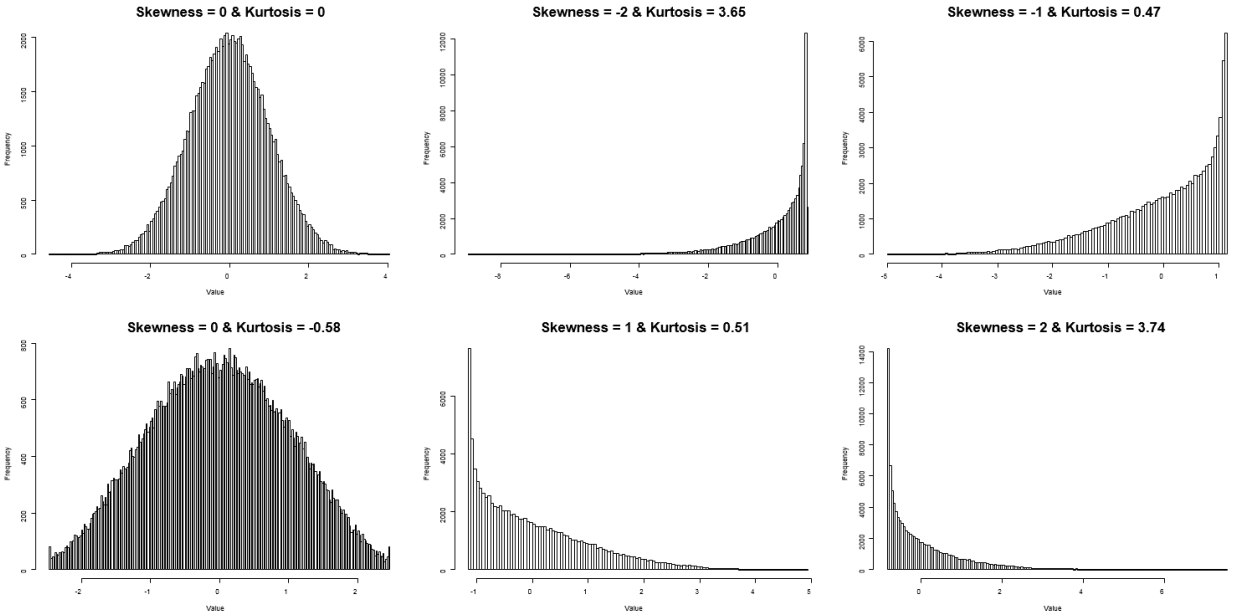


Fig. 3.3. Simulated scenarios for the shape of the distribution of parametric effects, assuming $\mu_\theta = 0$ and $\tau^2 = 1$.

To generate distributions of effect parameters with a given mean (μ_θ), variance (τ^2), skewness, and kurtosis, the Fleishman (1978) algorithm was applied. In particular, the Fleishman power transformation, $X = a + bZ + cZ^2 + dZ^3$, applied on a standard normal distribution, $Z \sim N(0,1)$, allows generating a non-normal random variable X with mean 0, variance 1, skewness γ_1, and kurtosis γ_2 . For a specific combination of γ_1 and γ_2 values, the equations to find the *a*, *b*, *c*, and *d* constants were calculated by solving the equation system presented in Fleishman (1978, p. 522-526). Table 3.1 presents the values of *a*, *b*, *c*, and *d* for the six combinations of γ_1 and γ_2 values in the simulated distributions of effect parameters. The linear transformation $Y = m + nX$ was then applied to generate distributions with the manipulated values of the mean of the effect parameters ($\mu_\theta = 0$, 0.2, 0.5, and 0.8) and the population between-studies variance ($\tau^2 = 0$, 0.03, 0.06, 0.11, 0.18, and 0.39), where $m = \mu_\theta$ and $n = \sqrt{\tau^2}$.

Table 3.1. Values of the *a, b, c, d* constants in the algorithm of Fleishman for the six combinations of skewness and kurtosis.

Skewness (γ_1)	Kurtosis (γ_2)	*a*	*b*	*c*	*d*
0	0	0	1	0	0
-2	3.65	0.349	0.862	-0.349	-0.018
-1	0.47	0.267	1.124	-0.267	-0.071
0	-0.58	0	1.093	0	-0.032
1	0.51	-0.256	1.112	0.256	-0.064
2	3.74	-0.360	0.862	0.360	-0.021

The average total sample size of the individual studies, $\overline{N}$, was manipulated with values 20, 30, 50, and 100. The primary studies were simulated within a two-group design with $n_E = n_C$. To simulate realistic scenarios, the distribution of individual sample sizes was based on the systematic review reported in Rubio-Aparicio et al. (in press) where the sample sizes distributions of the 50 meta-analyses exhibited a clear positive skewness, with average skewness = +1.423. To approach this distribution, a Chi-square distribution with 4 degrees of freedom was used to simulate the sample sizes, as the expected skewness for that distribution is $\sqrt{8/df} = 1.414$, very similar to that empirically obtained. Next, 16, 26, 46, and 96 were added to achieve the desired average values.

When $\tau^2 = 0$, the number of conditions was 64 [4 (*k* values) x 4 (μ_θ values) x 4($\overline{N}$ values)]. Regarding the other values of τ^2, the number of conditions was 1,920 [4 (*k*) x 4 (μ_θ) x 4 ($\overline{N}$) x 6 (shape of the distribution of θ_i values) x 5 (τ^2 values)]. The total number of conditions was then 1,984 and for each one 10,000 meta-analyses were generated. Thus, 19,840,000 meta-analyses were simulated. Furthermore, 1,000 samples per iteration were used for the non-parametric bootstrapping method.

Several criteria were used to compare the performance of the methods for estimating the mean effect and constructing CIs. First, the bias of each of the five methods to estimate the mean effect size was assessed as the diference between the mean of the 10,000

empirical values of each method and condition and the parametric mean effect size for that scenario, μ_θ. Second, variability in the estimates provided by these five methods was assessed by calculating the mean squared error with respect to the true value, μ_θ, across the 10,000 replications of one single condition. Third, the confidence width of the 10 methods to calculate a CI was estimated by averaging the confidence widths across 10,000 replications for each condition. Fourth, the empirical coverage probability for the 95% nominal confidence level of each method was calculated as the percentage of CIs that included the true mean effect size, μ_θ, through the 10,000 replications for each condition. The last criterion was the empirical bias of the estimated standard errors for the standard, Hartung, non-parametric bootstrapping, and fixed-effect methods to the standard deviation of the mean effect estimates distribution. For a particular condition, this criterion was computed as

$$\frac{SD(\hat{\mu}) - Md(SE(\hat{\mu}))}{SD(\hat{\mu})} * 100 \ , \tag{3.28}$$

with $SD(\hat{\mu})$ being the standard deviation of the mean effect estimates obtained in 10,000 replications of a given condition and Md(SE ($\hat{\mu}$)) being the median of the estimated standard errors for the mean effect estimates through the 10,000 replications of the same condition. The reason for using the median instead of the mean was to avoid the potential influence of extreme values. Positive percentages with this formula indicated a negative bias of the estimated standard errors, whereas negative percentages suggested a positive bias.

3.4. Results

For brevity, we only included the data when the grand mean of the distribution of effect parameters was of medium magnitude, 0.5, and the average total sample size was 30, since the pattern of results was very similar for the remaining levels of both factors. Moreover, the chosen value for the between-studies variance was the highest, 0.39, as the differences in the performance of the methods were more pronounced for that value,

although the trends observed in scenarios with less between-studies variation were analogous. The full set of results is available in the Open Science Framework (https://osf.io/z4vsg/).

This section is divided into five parts, corresponding to the comparative criteria: bias and mean squared error of the average effect estimators, empirical coverage probability and width of the CIs, and bias of the estimated standard errors.

3.4.1. Bias of the average effect estimators

Figure 3.4 shows the bias of the standard method with DerSimonian and Laird (DL), restricted maximum likelihood (REML), and empirical Bayes (EB) estimators of τ^2, non-parametric bootstrapping method (BOOT), and fixed-effect method (FE), as a function of the number of studies, *k*, and the shape of the distribution of θ_i .

All methods showed a small negative bias across all simulated scenarios for the shape of the distribution of parametric effects, regardless of the number of studies. The FE yielded the most negatively biased estimates across all conditions, as this model assumes a null between-studies variance ($\tau^2 = 0$).

Under the normality assumption (skewness = 0 & kurtosis = 0), the bias of DL, REML, EB and BOOT was very similar in all conditions for the number of studies. These methods provided the most negatively biased values with $k = 20$. For skewness = 0 and kurtosis = -.58, the performance shown for the five methods was quite similar to the normal condition. When the shape of the distribution of parametric effects was manipulated with skewness = -2 and kurtosis = 3.65, mean effects calculated under a RE model with the DL, REML, EB and BOOT methods were practically unbiased. Similar results were found with skewness = -1 and kurtosis = .47, although under this condition the five methods were more negatively biased. Under conditions with skewness = 1 and kurtosis = .51 and skewness = 2 and kurtosis = 3.74, the differences in bias among the DL, REML, EB and BOOT were practically negligible, with values of bias close to -.025 for all conditions of *k*. The mean effect estimate under the FE model yielded more negatively biased estimates than the rest of the methods.

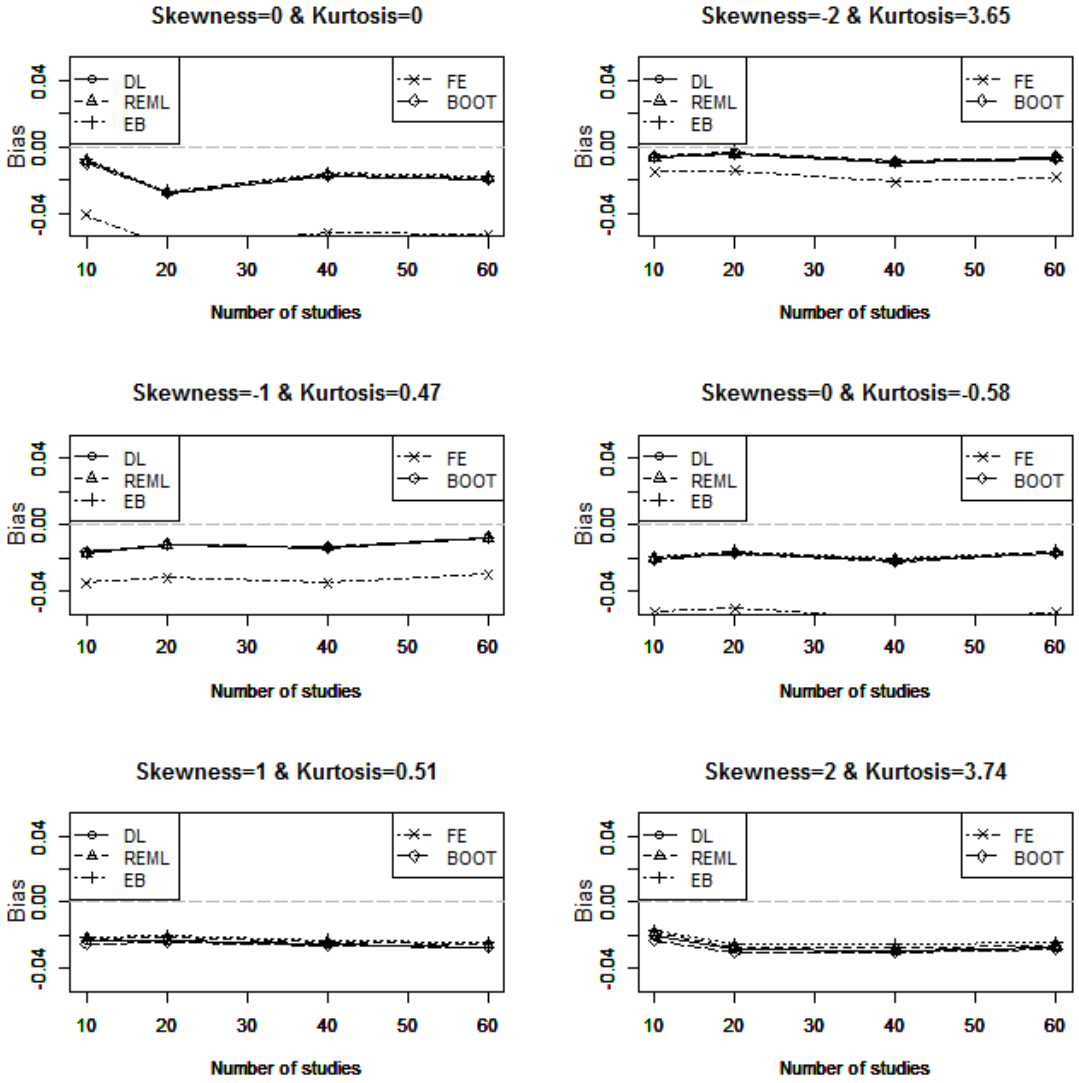


Fig. 3.4. Bias of the five methods to estimate μ_θ. DL = standard method with DerSimonian and Laird estimator; REML = standard method with restricted maximum likelihood estimator; EB = standard method with empirical Bayes estimator; FE = fixed-effect model; BOOT = non-parametric bootstrapping. These results are for: $\tau^2 = 0.39$, $\mu_\theta = 0.5$, and $\overline{N} = 30$. On average the standard error of the simulations was 0.0035

3.4.2. Mean Squared Error of the average effect estimators

Figure 3.5 shows the mean squared error (MSE) of the standard random-effects methods compared. As expected, an increase in the number of studies led to a decrease in the MSE values of the five estimators of μ_θ, regardless of the shape of the distribution of parametric effects. In addition, the results computed through all conditions of skewness

and kurtosis and the number of studies were generally similar in the five methods, without notable differences in their performance. The FE method showed slightly lower MSE values than those of the estimates based on the RE model with a small number of studies (k = 10).

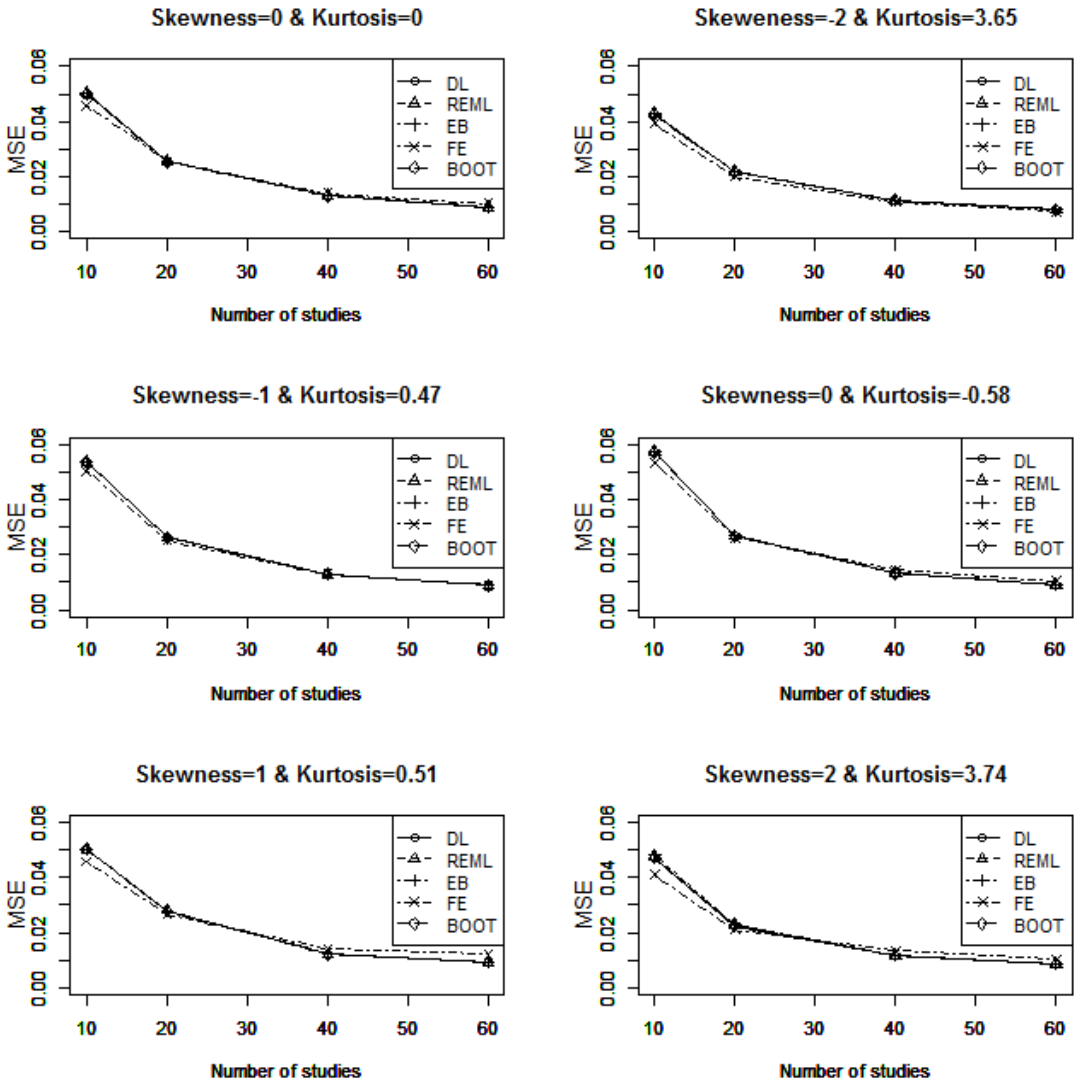


Fig. 3.5. Mean Squared Error (MSE) of the five methods to estimate μ_0. DL = standard method with DerSimonian and Laird estimator; REML = standard method with restricted maximum likelihood estimator; EB = standard method with empirical Bayes estimator; FE = fixed-effect model; BOOT = non-parametric bootstrapping. These results are for:

τ^2 = 0.39, μ_0= 0.5, and $\overline{N}$ = 30. On average the standard error of the simulations was 0.0022

3.4.3. Coverage Probability of the CIs

Figure 3.6 shows the empirical coverage probability of the six CIs compared. The standard and Hartung's methods were not influenced by the heterogeneity estimator used (the DL, REML, and EB estimators). Therefore, only results for the REML estimator are presented.

Most CIs calculated with SM, HM, BOOT_P, BOOT_Bca, and PL methods offered better coverage as the number of studies increased, this improvement being especially evident for *k* = 10 and *k* = 20. Under normality (skewness = 0 & kurtosis = 0), some differences in coverage probabilities among the CIs obtained by SM, HM, BOOT_P, BOOT_Bca, and PL methods were found for small numbers of studies (*k* = 10 and 20). In particular, CIs with HM and PL methods showed the best coverage of the nominal confidence level. For *k* = 10 and *k* = 20, HM method exhibited observed probabilities of .956 and .945, respectively, and PL method obtained .944 and .943. The same trend was found when the parametric effects were non-normally distributed.

The worst coverages of the nominal confidence level were found for skewness = 1 and kurtosis = 0.51, and for skewness = 2 and kurtosis = 3.74. Under these two conditions of shape of the distribution of θ_i, the CIs obtained by all methods generally showed empirical coverage probabilities slightly below the nominal confidence level, even for a large number of studies. Regarding the FE model, its empirical coverage was clearly under the nominal confidence level for all simulated scenarios.

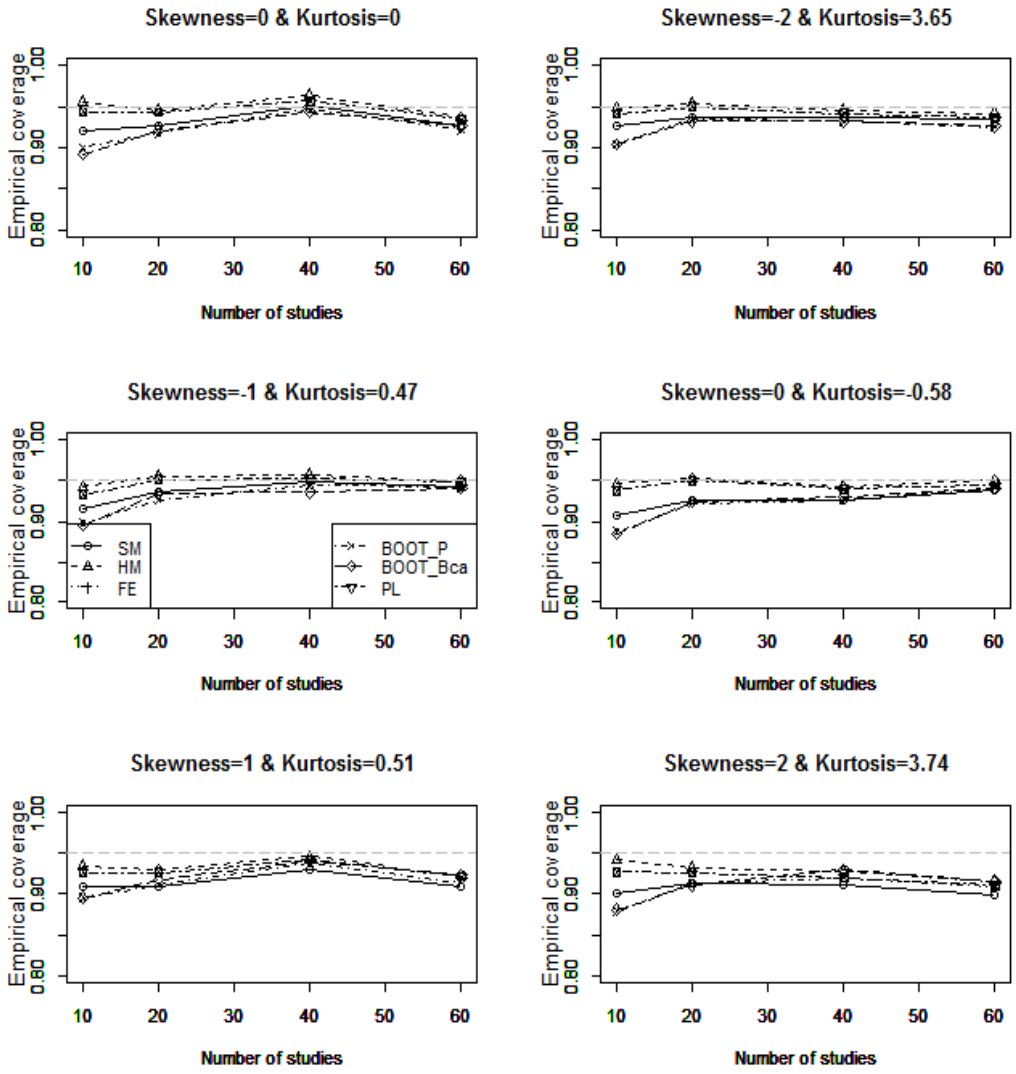


Fig. 3.6. Empirical coverage probability for the six confidence interval (CI) methods. SM = standard method; HM = Hartung's method; FE = fixed-effect model; BOOT_P = non-parametric bootstrapping with the percentile method; BOOT_Bca = non-parametric bootstrapping with the BCa method; PL= profile likelihood method. These results are for: $\tau^2 = 0.39$, $\mu_0 = 0.5$, and $\overline{N} = 30$. On average the standard error of the simulations was 0.0031

3.4.4. Width of the CIs

Figure 3.7 shows the width of the six 95% CIs for μ_θ compared. Only results for the REML estimator are presented. The interval width of the six CI procedures uniformly decreased as the number of studies increased. For k = 10 and 20, the CIs obtained with the HM (especially) and PL methods were wider than those yielded by the other methods. Although this pattern was consistent across all scenarios, the CIs were narrower in conditions with some degree of departure from normality. This was due to a slight undercoverage of the nominal confidence level under scenarios with departures from normality. For instance, with k = 10 and under the normal scenario, the CI widths for HM and PL were 1.004 and .992, with empirical coverage probabilities of .956 and .944, respectively. Under highly non-normal distributions (e.g., skewness = -2 & kurtosis = 3.65), the CI widths for HM and PL were .9456 and .9306, with empirical coverage probabilities of .948 and .941. The FE method consistently yielded the narrowest CIs at the expense of exhibiting a large undercoverage of the nominal confidence level.

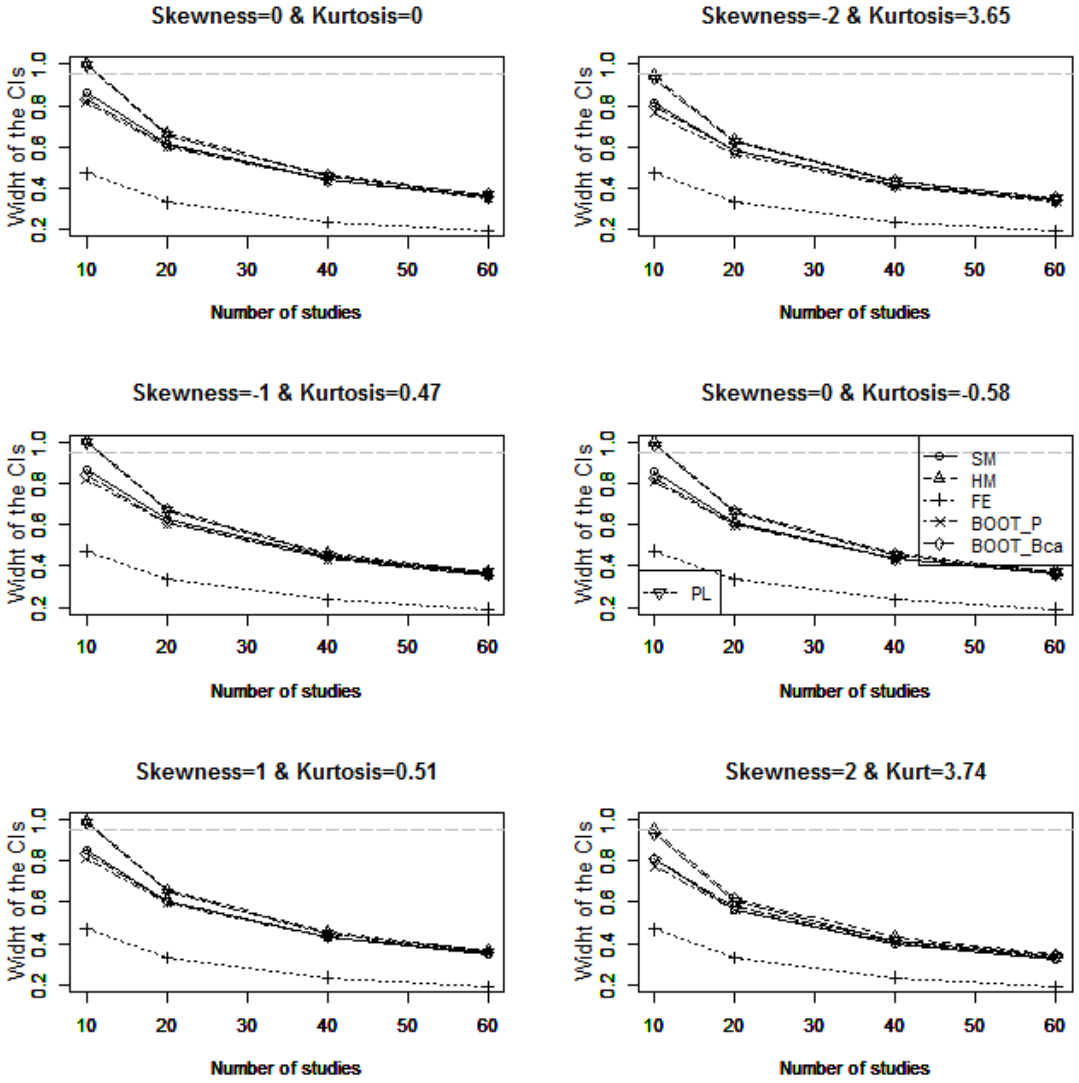


Fig. 3.7. Width of the 95% CI for μ_θ of the six confidence interval (CI) methods. SM = standard method; HM = Hartung's method; FE = fixed-effect model; BOOT_P = non-parametric bootstrapping with the percentile method; BOOT_Bca = non-parametric bootstrapping with the BCa method; PL= profile likelihood method. These results are for: $\tau^2 = 0.39$, $\mu_\theta = 0.5$, and $\overline{N} = 30$. On average the standard error of the simulations was 0.0062

3.4.5. Bias of the Standard Error

Figure 3.8 shows the bias (in %) of the standard error estimates using the REML estimator. On average, all methods yielded estimated standard errors lower than the

standard deviation of the distribution of overall effect estimates empirically constructed through 10,000 replications in a given condition. SM, HM, and BOOT methods exhibited nearly unbiased estimates of the standard error in all manipulated conditions. Moreover, the good performance of the standard error estimates of these methods improved with larger number of studies regardless of shape of the distribution of θ_i. The HM method systematically showed the best performance of the standard error estimates in contrast to the BOOT method, which exhibited the most pronounced negative bias (excluding the FE method). This same trend was found across all conditions of skewness and kurtosis regardless of the number of studies. On average, the negative bias of the standard errors for SM, HM, and BOOT was 3.52%, 1.89%, and 5.16%, respectively. These differences were larger for small *k* values. For instance, for $k = 10$ the average bias of the standard errors of SM, HM, and BOOT through the conditions of skewness and kurtosis was 5.90%, 4.79%, and 10.18%, respectively. The FE method exhibited the largest negative bias, close to 50% across all scenarios of number of studies and shape of distribution.

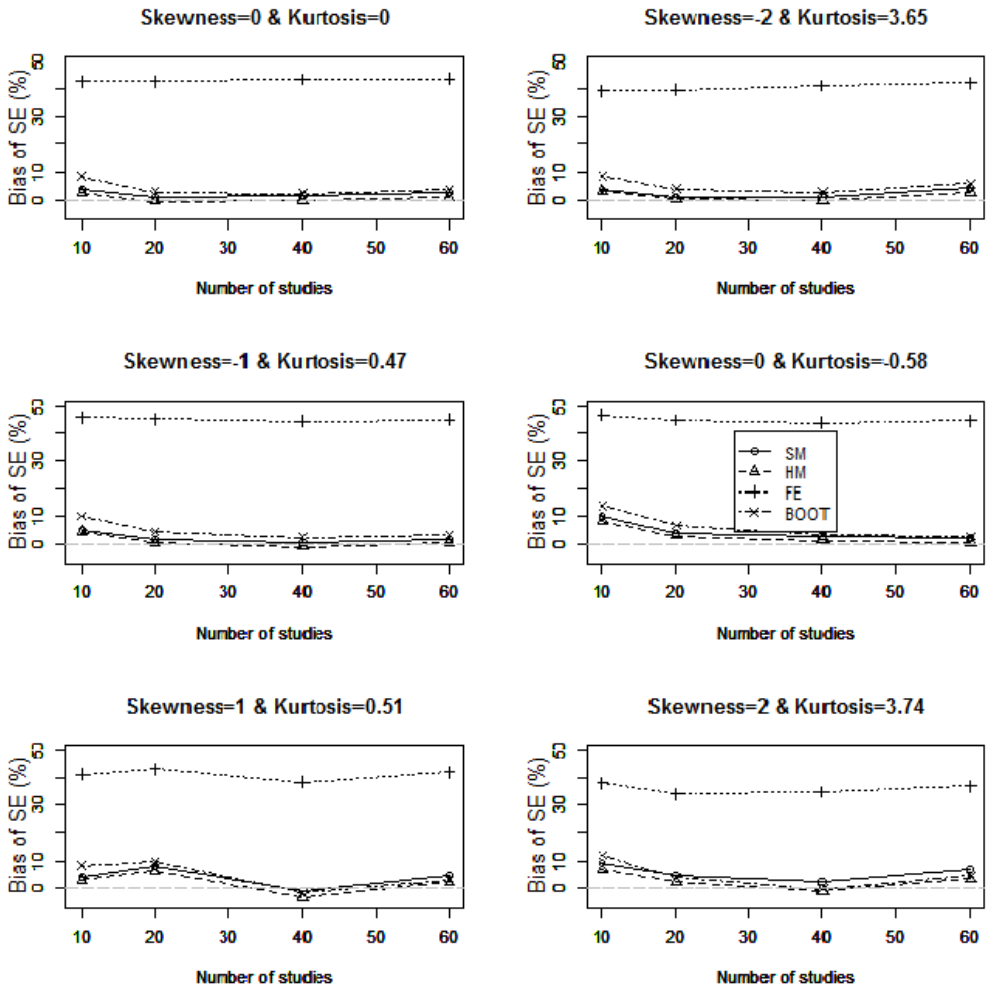


Fig. 3.8. Bias of the Standard Error of the four methods. SM = standard method; HM = Hartung's method; FE = fixed-effect model; BOOT = non-parametric bootstrapping. These results are for: $\tau^2 = 0.39$, $\mu_0 = 0.5$, and $\overline{N} = 30$. On average the standard error of the simulations was 0.0009

3.5. Discussion

In this study, we examined the bias and mean squared error of the average effect size, the empirical coverage and interval width of confidence intervals around the average effect size, and the bias of standard error estimates of various meta-analytic methods, when the normality assumption is not met in a random-effects model. A wide variety of

realistic scenarios in clinical psychology research was considered, and the standardized mean difference as the effect size measure.

In the random-effects model, normality of the effect parameter distribution is a usual model assumption, and several authors have raised concerns regarding the potential impact of non-normality on the performance of meta-analytic techniques (Borenstein et al., 2010; Brockwell & Gordon, 2001, 2007; Kontopantelis & Reeves, 2012a, 2012b; Sidik & Jonkman, 2002, 2007). In our study the performance of several meta-analytic methods was compared, and our results suggested that most were not substantially affected by the underlying distribution of effect parameters, even under severe departures from normality. In fact, an unexpected result in our study was the slightly lower bias of the mean effect size estimates for conditions with the most severe departure from normality (skewness = -2 and kurtosis = 3.65) in comparison with the other combinations of skewness and kurtosis. Thus, violation of the normality assumption does not seem to be critical in the estimation of an overall effect in random-effects meta-analysis.

Our findings are in agreement, in general, with the previous works of Kontopantelis and Reeves (2012a, 2012b) in the epidemiological field. In our study the manipulated conditions were related to the psychological field, where, for instance, it is more frequent to find meta-analyses with a larger number of studies and with the standardized mean difference as the effect size index. We also manipulated the average total sample size of the individual studies and the grand mean of the distribution of effect parameters. Furthermore, we considered other heterogeneity variance estimators different to DerSimonian and Laird and examined the non-parametric bootstrapping method. It is important to note that a limitation of Kontopantelis and Reeves (2012a, 2012b) works was that they used an inappropriate method to generate the individual log odd-ratios, a method that, contrary to what Kontopantelis and Reeves stated, cannot be applied to other effect metrics.

It was not surprising to find a weak performance of the average effect based on the fixed-effect model in scenarios where $\tau^2 > 0$, as this method assumes that the studies estimate the same effect parameter (i.e., there is no between-studies variability). Regarding random-effects methods, which account for between-studies variability, results were not found to be affected by the heterogeneity estimator used.

Some authors have criticized the standard random-effects method for not taking into account the uncertainty due to the variance estimation process, which increases the risk of false positive results (e.g., Thompson & Higgins, 2002). Our results showed that the Hartung's method outperformed the standard one, with a better coverage of the nominal confidence level. This trend was also reported in previous simulation studies restricted to normal scenarios (IntHout et al., 2014; Sánchez-Meca & Marín-Martínez, 2008; Viechtbauer et al., 2015). Nevertheless, for a small number of studies the CIs obtained by Hartung's method were very wide. Compared to Hartung's, the profile likelihood method provided narrower CIs. Both methods achieved coverage probabilities close to the nominal confidence level, with slightly lower values for the profile likelihood method.

The last method examined was non-parametric bootstrapping, which makes no distributional assumptions. Despite its theoretical advantage under non-normal scenarios, this method did not show a better performance than the standard, nor for Hartung's nor profile likelihood methods across the set of manipulated conditions and the comparative criteria considered in our study. This method requires substantially more computational resources, and no empirical results were found to encourage its use.

From the factors manipulated in this simulation, our results suggest that the number of studies exerts an important influence on the performance of the methods compared. With a small number of studies (less than 20) the performance of the methods was poorer and there were more notable differences among them than for a large number of studies. Similar results were observed in previous studies simulating normal scenarios (López-López et al., 2014; Rubio-Aparicio, Sánchez-Meca, López-López, Marín-Martínez, & Botella, 2017). Many meta-analyses in clinical psychology include fewer than 20 studies, and the picture is even more extreme in other health sciences (Davey, Turner, Clarke, & Higgins, 2011). Moreover, our results suggest that a large between-studies heterogeneity led to less accurate results and more pronounced differences among methods.

In conclusion, the results of our simulation study suggest that the most commonly used meta-analytic techniques are robust to violation of the normality assumption of the parametric effects distribution. All random-effects methods examined, as well as non-parametric bootstrapping, yielded similar results under optimal conditions (e.g. moderate

to large number of studies, small between-studies heterogeneity). However, we recommend using the Hartung's method and profile likelihood method to construct a CI for the average effect, due to their suitability in a wide range of scenarios and their computational simplicity. Nonetheless, the results of our study are limited to the manipulated conditions, so that future studies are warranted to improve the generalizability of these findings, extending the manipulated conditions and considering other effect size indices. Finally, our conclusions do not only apply to the estimation of an overall effect size alongside its confidence interval under random-effects models, but also to the analysis of the influence of moderator variables under mixed-effects models. Indeed, when the influence of a categorical moderator variable on the effect sizes is investigated, average effect sizes and CIs for each subgroup are calculated. Thus, our recommendation of using Hartung's or profile likelihood methods for that purpose can also be extended to the estimation of the true effect of each category of the moderator.

Chapter 4

Study 3:

"Analysis of categorical moderators in mixed-effects meta-analysis: Consequences of using pooled versus separate estimates of the residual between-studies variances

4.1. Introduction

One of the main purposes of meta-analysis is to examine whether the individual effect sizes are homogeneous around the average effect size. When there is more heterogeneity than expected from sampling error alone, the meta-analyst must search for study characteristics that can explain at least part of that variability. The moderators are considered as potential predictor variables and the effect sizes constitute the dependent variable (Borenstein et al., 2009). If the moderator variable is categorical, an analysis of variance, or subgroup analysis, can be formulated, while the continuous moderators are analysed using meta-analytic analogues to regression analysis. The

analysis of categorical moderators is usually referred to as 'subgroup analysis', and is the process of comparing the mean effect sizes in different subgroups of studies (Borenstein & Higgins, 2013).

Several statistical models are available to examine the relationship between a categorical moderator and the effect sizes through a subgroup analysis. On the one hand, applying the logic of the general fixed-effect model to subgroup analyses, a fixed-effects model can be assumed in which all studies within the same category of the moderator share a common effect size. In other words, if a fixed-effect model is assumed within each subgroup, such model is called a fixed-effects model. On the other hand, the mixed-effects model consists of assuming a random-effects model for each subgroup of studies. As a consequence, the mixed-effects model assumes that all studies within the same category of the moderator estimate a normal distribution of population effect sizes with a common mean effect size. The label 'mixed-effects model' is used because: (1) the moderator is considered a fixed-effects component, as the categories of the moderator are not a random sample of a larger number of categories, and (2) the effect sizes (i.e., the studies) include a random-effects component because they are considered a random sample of study effects pertaining to a population of studies in the same category (Borenstein et al., 2009; Viechtbauer, 2010).

In this study, we focus on the performance of the mixed-effects model, which is nowadays routinely applied in most meta-analytic studies.

4.1.1. Mixed-effects model

Suppose that the k studies in a meta-analysis are grouped into m mutually exclusive categories of the moderator variable. Moreover, k_1, k_2, ..., k_m denote the number of effect sizes of the categories 1, 2, ..., m, respectively, such that $k_1 + k_2 + ... + k_m = k$.

In a mixed-effects model the individual effect sizes, T_{ij}, within the same category j are assumed to estimate a distribution of true effect sizes with mean $\mu_{\theta j}$ and variance $\sigma_{ij}^2 + \tau_j^2$, with σ_{ij}^2 being the within-study variance for the ith study in the jth category of the moderator, and τ_j^2 the residual between-studies variance in that category.

We must assume a random-effects model within each category of the moderator variable, thus the statistical model applied in the *j*th category will be $T_{ij} = \mu_{\theta j} + \varepsilon_{ij} + e_{ij}$, where ε_{ij} and e_{ij} are the within-study and between-studies errors, respectively. It is very common to assume that these two errors are independent of each other and, therefore, the estimated effect sizes are normally distributed: $T_{ij} \sim N(\mu_{\theta j}, \sigma_{ij}^2 + \tau_j^2)$, where τ_j^2 is the common between-studies variance in *j*th category of the moderator. In addition, the parametric effect sizes of the *j*th category, θ_{ij}, follow a normal distribution with mean $\mu_{\theta j}$ and between-studies variance τ_j^2 : $\theta_{ij} \sim N(\mu_{\theta j}, \tau_j^2)$.

Under a mixed-effects model, the main goal in a subgroup analysis is to compare the parametric mean effect sizes from each category of the moderator variable, $\mu_{\theta j}$, in order to test if the moderator is statistically related to the effect sizes. Consequently, first we need to estimate the mean parametric effect size of the *j*th category of the moderator, $\mu_{\theta j}$, by means of

$$\overline{T}_j = \frac{\sum_i \hat{w}_{ij} T_{ij}}{\sum_i \hat{w}_{ij}} \tag{4.1}$$

where $\hat{w}_{ij}$ are the estimated weights computed through $\hat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_j^2)$, with $\hat{\sigma}_{ij}^2$ being the estimated within-study variance of the *i*th effect size and $\hat{\tau}_j^2$ the estimated residual between-studies variance of the *j*th category.

The sampling variance of the mean effect size in the *j*th category is estimated as

$$V(\overline{T}_j) = \frac{1}{\sum_i \hat{w}_{ij}} \ . \tag{4.2}$$

4.1.2. Omnibus Test of Between-Groups Differences

It is possible to test the statistical significance of a categorical moderator by means of the omnibus Wald-type χ^2 test, the Q_B test, obtained with (Borenstein et al., 2009)

$$Q_B = \sum_{j=1}^{m} \hat{w}_{+j} \left(\overline{T}_j - \overline{T} \right)^2 , \qquad (4.3)$$

where $\hat{w}_{+j}$ is the inverse of Eq. 4.2 applied to the *j*th category of the moderator, $\overline{T}_j$ is the mean effect size of the *j*th category calculated by Eq. 4.1 and $\overline{T}$ represents the weighted grand mean of all effect sizes and is given by

$$\overline{T} = \frac{\sum_i \sum_j \hat{w}_{ij} T_{ij}}{\sum_i \sum_j \hat{w}_{ij}}, \qquad (4.4)$$

where the total between-studies variance estimate, $\hat{\tau}^2$, is used to compute $\hat{w}_{ij}$.

Under the null hypothesis of no difference between the mean effect sizes for each of the *m* categories (H_0: $\mu_{\theta 1} = \mu_{\theta 2} = \ldots = \mu_{\theta m}$), the Q_B statistic follows a Chi-square distribution with $m - 1$ degrees of freedom. Therefore, the null hypothesis will be rejected when Q_B exceeds the $100(1 - \alpha)$ percentile point of the chi-square distribution. A statistically significant result for Q_B provides evidence that the moderator is statistically related to the effect sizes.

4.1.3. Estimating the residual between-studies variance

Several methods have been proposed to estimate the total heterogeneity variance in the random-effects model. The most commonly used is that proposed by DerSimonian and Laird (1986), a heterogeneity variance estimator derived from the moment method.

At this point, it could be useful to make a distinction between the total between-studies variance and the residual between-studies variance. On the one hand, when we apply the random-effects model to estimate the mean effect in a meta-analysis (i.e., without moderators being added to the model) there is an amount of heterogeneity due to sampling error in the selection of the studies in the meta-analysis. This heterogeneity is

estimated through the total between-studies variance, which represents the excess variation among the effects over that expected from within-study sampling error alone. On the other hand, in the mixed-effects model we include moderator variables aiming to explain at least part of the total heterogeneity in the effect sizes. Thus, after adding moderator variables the amount of heterogeneity that remains to be explained is the residual heterogeneity or the heterogeneity that cannot be explained by the moderators included in the model.

In the mixed-effects model, two approaches can be adopted to estimate the residual between-studies variance. One is to estimate the residual between-studies variance separately within each category of the moderator, and the other one is to calculate a pooled estimate across categories (Borenstein et al., 2009).

Separate estimates of the residual between-studies variance

This procedure consists of estimating the residual between-studies variance within each category of the moderator. Thus, in a moderator variable with *m* categories, we need to calculate the residual between-studies variance estimates $\hat{\tau}_1^2$, $\hat{\tau}_2^2$, …, and $\hat{\tau}_m^2$ The residual between-studies variance for the *j*th category of the moderator, $\hat{\tau}_j^2$, can be computed applying the Dersimonian and Laird estimator with the expression

$$\hat{\tau}_j^2 = \frac{Q_{wj} - (k_j - 1)}{c_j}, \tag{4.5}$$

where k_j is the number of studies of the *j*th category, Q_{wj} is the within-group homogeneity statistic of the *j*th category computed through

$$Q_{w_j} = \sum_{i=1}^{k_j} \hat{w}_{ij}^{*} \left(T_{ij} - \overline{T}_j^{*} \right), \tag{4.6}$$

with $\hat{w}_{ij}^{*}$ being the estimated weights assuming a fixed-effect model, $\hat{w}_{ij}^{*} = 1/\hat{\sigma}_{ij}^{2}$, and $\overline{T}_{j}^{*}$ the mean effect size of the *j*th category of the moderator also assuming a fixed-effect model, that is, applying Eq. 4.1 but using $\hat{w}_{ij}^{*}$ as weighting factor; and c_j is given by

$$c_j = \sum_i \hat{w}_{ij}^{*} - \frac{\sum_i \left(\hat{w}_{ij}^{*}\right)^2}{\sum_i \hat{w}_{ij}^{*}} \quad . \tag{4.7}$$

Therefore, Eq. 4.5 allows a separate estimate of the between-studies variance of each category, $\hat{\tau}_j^2$, to be obtained, and these are used to calculate the weights, $\hat{w}_{ij}$, for each category of the moderator. This implies that in each category a different between-studies variance is used to calculate the weights: $\hat{\tau}_1^2$ for category 1, $\hat{\tau}_2^2$ for category 2, and so on, that is, $\hat{w}_{ij} = 1/\left(\hat{\sigma}_{ij}^2 + \hat{\tau}_j^2\right)$. Here we will name the Q_B statistic calculated with separate between-studies variances as $Q_{B(S)}$.

Pooled estimate of the residual between-studies variance

An alternative method to estimate the residual heterogeneity variance consists of averaging the residual between-studies variances of the *m* categories of the moderator variable, through the equation (Borenstein et al., 2009)

$$\hat{\tau}_{+}^2 = \frac{\sum_j^m Q_{wj} - \sum_j^m (k_j - 1)}{\sum_j^m c_j} \quad . \tag{4.8}$$

Eq. 8 provides a pooled estimate of the residual between-studies variance, so that the weights, $\hat{w}_{ij}$, are obtained using a common between-studies variance through the

different categories of the moderator, that is, $\hat{w}_{ij} = 1/\left(\hat{\sigma}_{ij}^2 + \hat{\tau}_{+}^2\right)$. Here we will use the term $Q_{B(P)}$ to refer to the Q_B statistic calculated with a pooled estimate of the residual between-studies variance, $\hat{\tau}_{+}^2$.

4.1.4. An example

To illustrate how the Q_B statistic is calculated with the two different methods to estimate the residual between-studies variance (pooled vs. separate estimates), an example extracted from a real meta-analysis is presented here. The data were obtained from a meta-analysis about the efficacy of psychological treatments for panic disorder with or without agoraphobia (Sánchez-Meca, Rosa-Alcázar, Marín-Martínez, & Gómez-Conesa, 2010). The effect size index in this meta-analysis was the standardized mean difference (*d*) between two groups (treated vs. control groups) defined in Eq. 4.1. Out of all the moderator variables analyzed in this meta-analysis, a dichotomous characteristic was selected to illustrate a subgroup meta-analysis: whether or not the assignment of the participants to the treated and control groups was at random. The database composed of 50 studies is presented in Appendix 4A.

Table 4.1. Results of the subgroup analysis for the moderator variable 'random assignment' in the Sánchez-Meca et al. (2010) meta-analysis by using separate estimates of the residual between-studies variance, $\hat{\tau}_j^2$.

				95% CI		
Random assignment	k_j	$\bar{d}_j$	$V(\bar{d}_j)$	d_l	d_u	$\hat{\tau}_j^2$
No	8	0.545	0.024	0.242	0.847	0.053
Yes	42	0.966	0.011	0.765	1.167	0.303
Separate estimates of $\hat{\tau}_j^2$:	$Q_{B(S)}(1) = 5.165$, $p = .023$					

Note. k_j = number of studies in each category of the moderator; $\bar{d}_j$ = mean effect size for each category, obtained with Eq. 4.1; $V(\bar{d}_j)$ = estimated sampling variance of the mean effect size for each category, obtained with Eq. 4.2; d_l and d_u = lower and upper confidence limits (for a 95% confidence level) for each mean effect size, obtained by means of $\bar{d}_j \pm 1.96 \times \sqrt{V(\bar{d}_j)}$ (1.96

being the 97.5% percentile of the standard normal distribution); $\hat{\tau}_j^2$ = residual between-studies variance for each category, estimated with Eq. 4.5.

Table 4.2. Results of the subgroup analysis for the moderator variable 'random assignment' in the Sánchez-Meca et al. (2010) meta-analysis by using a pooled estimate of the residual between-studies variance, $\hat{\tau}_+^2$.

Random assignment	k_j	$\bar{d}_j$	$V(\bar{d}_j)$	95% CI d_l	95% CI d_u	$\hat{\tau}_+^2$
No	8	0.559	0.053	0.109	1.009	0.270
Yes	42	0.961	0.010	0.768	1.155	0.270
Pooled estimate of $\hat{\tau}_j^2$:	$Q_{B(P)}(1) = 2.588$, $p = .108$					

Note. k_j = number of studies in each category of the moderator; $\bar{d}_j$ = mean effect size for each category, obtained with Eq. 4.1; $V(\bar{d}_j)$ = estimated sampling variance of the mean effect size for each category, obtained with Eq. 4.2; d_l and d_u = lower and upper confidence limits (for a 95% confidence level) for each mean effect size, obtained by means of $\bar{d}_j \pm 1.96 \times \sqrt{V(\bar{d}_j)}$ (1.96 being the 97.5% percentile of the standard normal distribution); $\hat{\tau}_+^2$ = pooled estimate of the residual between-studies variances of the two categories, calculated with Equation 4.8.

Tables 4.1 and 4.2 present the results yielded by the Q_B statistic with the two methods here compared, as well as the mean effects for each category of the moderator, the sampling variances, the residual between-studies variances and the 95% confidence intervals for each mean effect. Separate estimates of the residual between-studies variances for each category ($\hat{\tau}_j^2$) were calculated using Eq. 4.5. As shown in Table 4.1, their values were 0.053 and 0.303 for non-random and random assignment, respectively. On the other hand, the pooled estimate of the residual between-studies variances calculated using Eq. 4.8 was $\hat{\tau}_+^2 = 0.270$ (Table 4.2). When the Q_B statistic was calculated taking separate estimates of the residual between-studies variances, the estimated weights for each study were obtained by means of $\hat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_j^2)$. Conversely, when the Q_B statistic was calculated taking a pooled estimate of the residual between-studies variances ($\hat{\tau}_+^2$), the estimated study weights were $\hat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_+^2)$. This distinction affects the Q_B statistic, here referred as $Q_{B(S)}$ and $Q_{B(P)}$, respectively, as well as the mean effect from

each category of the moderator, their sampling variances ($V(\bar{d}_j)$), and their confidence limits.

The mean effects for non-random and random assignment were 0.545 and 0.966, respectively (Table 4.1), when separate estimates of the residual between-studies variances were used ($\hat{\tau}_j^2$), and 0.559 and 0.961 when a pooled estimate ($\hat{\tau}_+^2$) was used (Table 4.2). The sampling variances and the confidence limits also varied depending on the residual between-studies variances used in the calculations. However, the most dramatic discrepancy among methods involved the two versions of the Q_B statistic: the $Q_{B(S)}$ and $Q_{B(P)}$ statistics. The null hypothesis of equal mean effect sizes was rejected when separate estimates of the between-studies variances were used (Table 4.1: $Q_{B(S)} = 5.165$, $p = .023$), but not when a pooled estimate was considered (Table 4.2: $Q_{B(P)} = 2.588$, $p = .108$).

This example illustrates how results and their interpretation can be affected by the meta-analytic methods selected to undertake the statistical analyses. The choice of the meta-analyst will often be conditioned by the software used for the calculations and he/she will not be aware of which method was implemented. In fact, the most commonly used statistical programs for meta-analysis do not enable users to choose among the two methods to calculate the individual weights in a mixed-effects model. For instance, if the meta-analyst would use *metafor* (Viechtbauer, 2010), *Comprehensive Meta-analysis 2.0* (Borenstein, Hedges, Higgins, & Rothstein, 2005) or the SPSS macros elaborated by David B. Wilson to replicate this example, he/she would obtain the results presented in Table 4.2, whereas if using *RevMan 5.3* (Review Manager, 2014), the results will be those presented in Table 4.1. On the other hand, *Comprehensive Meta-analysis 3.0* (Borenstein, Hedges, Higgins, & Rothstein, 2014) incorporates both methods so that the meta-analyst can use either to estimate the weights (in fact, the results in Tables 4.1 and 4.2 were obtained with this program).

4.1.5. Purpose of the study

It is not clear which of these two procedures (separate or pooled estimates) should be preferred in order to estimate the residual between-studies variance, which is involved in the subgroup analysis in a mixed-effects meta-analysis. At this point, it is useful to

revise the analogy between the subgroup analysis in a meta-analysis and the analysis of variance (ANOVA) for comparing means in a primary study. On the one hand, in the simplest case of a primary study with a two-independent group design (e.g. experimental *vs*. control group), the means of two samples of subjects are compared performing a *t*-test or an ordinary least squared ANOVA. On the other hand, in a meta-analysis with two subgroups of studies, the mean effect sizes in each subgroup are compared by performing a weighted least squared ANOVA, the weights being the inverse-variance of each effect size.

Both the *t*-test or ANOVA for comparing the means of two or more independent groups of subjects assume homogeneity between variances in the two populations. The pooled variance is estimated through the mean squared error in the ANOVA. When the two population variances are heterogeneous, the so-called Behrens-Fisher problem arises, which requires an alternative procedure to the classic *t*-test or ANOVA. In practice, the usual solution to the Behrens-Fisher problem is to apply the Welch-Satterthwaite approach to correct the classical *t*-test (Welch, 1947).

In the meta-analytic arena, the picture is a little more complex, as we are working with aggregate scores (e.g. effect sizes summarizing individual scores) studies instead of groups of subjects. While in a primary study each subject provides a score, in a meta-analysis, each study provides an effect size. The effect sizes of the studies in a meta-analysis will exhibit different precision depending on the sample size of the study. Effect sizes obtained from large samples will be more accurate (less variable) than those obtained from small ones. As a consequence, the appropriate mean of a set of effect sizes is a weighted average, the weights being the inverse-variance of each effect size. This weighting scheme affects all statistical calculations in a meta-analysis.

The pooled estimation of the residual between-studies variance from two or more subgroups of studies in a meta-analysis, is akin to the estimation of the mean squared error in the ANOVA in a primary study, as both procedures assume the variance between groups to be homogeneous. When this assumption is not tenable, a problem similar to that of Behrens-Fisher emerges, which may lead to inaccurate estimation of the residual between-studies variance. To circumvent this problem, an alternative is the separate estimation of the residual between-studies variance for each subgroup of studies.

However, this approach can also yield inaccurate estimates if the number of studies in the subgroups is small (which will often be the case).

In a mixed-effects meta-analysis, the residual between-studies variance is included in the weighting scheme. Thus, the estimation procedure for the residual between studies variance may have an impact on a wide range of meta-analytic outputs, including, such as: (1) the estimate of the average effect size for each category of the moderator (see Eq. 4.1); (2) their sampling variances; (3) the confidence intervals; and, (4) relevant to the present work, the computation of the between-group heterogeneity statistic, Q_B (see Eq. 4.3).

The large number of factors that can affect the performance of the $Q_{B(P)}$ and $Q_{B(S)}$ statistics lead to the need for simulation studies to determine which of them is a better option under different meta-analytic conditions. Previous simulation studies have examined the statistical performance of the *t*-test and ANOVA *F*-test in a primary study, assuming homogeneous and heterogeneous population variances. However, those studies do not address the more complex picture of subgroup analyses in meta-analysis, and therefore their findings might not be generalizable to the meta-analytic arena.

The purpose of this work was to directly compare, by means of Monte Carlo simulation, the statistical performance of the Q_B statistic applied in meta-analysis, when two alternative procedures for estimating the residual between-studies variance (separate estimates and pooled estimate) are used. With that aim, the present work is the first simulation study where the $Q_{B(S)}$ and $Q_{B(P)}$ tests were compared, assessing their Type I error and statistical power in different meta-analytic scenarios.

The existence of previous simulation studies addressing the heterocedasticity problem in primary studies enables us to formulate some expectations (Glass & Hopkins, 1996; Glass, Peckham & Sanders, 1972; Hinkle, Wiersma & Jurs, 2003; Senn, 2008). First, in scenarios with balanced sample sizes, we expect $Q_{B(P)}$ to provide an adequate adjustment of the Type I error, even with heterogeneous variances between subgroups. Second, in unbalanced scenarios with heterogeneous variances where the larger variance is associated with the bigger subgroup, the $Q_{B(P)}$ test will be too conservative, and too liberal if the smaller variance is associated with the subgroup with the bigger subgroup instead.

4.2. Method of the Simulation Study

A simulation study was carried out in R using the *metafor* package (Viechtbauer, 2010) and the two procedures (pooled versus separate) for estimating the residual between-studies variance were programmed. Meta-analyses of *k* studies were simulated with the standardized mean difference as the effect size index. Each individual study included in a meta-analysis compared two groups (experimental and control) with respect to some continuous outcome. Both populations were normally distributed with homogeneous variances, [$N(\mu_E, \sigma^2)$, $N(\mu_C, \sigma^2)$]. The population standardized mean difference, θ, was defined as (Hedges & Olkin, 1985)

. $$\theta = \frac{\mu_E - \mu_C}{\sigma} \tag{4.9}$$

The parametric effect size, θ, can be estimated by means of

$$\hat{\theta} = c(m)\frac{\bar{y}_E - \bar{y}_C}{s}, \tag{4.10}$$

where $\bar{y}_E$ and $\bar{y}_C$ are the sample means of experimental and control groups, *S* is a pooled standard deviation computed through

$$S = \sqrt{\frac{(n_E - 1)S_E^2 + (n_C - 1)S_C^2}{n_E + n_C - 2}}\ , \tag{4.11}$$

n_E and n_C being the experimental and control sample sizes, respectively, S_E^2 and S_C^2 being the unbiased variances of the two groups, and c(*m*) is a correction factor for small sample sizes, given by

$$c(m) = 1 - \frac{3}{4N - 9} \tag{4.12}$$

being $N = n_E + n_C$. The estimated within-study variance of $\hat{\theta}$, assuming equal variances and normality within each study, is given by

$$\hat{\sigma}_d^2 = \frac{n_E + n_C}{n_E n_C} + \frac{\hat{\theta}^2}{2(n_E + n_C)} \ . \tag{4.13}$$

We simulated a mixed-effects model involving a moderator variable with two categories. In each category of the moderator variable a population of parametric effect sizes was assumed, in addition to the within-group variability.

The number of studies of each simulated meta-analysis was defined as $k = k_1 + k_2$, with k_1 and k_2 being the number of studies falling into the first and second categories of the moderator, respectively.

The manipulated conditions in the present study were intended to represent the most realistic scenarios found in meta-analysis. For the number of studies, k, we considered four values, namely 12, 20, 40, and 60. Furthermore, we manipulated the distribution of k within each category of the moderator, so that in some conditions there was a balanced distribution (e.g. $k_1 = k_2$), while in the remaining conditions there was an unbalanced distribution between the two categories with the second category containing three times as many studies as the first category.

We also manipulated the residual between-studies variance of each category of the moderator in two different ways. First, we considered two values for this parameter, namely 0.08 and 0.16. Second, we simulated a set of scenarios with homogeneous residual between-studies variances for both categories ($\tau_1^2 = \tau_2^2$), and also another set of heterogeneous conditions, with values $\tau_1^2 = 0.08$ and $\tau_2^2 = 0.16$ or $\tau_1^2 = 0.16$ and $\tau_2^2 = 0.08$.

The average sample size of the k studies in a meta-analysis was set to 60. Note that, for each study, $N = n_E + n_C$, with $n_E = n_C$. The selection of the sample sizes for the individual studies in each meta-analysis was performed from the generation of skewed distributions, applying the Fleishman's algorithm (1978) with an average value of 60, a skewness index of 1.386, a kurtosis index of 1.427 and a standard deviation of 5.62. The

parameters of this distribution are similar to the distribution of sample sizes found in a recent review of 50 real meta-analyses on the effectiveness of psychological treatments (Rubio-Aparicio et al., in press).

The parametric mean effect size of each category of the moderator was also manipulated. In some conditions the two parametric mean effects were equal to 0.5 ($\mu_{\theta 1} = \mu_{\theta 2} = 0.5$), whereas for other conditions they were set to different values: $\mu_{\theta 1} = 0.5$ and $\mu_{\theta 2} = 0.3$ or $\mu_{\theta 1} = 0.5$ and $\mu_{\theta 2} = 0.1$. Moreover, when the parametric mean effect sizes were different for each category, their position was also manipulated, and hence we also generated scenarios with $\mu_{\theta 1} = 0.3$ and $\mu_{\theta 2} = 0.5$ or $\mu_{\theta 1} = 0.1$ and $\mu_{\theta 2} = 0.5$. The conditions with equal parametric mean effect sizes across categories allowed us to study the Type I error rate of the $Q_{B(S)}$ and $Q_{B(P)}$ statistics, whereas the conditions with different parametric mean effect sizes enabled us to assess their statistical power.

To assess the Type I error rate, the total number of conditions was: 4 (number of studies) x 2 (balanced-unbalanced number of studies in the two categories) x 4 (residual between-studies variance) = 32. With respect to the statistical power, the conditions were quadrupled regarding those of the Type I error by including two different parametric mean effect sizes and manipulating their position across categories, so that there were 32 x 4 = 128 conditions defined. To sum up, the total number of conditions was 160 and for each one 10,000 replications were generated. Thus, 1,600,000 meta-analyses were simulated.

The $Q_{B(S)}$ test (Eq. 4.3) using separate estimates of τ^2 for each subgroup (Eq. 4.5) and the $Q_{B(P)}$ test when using a pooled estimate of τ^2 (Eq. 4.8) were applied to each one of these replications. In each of the 160 conditions of our simulation study, the proportion of rejections of the null hypothesis of equality of the parametric mean effect sizes of the moderator enabled us to estimate the Type I error rate and the statistical power.

Appendix 4B shows the R code of our simulation study.

4.3. Results

4.3.1. Type I error rate

Table 4.3 presents Type I error rates for the $Q_{B(S)}$ and $Q_{B(P)}$ statistics when using the two estimation procedures of the residual between-studies variance in the manipulated conditions. Table 4.4 summarizes the average Type I error rates as a function of the number of studies, balanced and unbalanced distribution of number of studies within each category of the moderator, and residual between-studies variance of each category of the moderator. Note that the nominal significance level was set to $\alpha = .05$.

Table 4.3. Type I error for the two estimation procedures of the residual between studies variance

		Balanced		Unbalanced	
$\tau_1^2 : \tau_2^2$	k	$Q_{B(S)}$	$Q_{B(P)}$	$Q_{B(S)}$	$Q_{B(P)}$
0.08 : 0.08	12	0.0611	0.0655	.0801	0.0719
	20	0.0595	0.0609	0.0743	0.0672
	40	0.0584	0.0581	0.0639	0.0577
	60	0.0543	0.0548	0.0564	0.0527
0.16 : 0.16	12	0.0737	0.0761	0.0950	0.0976
	20	0.0648	0.0650	0.0783	0.0652
	40	0.0554	0.0548	0.0696	0.0612
	60	0.0567	0.0566	0.0640	0.0579
0.08 : 0.16	12	0.0705	0.0733	0.0758	0.0524
	20	0.0602	0.0611	0.0709	0.0456
	40	0.0584	0.0580	0.0623	0.0377
	60	0.0510	0.0505	0.0552	0.0349
0.16 : 0.08	12			0.0956	0.1013
	20			0.0886	0.0949
	40			0.0716	0.0890
	60			0.0606	0.0801

Note. τ_1^2 = residual between-studies variance of the first category of the moderator; τ_2^2 = residual between-studies variance of the second category of the moderator; *k* = number of studies; Balanced = balanced distribution of *k* within each category of the moderator; Unbalanced = unbalanced distribution of *k* within each category of the moderator, with fewer studies in the first category; $Q_{B(S)} = Q_B$ test using separate estimates of τ^2 for each subgroup; $Q_{B(P)} = Q_B$ test using a pooled estimate of τ^2.

Table 4.4. Average Type I rates by number of studies (*k*), by balanced and unbalanced distribution of *k*, and by the residual between-studies variance of each category of the moderator ($\tau_1^2 : \tau_2^2$)

	$Q_{B(S)}$	$Q_{B(P)}$
k		
12	0.0788	0.0738
20	0.0709	0.0657
40	0.0628	0.0595
60	0.0569	0.0553
Distribution of *k*		
Balanced	0.0577	0.0577
Unbalanced	0.0679	0.0620
$\tau_1^2 : \tau_2^2$		
0.08 : 0.08	0.0612	0.0585
0.16 : 0.16	0.0648	0.0601
0.08 : 0.16	0.0597	0.0479
0.16 : 0.08	0.0736	0.0880

Note. $Q_{B(S)} = Q_B$ test using separate estimates of τ^2 for each subgroup; $Q_{B(P)} = Q_B$ test using a pooled estimate of τ^2.

First, in most conditions results showed the empirical rejection rates of both estimation procedures above the nominal significance level (Tables 4.3 and 4.4). As expected, as the number of studies increased, the proportion of rejections of the null hypothesis of equality for $Q_{B(S)}$ and $Q_{B(P)}$ converged to the nominal significance level (Table 4.4)

In general, when the number of studies was balanced across categories, both estimation procedures showed a good adjustment to the nominal level, with negligible differences among the empirical error rates. By contrast, under the conditions with an unbalanced distribution of studies between the two categories, the differences in error rates for both estimation procedures were most notable (Table 4.3).

As can be seen in Table 4.3, and focusing on unbalanced distribution of the number of studies within each category of the moderator, when the residual between-studies variances of each category were homogeneous ($\tau_1^2 = \tau_2^2 = 0.08$ or $\tau_1^2 = \tau_2^2 = 0.16$), the $Q_{B(P)}$ test presented a better control of the Type I error rate than $Q_{B(S)}$. In contrast, when variances were heterogeneous, specifically under the condition where the value of the smallest residual between-studies variance, $\tau^2 = 0.08$, was associated with the category with the smallest number of studies ($\tau_1^2 = 0.08$; $\tau_2^2 = 0.16$), the $Q_{B(P)}$ test showed Type I error rates below 0.05, whereas the $Q_{B(S)}$ test yielded rates over nominal except for a large number of studies, $k = 60$ ($k_1 = 15$ and $k_2 = 45$). Under the condition where the value of the largest residual between-studies variance, $\tau^2 = 0.16$, was associated with the category with the smallest number of studies, ($\tau_1^2 = 0.16$; $\tau_2^2 = 0.08$), the $Q_{B(P)}$ test showed empirical rejection rates above the nominal significance level, while the $Q_{B(S)}$ test only showed results close to the nominal level with $k = 60$ ($k_1 = 15$ and $k_2 = 45$).

4.3.2. Statistical power

Table 4.5 shows the empirical power rates for $Q_{B(S)}$ and $Q_{B(P)}$ tests in the manipulated conditions. Table 4.6 summarizes the average power rates as a function of the magnitude of the difference between the parametric mean effect sizes of each category of the moderator, number of studies, balanced and unbalanced distribution of number of studies within each category of the moderator, and the residual between-studies variance for each category of the moderator. In general, the influence of the different conditions manipulated was equivalent for the $Q_{B(S)}$ and $Q_{B(P)}$ tests and, in most conditions, both tests yielded statistical power rates far below .80 (Tables 4.5 and 4.6).

Table 4.6 shows that, as expected, $Q_{B(S)}$ and $Q_{B(P)}$ tests increased their statistical power as the number of studies and the magnitude of the difference between the parametric effect size of each category increased. Furthermore, under the conditions with

a balanced distribution of the studies across categories, the $Q_{B(S)}$ and $Q_{B(P)}$ tests showed greater power than under the condition with an unbalanced distribution of the studies (see also Table 4.5). In relation to the conditions with homogeneous residual between-studies variances, large amounts of residual τ^2 values correspond to smaller rejection rates for both tests. Accordingly, the highest power rates, $Q_{B(S)} = 0.9760$ and $Q_{B(P)} = 0.9759$, were obtained under optimal scenarios, that is, maximum difference between the parametric mean effect size of each category ($|\mu_{\theta 1} - \mu_{\theta 2}| = 0.4$), large number of studies ($k = 60$), balanced distribution of studies within each category and small and homogeneous values of the residual between-studies variance of each category ($\tau_1^2 = 0.08$ and $\tau_2^2 = 0.08$) (Table 4.5).

As shown in Table 4.5, under a balanced distribution of the number of studies within each category of the moderator, the $Q_{B(S)}$ and $Q_{B(P)}$ tests performed very similarly, even when the assumption of homogeneity variances was not fulfilled. By contrast, when the number of studies was distributed unequally within each category of the moderator and the residual between-studies variances of each category were homogeneous, the $Q_{B(S)}$ test yielded a slightly higher power than $Q_{B(P)}$ test.

Table 4.5. Statistical Power Rates for the two estimation procedures of the residual between-studies variance

		$\lvert\mu_{\theta 1} - \mu_{\theta 2}\rvert = 0.2$				$\lvert\mu_{\theta 1} - \mu_{\theta 2}\rvert = 0.4$			
		Balanced		Unbalanced		Balanced		Unbalanced	
$\tau_1^2 : \tau_2^2$	k	$Q_{B(S)}$	$Q_{B(P)}$	$Q_{B(S)}$	$Q_{B(P)}$	$Q_{B(S)}$	$Q_{B(P)}$	$Q_{B(S)}$	$Q_{B(P)}$
0.08 : 0.08	12	0.161	0.1701	0.1599	0.151	0.4383	0.4479	0.3645	0.3638
	20	0.2203	0.2235	0.1894	0.1827	0.6341	0.6385	0.5293	0.5298
	40	0.3796	0.3783	0.3028	0.2953	0.8988	0.9000	0.8028	0.8068
	60	0.5224	0.5220	0.4168	0.4116	0.9760	0.9759	0.9296	0.9323
0.16 : 0.16	12	0.1446	0.1483	0.1505	0.1294	0.3298	0.3329	0.3012	0.2792
	20	0.1752	0.1768	0.1642	0.1489	0.4803	0.4804	0.4004	0.3893
	40	0.2756	0.2753	0.2269	0.2175	0.7501	0.7502	0.6305	0.6285
	60	0.3710	0.3700	0.3139	0.3060	0.8979	0.8971	0.7972	0.7994
0.08 : 0.16	12	0.1512	0.1567	0.1405	0.1046	0.3759	0.3831	0.3342	0.2635
	20	0.1986	0.2025	0.1749	0.1261	0.5392	0.5443	0.4772	0.4022
	40	0.3136	0.3198	0.2802	0.2130	0.8275	0.8299	0.7542	0.6905
	60	0.4377	0.4432	0.3787	0.3024	0.9478	0.9493	0.9007	0.8615
0.16 : 0.08	12	0.1466	0.1512	0.3808	0.1749	0.3677	0.3729	0.3204	0.3541
	20	0.1918	0.1922	0.1778	0.2062	0.5441	0.5443	0.4271	0.4823
	40	0.3146	0.3098	0.2489	0.2960	0.8241	0.8213	0.6763	0.7373
	60	0.4355	0.4274	0.3249	0.3832	0.9432	0.9422	0.8268	0.8748

Note. $\mu_{\theta 1}$ = parametric mean effect size of the first category of the moderator; $\mu_{\theta 2}$ = parametric mean effect size of the second category of the moderator; τ_1^2 = residual between-studies variance of the first category of the moderator; τ_2^2 = residual between-studies variance of the second category of the moderator; k = number of studies; Balanced = balanced distribution of *k* within each category of the moderator; Unbalanced = unbalanced distribution of *k* within each category of the moderator, where the number of studies in the first category is the lowest one; $Q_{B(S)} = Q_B$ test using separate estimates of τ^2 for each subgroup; $Q_{B(P)} = Q_B$ test using a pooled estimate of τ^2.

Table 4.6. Average power values rates by difference between the parametric mean effect size of each category of the moderator ($|\mu_{\theta 1} - \mu_{\theta 2}|$), by number of studies (*k*), by balanced and unbalanced distribution of *k*, and by the residual between-studies variance of each category of the moderator ($\tau_1^2 : \tau_2^2$)

	$Q_{B(S)}$	$Q_{B(P)}$
$\|\mu_{\theta 1} - \mu_{\theta 2}\|$		
0.2	0.2843	0.2783
0.4	0.7102	0.7095
k		
12	0.2674	0.2418
20	0.3359	0.3307
40	0.5179	0.5148
60	0.6378	0.6362
Distribution of *k*		
Balanced	0.5458	0.5464
Unbalanced	0.4729	0.4676
$\tau_1^2 : \tau_2^2$		
0.08 : 0.08	0.5540	0.5530
0.16 : 0.16	0.4453	0.4405
0.08 : 0.16	0.5109	0.4711
0.16 : 0.08	0.4787	0.5109

Note. $Q_{B(S)} = Q_B$ test using separate estimates of τ^2 for each subgroup; $Q_{B(P)} = Q_B$ test using a pooled estimate of τ^2.

4.4. Discussion

This study compares the impact of two procedures for estimating the residual between-studies variance, separate estimates and pooled estimate, on the statistical performance of the Q_B test for subgroup analyses assuming a mixed-effects meta-analysis. Our work is the first simulation study addressing the question of which estimation procedure of the residual between-studies variance yields the most accurate

results for the Q_B test under a set of realistic scenarios, and also allows us to explore the practical consequences of using separate estimates or a pooled estimate.

Under a balanced distribution of the number of studies across categories, we expected good performance of the $Q_{B(P)}$ test even when the assumption of homogeneity of the residual between-studies variances was not fulfilled. This is a similar situation to that of the typical ANOVA *F*-test with equal sample sizes between groups of subjects, where the *F*-test is robust to violations of the homoscedasticity assumption (Glass & Hopkins, 1996; Senn, 2008). Our results showed similar Type I error rates for the $Q_{B(P)}$ test in the conditions with homogeneous and heterogeneous residual between-studies variances. However, the empirical Type I error rates showed a good adjustment to the nominal level only in meta-analyses with large number of studies (40 or more studies), the adjustment becoming slightly more liberal as the number of studies decreased.

Comparing the performance of the $Q_{B(S)}$ and $Q_{B(P)}$ tests, their Type I error and statistical power rates were similar through all the conditions of subgroups with equal number of studies. This suggests that when the studies are distributed equally within each category of the moderator the meta-analyst may apply any of the procedures in order to estimate the residual between-studies variance. Nevertheless, if the number of studies and the residual between-studies variances are roughly similar across categories, using a pooled estimate would be expected to provide more accurate results for most scenarios, as it takes into account a larger number of studies. This can be particularly important if the total number of studies is small (e.g., <20), which has been found to be the case for most Cochrane Reviews (Davey et al., 2011).

When the number of studies was distributed unequally across categories, the practical consequences of having heterogeneous residual between-studies variances were more evident, with both tests leading to the wrong statistical conclusion more often than in the conditions with balanced subgroups. Specifically, under the condition of heterogeneity where the value of the smallest residual between-studies variance (τ^2 = 0.08) was associated with the category with the smallest number of studies, the $Q_{B(S)}$ test showed adequate control of the Type I error rate with at least 60 studies, whereas that the $Q_{B(P)}$ test yielded over-conservative Type I error rates and poor performance in terms of statistical power regardless of the number of studies. Under conditions where the value of the largest residual between-studies variance (τ^2 = 0.16) was associated with the

category with the smallest number of studies, both tests provided inflated Type I error rates, with the $Q_{B(P)}$ test showing a greater departure from the nominal significance level. Note that the performance of the $Q_{B(P)}$ test was similar to that expected for the *F*-test in a typical ANOVA with unbalanced sample sizes, when the homoscedasticity assumption was not met (Glass et al., 1972; Hinkle et al., 2003).

Lastly, our results also reflect that the $Q_{B(P)}$ test yielded more accurate control of error rates when the residual between-studies variances homogeneity assumption was fulfilled. In practice, the Q_B test is usually calculated using a pooled estimate (Borenstein et al., 2009; Viechtbauer, 2010). Borenstein et al. (2009) and Viechtbauer (2010) suggested using a pooled estimate of the residual between-studies variance except when the meta-analyst suspects that the true value of the residual between-studies may vary from one category to the next.

As pointed out in the introduction, the most popular statistical packages for meta-analysis estimate the residual between-studies variance implementing only one of the two procedures described and compared throughout this study, so that choice of software determines the method to be used. Our results showed some evidence that pooled or separate estimates might lead to a different performance of the Q_B test under some scenarios. Therefore, it would be helpful for the different meta-analysis software options to allow users to implement either method based on the characteristics of the database, as it is already the case for *Comprehensive Meta-analysis 3.0* (Borenstein et al., 2014). That would also allow undertaking sensitivity analyses if the meta-analyst suspects that the choice of procedure may have an impact on the results.

Results from our simulation study also shed some light on the accuracy of hypothesis testing for categorical moderators in meta-analysis, beyond the choice of pooled or separate variance estimates. The overall picture suggests that statistical tests can be expected to perform close to the nominal significance level in terms of Type I error, although greater between-studies variances and unbalanced category sizes may lead to inflated rates. Conversely, statistical power rates can be lower than desirable unless the difference among category effects and the number of studies are large enough. While the former may vary widely, the number of studies is often below 40 when the influence of a categorical moderator is statistically tested. Therefore, our results remark that most of those analyses might be underpowered.

In conclusion, the results of our simulation study suggest that similar performance can be expected when using a pooled estimate or separate estimates of the residual between-studies variance to test the statistical association of a dichotomous moderator with the effect sizes, as long as there are at least 20 studies and these are roughly balanced across categories. Our results stress the need for a relatively large number of studies for the methods to have enough power to detect small to moderate differences among effect sizes from different subgroups. A pooled estimate will be preferable for most scenarios, unless the residual between-studies variances are clearly different and there are enough studies in each category to get precise separate estimates. Researchers are also encouraged to report the between-studies variance estimate(s) alongside its (their) confidence limits.

4.4.1. Limitations and future research

There are some limitations to this study. First, results can onlybe generalized to the specific manipulated conditions. Although this study was focused on standardized mean differences as the effect size index, our findings may be generalized to other effect size measures which follow an approximately normal distribution. In future simulation studies, it would be advisable to extend the manipulated conditions, for example, using other effect size indices, increasing the number of categories of the moderator and varying the average sample size of each meta-analysis.

In future research, other estimators of the residual between-studies variance could be applied, such as the restricted maximum likehood estimator (Viechtbauer, 2005) and they may also consider alternatives to the normal distribution to generate parametric effects, in order to mimic realistic scenarios more closely.

Finally, the Type I error and statistical power rates yielded by the methods considered in this study were suboptimal for many of the examined conditions. Previous simulation studies have demonstrated that the method proposed by Knapp and Hartung (2003) outperforms the standard method for testing the statistical significance of a continuous moderator (Viechtbauer et al., 2015). It should be interesting to evaluate the performance of this method to test for categorical moderators (see Chapter 5).

Chapter 5

Study 4:

"A comparison of hypothesis tests for categorical moderators in meta-analysis using mixed-effects models"

5.1. Introduction

In the present study, we are still interested in subgroup analyses, which are commonly used to examine the association between categorical moderator variables and the magnitude of the effect size. Based on a subgroup analysis, we can estimate the (average) effect size for each level of the moderator and test for between-group differences. Such analyses may provide valuable insights into circumstances and conditions under which an effect (e.g., the effectiveness of a treatment or intervention) is particularly large or small.

A general recommendation when conducting such moderator analyses is to adopt a mixed-effects model which explicitly models potential 'residual heterogeneity' in the effects, that is, heterogeneity in the true effects not accounted for by the moderator

variable(s) included in the model (Thompson & Higgins, 2002). For models with a categorical moderator, residual heterogeneity simply denotes heterogeneity in the true effects within the various levels of the moderator.

Two approaches can be used to estimate the amount of residual heterogeneity in the context of such models. One is to allow for and estimate a different between-studies variance component within each level of the moderator, while the other consists of assuming a common amount of residual heterogeneity across categories and to calculate a pooled estimate thereof (Borenstein et al., 2009).

Rubio-Aparicio, et al. (2017) recently carried out a simulation study (described in Chapter 4) to compare the statistical performance of the omnibus Wald-type χ^2 test, the Q_B test, for between-group differences in the (average) effect sizes in terms of its Type I error and statistical power rates when the two alternative procedures for estimating the residual between-studies variance (i.e., separate vs. pooled estimation) are used. Results indicated that pooled estimation is preferable for most scenarios, unless the residual between-studies variance is different across categories and the number of studies in each category is large enough to obtain precise separate estimates. However, the Type I error rate of the Q_B test was not nominal for many of the conditions examined, regardless of the approach used in the estimation of the residual between-studies variance. A potential explanation is that the test does not take into account the uncertainty derived from the estimation process of the residual between-studies variance, which typically results in inflated rejection rates under the null hypothesis.

To address that limitation, Hartung, Makambi, and Argaç (2001), and Hartung, Argaç, and Makambi (2002) proposed an alternative method that accounts for the imprecision in the estimated amount of residual heterogeneity in subgroup analyses. This method is based on the same rationale that also underlies the improved method for meta-regression proposed by Knapp and Hartung (2003), which has repeatedly been found to provide adequate control of the Type I error rate in several simulation studies (Huizenga, Visser, & Dolan, 2011; Knapp & Hartung, 2003; Sidik & Jonkman, 2005; Viechtbauer, et al., 2015) and is routinely recommended nowadays (Gonzalez-Mulé & Aguinis, in press). Nonetheless, the implementation of the alternative method is still relatively uncommon when testing for categorical moderators in contrast with growing popularity of the improved method for continuous moderators.

The purpose of the present study was to examine the Type I error and statistical power rates of the improved method proposed by Hartung et al. (2001) for subgroup analyses under a mixed-effects model, as well as to compare its performance with that of the standard Q_B test. Furthermore, the impact of using a pooled estimate versus separate estimates of the residual between-studies variance on the statistical performance of both tests was also explored.

This study is focused on the performance of the mixed-effects model, already described in Chapter 4 (see section 4.1.1). We now present the hypothesis tests and residual heterogeneity estimators that we examined in this study, in the context of the mixed-effects model. Then, the methods and results from a Monte Carlo simulation study comparing the performance of the different procedures are detailed. Last, a discussion of the main results and implications arising from them is provided.

5.1.1. Tests of between-groups differences

The statistical association of a categorical moderator with the effect sizes can be tested by means of a standard Wald-type χ^2 test, the Q_B test. The computation of this statistic can be found in the section 4.1.2 in Chapter 4.

An alternative method to test the statistical significance of a categorical moderator is computed with (Hartung et al., 2001)

$$F = \frac{\frac{Q_B}{m-1}}{\frac{Q_W}{k-m}}, \tag{5.1}$$

where $Q_W = \sum_j Q_{w_j}$ and

$$Q_{w_j} = \sum_{i=1}^{k_j} \widehat{w}_{ij}\left(T_{ij} - \bar{T}_j\right)^2. \tag{5.2}$$

Under the null hypothesis of no difference between the mean effect sizes across categories ($H_0: \mu_{\theta_1} = \mu_{\theta_2} = \cdots = \mu_{\theta_m}$), the *F* statistic follows asymptotically an *F* distribution with (m – 1) and (k – m) degrees of freedom. The equivalence between this *F* statistic for subgroup analyses and the method proposed by Knapp and Hartung (2003) for meta-regression is shown in Appendix 5A.

5.1.2. Estimating the residual between-studies variance

Most of methods proposed to estimate the between-studies variance in the context of the random-effects model estimators have also been extended to the mixed-effects model, and we selected three methods that are commonly implemented and have been found to perform adequately in previous simulation studies (López-López et al., 2014; Veroniki et al., 2016). Concretely, DerSimonian and Laird (DL) estimator, Restricted Maximum Likelihood (REML) estimator and Paule and Mandel (PM) estimator were used in the present study. In Chapter 4 (see section 4.1.3) we described the DL estimator using both separate estimates and pooled estimate of the residual between-studies variance. In this section, we describe the other two estimators (REML and PM) and their computation using both separate estimates and pooled estimate of the residual between-studies variance.

Restricted Maximum Likelihood (REML) estimator

The second method for estimating the residual between-studies variance is based on restricted maximium likelihood estimation. The REML estimator for the *j*th category of the moderator can be obtained iteratively from

$$\hat{\tau}_j^2(REML) = \frac{\sum_i \widehat{w}_{ij}^2 \left[\left(T_{ij} - \bar{T}_j \right)^2 - \hat{\sigma}_{ij}^2 \right]}{\sum_i \widehat{w}_{ij}^2} + \frac{1}{\sum_i \widehat{w}_{ij}} \qquad (5.3)$$

by first computing the right-hand side using initial values for the weights (e.g., by setting $\hat{\tau}_j^2$ in $\widehat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_j^2)$ equal to the estimate obtained using the non-iterative

DL estimator), then updating the weights (and hence also $\bar{T}_j$) using the estimate of $\hat{\tau}_j^2$ obtained, and then iterating this process until convergence. Should $\hat{\tau}_j^2$ ever become negative during this process, the estimate is truncated to zero.

The pooled REML estimate of the residual variance is again computed iteratively, but now using

$$\hat{\tau}_+^2(REML) = \frac{\sum_j \sum_i \hat{w}_{ij}^2 \left[\left(T_{ij} - \bar{T}_j\right)^2 - \hat{\sigma}_{ij}^2 \right]}{\sum_j \sum_i \hat{w}_{ij}^2} + \frac{m}{\sum_j \sum_i \hat{w}_{ij}}, \tag{5.4}$$

with weights $\hat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_+^2)$.

Paule and Mandel (PM) estimator

The third estimator that we included in our simulation study was proposed by Paule and Mandel (1982). The PM estimate for the *j*th category is given by the solution to

$$\sum_i \hat{w}_{ij}\left(T_{ij} - \bar{T}_j\right)^2 - \left(k_j - 1\right) = 0. \tag{5.5}$$

The left-hand side of Eq. 5.5 is a monotonically decreasing function of $\hat{\tau}_j^2$ and can be easily solved for 0 using any standard root finding algorithm. We denote the resulting estimate with $\hat{\tau}_j^2(PM)$. Should Eq. 5.5 be negative for $\hat{\tau}_j^2 = 0$, then the estimate is truncated to zero.

To obtain the pooled estimate for the PM estimator, $\hat{\tau}_+^2(PM)$, we must solve

$$\sum_j \sum_i \hat{w}_{ij}\left(T_{ij} - \bar{T}_j\right)^2 - \sum_j \left(k_j - 1\right) = 0, \tag{5.6}$$

with weights $\hat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_+^2)$.

5.2. Method of the Simulation Study

In the previous section, we presented two methods for testing the statistical significance of a categorical moderator (i.e., the Q_B and F tests) and three methods (i.e., the DL, REML, and PM estimators) which can be used to obtain either a pooled estimate or separates estimates for the residual between-studies variance. This yields 12 different ways of testing the statistical significance of a categorical moderator in a mixed-effects model subgroup analysis, namely the $Q_{B(S)}$ test using separate estimates of the heterogeneity variance combined with either the DL, REML, or PM estimator ($Q_{B(S)_{DL}}$, $Q_{B(S)_{REML}}$, and $Q_{B(S)_{PM}}$, respectively), the $Q_{B(P)}$ test when using a pooled estimate using either the DL, REML, or PM estimator ($Q_{B(P)_{DL}}$, $Q_{B(P)_{REML}}$, and $Q_{B(P)_{PM}}$, respectively), the $F_{(S)}$ test using separate estimates ($F_{(S)_{DL}}$, $F_{(S)_{REML}}$, and $F_{(S)_{PM}}$, respectively), and the $F_{(P)}$ test when using a pooled estimate ($F_{(P)_{DL}}$, $F_{(P)_{REML}}$, and $F_{(P)_{PM}}$, respectively). To compare the performance of these methods, we conducted a Monte Carlo simulation study programmed in R using the *metafor* package (Viechtbauer, 2010). Appendix 5B contains the full R code of our simulation study.

Meta-analyses of k studies were simulated with the standardized mean difference as the effect size index. Each individual study included in a meta-analysis compared two groups (experimental and control) with respect to some continuous outcome. For a given study, values of the outcome were sampled from normal distributions with equal variances (i.e., $N(\mu_E, \sigma^2)$ and $N(\mu_C, \sigma^2)$). For each study, the population standardized mean difference, θ, was defined as (Hedges & Olkin, 1985)

$$\theta = \frac{\mu_E - \mu_c}{\sigma}. \tag{5.7}$$

Without loss of generality, the normal distributions of the experimental and control populations were defined as $N(\theta, 1)$ and $N(0,1)$, respectively.

The effect size was estimated by means of the nearly unbiased estimator proposed by Hedges and Olkin (1985, p. 81)

$$\hat{\theta} = c(m)\frac{\bar{y}_E - \bar{y}_C}{s}, \tag{5.8}$$

where $\bar{y}_E$ and $\bar{y}_C$ are the sample means of the experimental and control groups, s is the pooled standard deviation computed with

$$s = \sqrt{\frac{(n_E - 1)s_E^2 + (n_C - 1)s_C^2}{n_E + n_C - 2}}, \tag{5.9}$$

n_E and n_C being the experimental and control group sample sizes, respectively, s_E^2 and s_C^2 the variances of the two groups, and $c(m)$ is a correction factor for small sample sizes given by

$$c(m) = 1 - \frac{3}{4N - 9}, \tag{5.10}$$

where $N = n_E + n_C$. The estimated within-study variance of $\hat{\theta}$, assuming equal variances and normality within each study, is given by

$$\hat{\sigma}^2 = \frac{n_E + n_C}{n_E n_C} + \frac{\hat{\theta}^2}{2(n_E + n_C)}. \tag{5.11}$$

The k studies were assumed to fall into two categories (with k_1 and k_2 studies in each group). The true standardized mean differences within each subgroup were simulated from $N(\mu_{\theta_j}, \tau_j^2)$ according to a mixed-effects model.

A systematic review of 50 meta-analyses on the efficacy of psychological interventions (Rubio-Aparicio, et al., in press) enabled us to identify a range of representative values for our simulation. We set the number of studies, *k*, to values of 12, 20, 40, and 60. Moreover, we manipulated how *k* was distributed within each category of the moderator, so that in some conditions there was a balanced distribution (i.e., $k_1 = k_2$), while in the remaining conditions there was an unbalanced distribution (i.e., $k_1 \neq k_2$) between the two categories with the second category containing three times as many studies as the first category. For instance, when *k* = 12 we set $k_1 = k_2 = 6$ in the balanced conditions, and $k_1 = 3$ and $k_2 = 9$ in the unequal conditions.

Furthermore, τ_j^2 was manipulated in two different ways. First, we considered three values for this parameter, 0.08, 0.16, and 0.32. Second, we simulated a set of scenarios with homoscedastic variances across categories ($\tau_1^2 = \tau_2^2$), as well as another set of heteroscedastic conditions, with pairs of values $\tau_1^2 = 0.08$ and $\tau_1^2 = 0.16$, $\tau_1^2 =$

0.16 and $\tau_1^2 = 0.08$, $\tau_1^2 = 0.08$ and $\tau_1^2 = 0.32$, $\tau_1^2 = 0.32$ and $\tau_1^2 = 0.08$, $\tau_1^2 = 0.16$ and $\tau_1^2 = 0.32$, and $\tau_1^2 = 0.32$ and $\tau_1^2 = 0.16$.

The average total sample size of the individual studies $\bar{N}$ was set to 20, 40, 60, and 80. The data in the primary studies were simulated assuming $n_E = n_C$. A χ^2 distribution with 4 degrees of freedom was used, so that the skewness of the distribution was +1.414. In addition, values equal to 16, 36, 56 or 76 were added to get the desired average value.

The mean effect size of each category of the moderator was also manipulated. In some conditions the two parametric mean effects were both equal to 0.5 ($\mu_{\theta_1} = \mu_{\theta_2} = 0.5$), whereas for other conditions they were set to different values: $\mu_{\theta_1} = 0.5$ and $\mu_{\theta_2} = 0.3$, $\mu_{\theta_1} = 0.5$ and $\mu_{\theta_2} = 0.1$, and $\mu_{\theta_1} = 0.7$ and $\mu_{\theta_2} = 0.1$. The conditions with equal mean effect sizes across categories allowed us to study the Type I error rate, whereas the conditions with different mean effect sizes enabled us to assess the statistical power.

To assess the Type I error rate, the total number of conditions was: 4 (number of studies) × 2 (balanced-unbalanced number of studies in the two categories) × 4 (average total sample size) × 9 (residual between-studies variance) = 288. With respect to the statistical power, 288 × 3 = 864 conditions examined. Overall, the total number of conditions was therefore 1,152 and for each condition we generated 10,000 replications. Thus, 11,520,000 meta-analyses were simulated. The 12 methods ($Q_{B(S)_{DL}}$, $Q_{B(S)_{REML}}$, $Q_{B(S)_{PM}}$, $Q_{B(P)_{DL}}$, $Q_{B(P)_{REML}}$, $Q_{B(P)_{PM}}$, $F_{(S)_{DL}}$, $F_{(S)_{REML}}$, $F_{(S)_{PM}}$, $F_{(P)_{DL}}$, $F_{(P)_{REML}}$, and $F_{(P)_{PM}}$) were applied to each one of these replications. In each of the 1,152 conditions of our simulation study, the proportion of rejections of the null hypothesis of equality of the mean effect sizes across categories of the moderator was examined.

5.3. Results

In this section, we describe and compare the performance of the methods under the simulated conditions. For brevity, we only present the results for the PM estimator since the pattern of results was very similar for the remaining estimators. Nevertheless, Appendix 5C presents figures using the DL and REML estimators, and the full set of results is available in the Open Science Framework (https://osf.io/6ubxz/). This section is

divided into two parts, corresponding to the Type I error and statistical power rates, respectively.

5.3.1. Type I error rate

Setting $\mu_{\theta_1} = \mu_{\theta_2} = 0.5$ allowed comparing the methods in terms of their Type I error rates. Figures in this section include dashed horizontal lines delimiting the range of values that can be considered as equivalent to the nominal significance level of 5%, after accounting for Monte Carlo error [.0543; .0457]. Therefore, empirical rejection rates within this interval indicate adequate control of the Type I error rate.

Figure 5.1 shows the average Type I error rates as a function of the number of studies, balanced and unbalanced distribution of number of studies within each category of the moderator, average sample size per study, and the amount of residual heterogeneity, in scenarios with homoscedastic residual between-studies variances across the categories of the moderator. As *k* increased (Figure 5.1A), the proportion of rejections of the null hypothesis of equality for $Q_{B(S)}$, $Q_{B(P)}$, and $F_{(S)}$, converged to the nominal significance level, whereas $F_{(P)}$ showed nominal levels regardless of the number of studies. Focusing on the balanced versus unbalanced distribution of the number of studies across categories (Figure 5.1B), $Q_{B(P)}$ and $F_{(P)}$ were not influenced by this factor, whereas $Q_{B(S)}$ and $F_{(S)}$ showed higher empirical rejection rates (above .05) when the number of studies was unbalanced across categories. Last, sample size and the amount of residual heterogeneity did not seem to have a strong influence on the rejection rates (Figures 5.1C and 5.1D), with $F_{(P)}$ consistently yielding the best control of the Type I error rate.

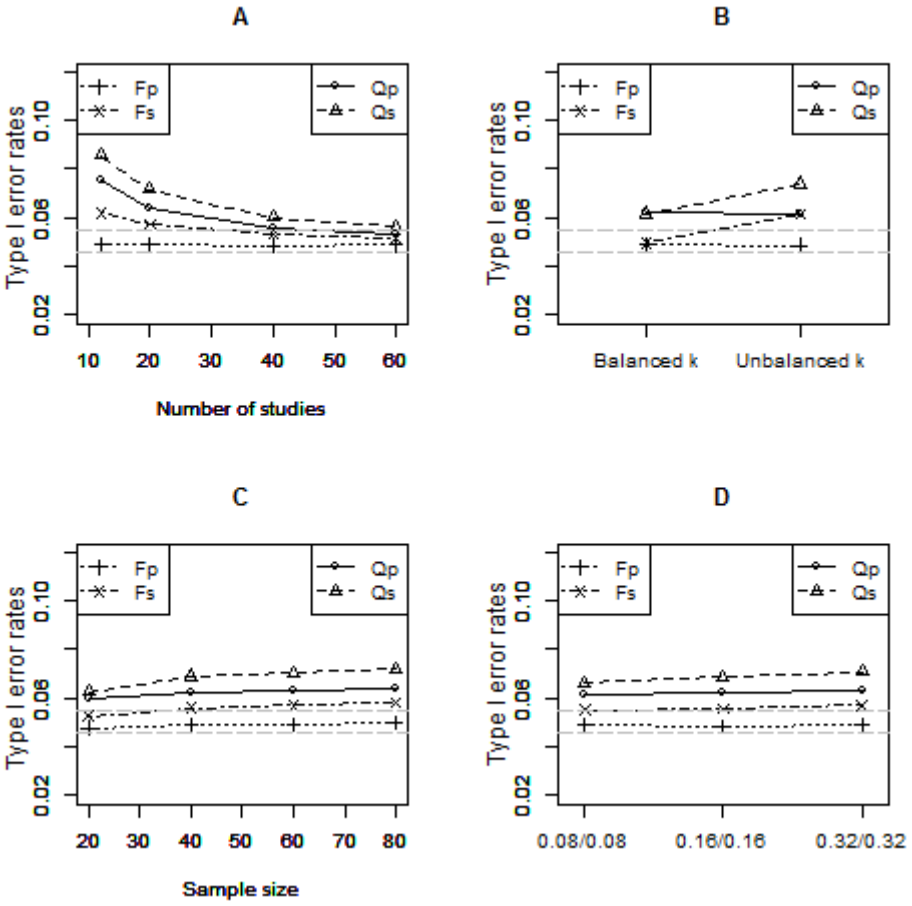


Fig. 5.1. Average Type I error rates in scenarios with homoscedastic residual between-studies variances across the categories of the moderator.

Figure 5.2 presents the average Type I error rates in conditions where the residual between-studies variances were heteroscedastic across the categories of the moderator, and the category with less studies had the smaller variance. The influence of the number of studies (Figure 5.2A) was more pronounced for the Q_B test, with lower Type I error rates as *k* increased, and $Q_{B(S)}$ showing inflated rates with less than 40 studies. The *F* test was less affected, with $F_{(S)}$ showing an adequate control and $F_{(P)}$ yielding overly conservative results, regardless of the number of studies. Regarding the distribution of the number of studies (Figure 5.2B), $Q_{B(S)}$ and $F_{(S)}$ were not influenced by this factor, whereas $Q_{B(P)}$ and $F_{(P)}$ showed error rates below .05 under unbalanced distribution of the number of studies. Furthermore, results did not show important variations as a function of the average sample size and the amount of residual heterogeneity (Figures 5.2C and

5.2D), with $F_{(S)}$ and $Q_{B(P)}$ leading to a good adjustment to the nominal level on average, $F_{(P)}$ yielding over-conservative results, and $Q_{B(S)}$ showing inflated Type I error rates.

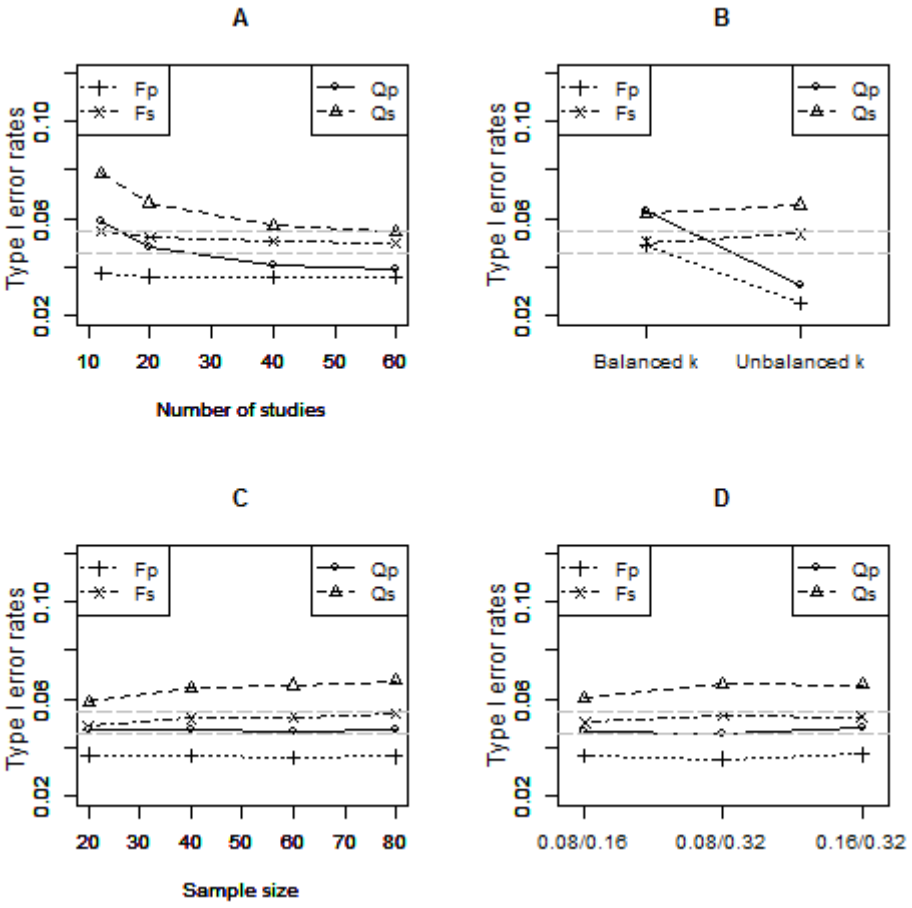


Fig. 5.2. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the categories of the moderator and smaller variance in the smaller category.

Figure 5.3 shows the average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the categories of the moderator and larger variance for the category with less studies. When looking at the results as a function of the number of studies (Figure 5.3A), the rejection rates generally fell above the nominal significance level, with accurate rates provided only by $Q_{B(S)}$ and $F_{(S)}$ with at least 60 and 40 studies, respectively. Regarding the distribution of the number of studies in each category of the moderator, only $F_{(P)}$ and $F_{(S)}$ achieved good adjustment when the number

of studies was balanced across categories, with inflated Type I error rates for all methods in the unbalanced scenarios. The influence of the average sample size and the amount of residual heterogeneity were relatively minor (Figures 5.3C and 5.3D), and all methods yielded rejection rates that were too liberal. The $F_{(S)}$ test consistently provided the closest performance to the nominal significance level.

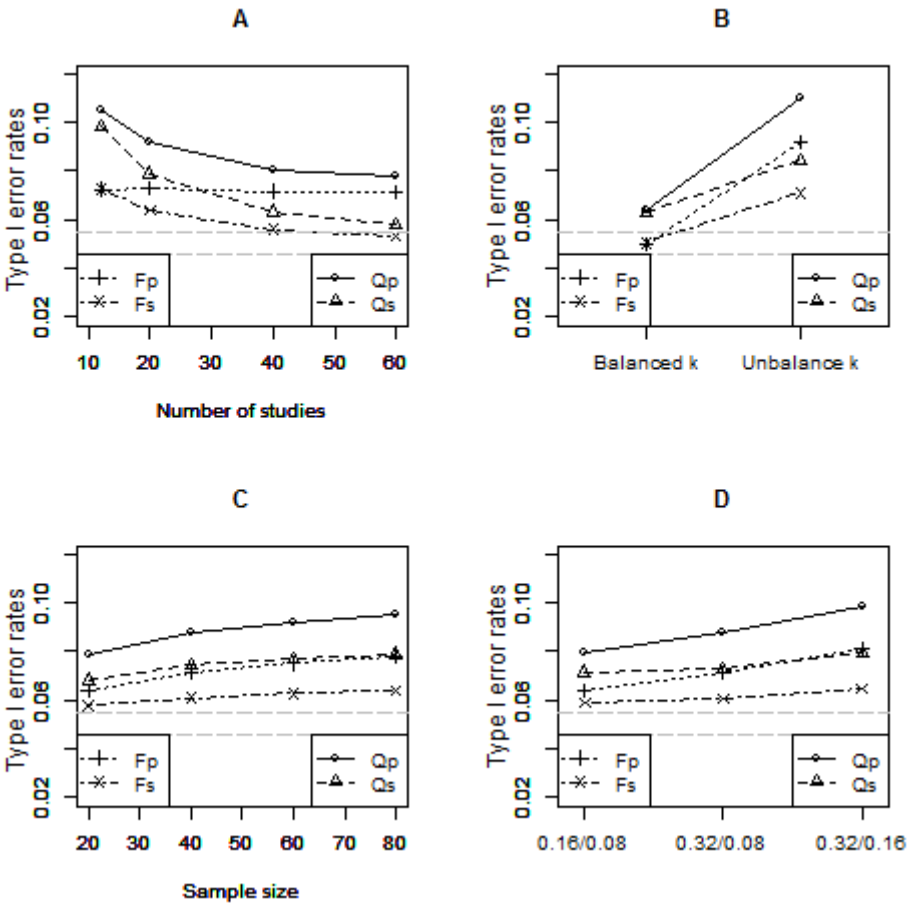


Fig. 5.3. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances in each category of the moderator and larger variance in the smaller category.

5.3.2. Statistical Power

Statistical power reflects the probability of a method rejecting the null hypothesis that is in fact false (i.e., $\mu_{\theta_1} = 0.5$ and $\mu_{\theta_2} = 0.3$, $\mu_{\theta_1} = 0.5$ and $\mu_{\theta_2} = 0.1$, and $\mu_{\theta_1} =$

0.7 and $\mu_{\theta_2} = 0.1$ in our simulation study). In general, power rates equal to or greater than 0.8 are often considered as acceptable in field of psychology (Cohen, 1988).

Figure 5.4 presents the average power rates in scenarios with homoscedastic residual between-studies variances across the categories of the moderator. First, the influence of the different conditions manipulated was equivalent for $Q_{B(S)}$, $Q_{B(P)}$, $F_{(S)}$, and $F_{(P)}$ and, in most conditions, yielding statistical power below 0.8. As expected, for all methods, power increased as the number of studies (Figure 5.4A) and the magnitude of the difference between the mean effect sizes of the two categories (Figure 5.4E) increased, with at least 60 studies and a difference between the mean effect sizes equal to 0.6 ($\mu_{\theta_1} = 0.7$ and $\mu_{\theta_2} = 0.1$) being needed for the methods to provide power rates close to 0.8. Furthermore, larger residual heterogeneity resulted in lower power rates (Figure 5.4D), whereas the distribution of the number of studies across categories (Figure 5.4B) and the average sample size per study (Figure 5.4C) did not show a substantial impact on the power rates of the methods under assessment. The Q_B test yielded slightly higher power rates than the F test across all manipulated conditions.

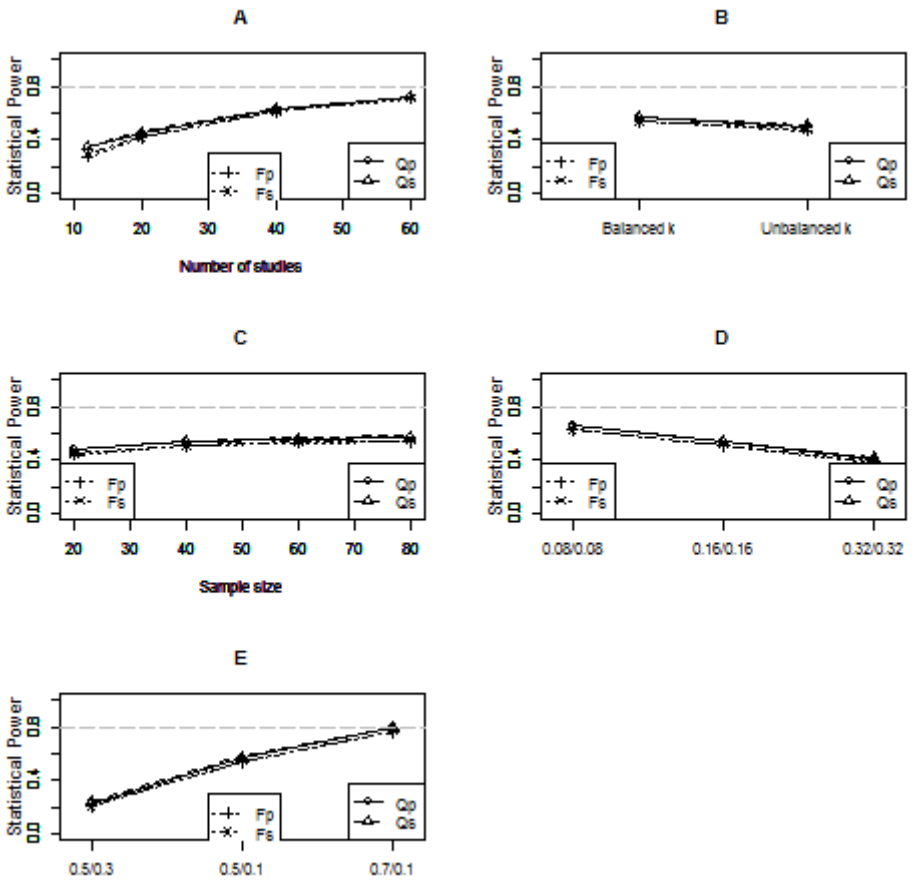


Fig. 5.4. Average power rates in scenarios with homoscedastic residual between-studies variances across the categories of the moderator.

Figures 5.5 and 5.6 present the average power rates in scenarios were the residual between-studies variances were heteroscedastic across the categories of the moderator, with the largest variance either falling in the category with more (Figure 5.5) or with less studies (Figure 5.6). The influence of the different conditions manipulated on the power rates of $Q_{B(S)}$, $Q_{B(P)}$, $F_{(S)}$, and $F_{(P)}$ was very similar to those under homoscedastic residual between-studies variances (see Figure 5.4), with larger k and larger differences among the mean effects leading to higher power rates. It is worth noting the effect of the residual between-studies variance on the power rates. On the one hand, when the category with less studies had less heterogeneous effect sizes (Figure 5.5D), $Q_{B(S)}$, $Q_{B(P)}$, $F_{(S)}$, and $F_{(P)}$ yielded power rates relatively higher under the condition of $\tau_1^2 = 0.08$ and $\tau_1^2 = 0.32$. On the other hand, when the category with less studies was more heterogeneous (Figure

5.6D), power rates for all of methods were slightly higher under the condition of $\tau_1^2 = 0.16$ and $\tau_1^2 = 0.08$.

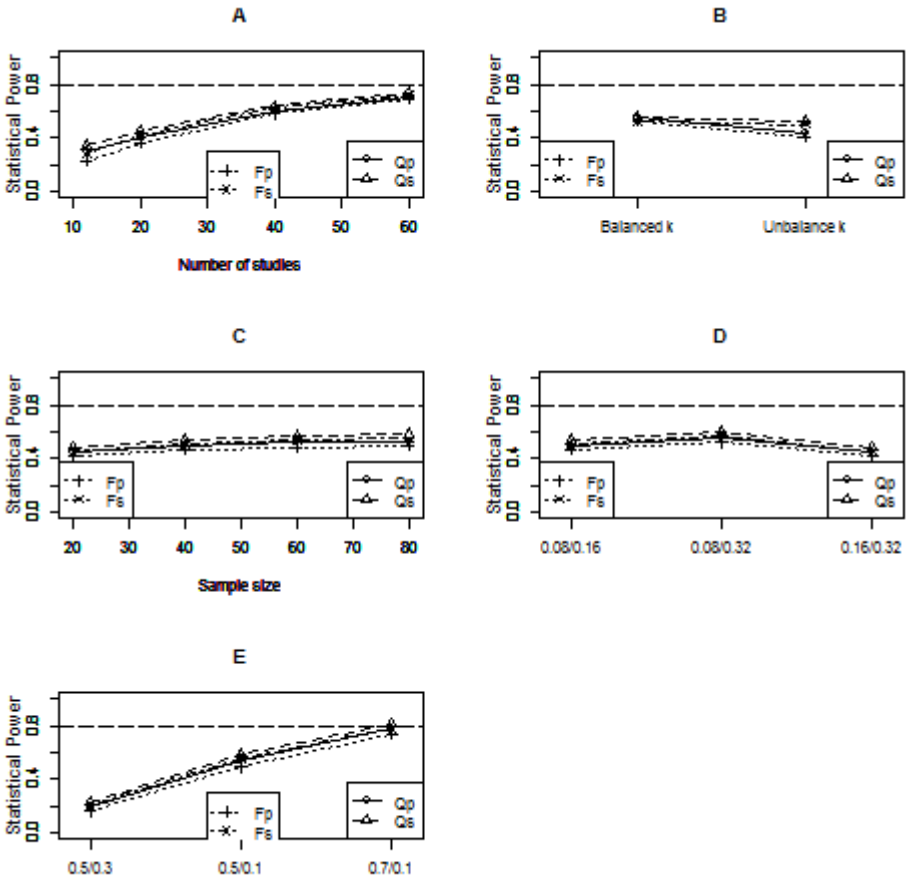


Fig. 5.5. Average power rates in scenarios with heteroscedastic residual between-studies variances across the categories of the moderator and smaller variance in the smaller category.

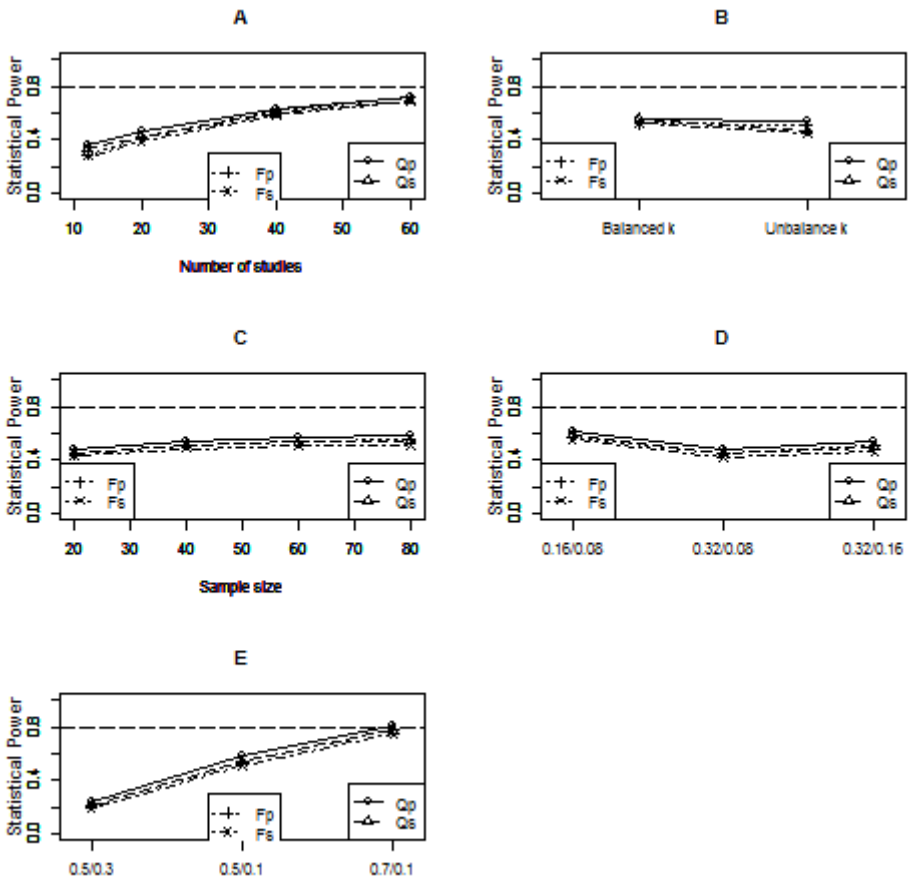


Fig. 5.6. Average power rates in scenarios with heteroscedastic residual between-studies variances across the categories of the moderator and larger variance in the smaller category.

5.4. Discussion

This study compared a variety of methods in the context of subgroup analyses using mixed-effects models. Specifically, two methods for testing the statistical significance of the categorical moderator (i.e., the Q_B and F tests), two procedures for estimating the residual between-studies variance (pooled or separates estimates), and three residual heterogeneity variance estimators (DL, REML, and PM) were combined to provide twelve analysis approaches that were examined in a Monte Carlo simulation study, with standardized mean differences as the effect size measure. Two comparative criteria, empirical Type I error and statistical power rates, were considered for assessing

the adequacy of each method across a wide variety of realistic scenarios in clinical psychology.

Results were not found to be affected by the residual between-studies variance estimator used. However, some notable differences were observed depending on the method employed for testing the statistical association of a categorical moderator and on the approach implemented to estimate the amount of residual heterogeneity in each category (pooled versus separate estimates).

Some authors have criticized that the standard random-effects method does not take into account the uncertainty derived from the variance estimation process, which can lead to wrong statistical conclusions (e.g., Thompson & Higgins, 2002). This led to the development of improved hypothesis tests by Hartung and colleagues in the context of random-effects meta-analysis (Hartung, 1999) and mixed-effects meta-regression (Knapp & Hartung, 2003). These tests are known to outperform the standard methods in terms of their control of the Type I error rate (Huizenga et al., 2011; Sánchez-Meca & Marín-Martínez, 2008; Sidik & Jonkman, 2005; Viechtbauer et al., 2015) and are recommended for routine use nowadays. Hartung and colleagues (2001) also proposed an improved method for subgroup analyses using mixed-effects models using an F test, and we examined its performance compared to the typically implemented Q_B test. The empirical Type I error rates obtained by both methods suggest that the improved F test has clear advantages over the standard Q_B test. This leads us to encourage meta-analysts to apply the F test instead of the standard Q_B test in subgroup analyses.

The F test for subgroup analyses can be considered to be a special case of the improved method for meta-regression. In the meta-regression context, Knapp and Hartung (2003) proposed a multiplicative adjustment factor for the estimated variances of the model coefficients, and suggested to truncate this factor to one if a smaller value was obtained, in order to minimize false positive findings. Several pieces of meta-analytic software currently incorporate such truncation, including Comprehensive Meta-Analysis 3.0 (Borenstein et al., 2014) and the *metareg* macro for Stata (Harbord & Higgins, 2008), whereas other alternatives like the *metafor* package for R (Viechtbauer, 2010) use the untruncated factor by default. This adjustment factor is equal to the denominator of the F formula (see Appendix 5A), hence implementing the truncation in the context of a subgroup analysis would be straightforward. However, Viechtbauer and colleagues

(2015) showed that the improved method for meta-regression provides an adequate adjustment of the nominal significance level without truncating, whereas overly conservative results may be obtained if the truncation is applied. Consequently, in the present study, we allowed the denominator of the *F* test to be smaller than one, and we generally would recommend this version of the test.

When comparing the performance of the $F_{(P)}$ and $F_{(S)}$ tests, under homoscedastic variances across categories, $F_{(P)}$ yielded the best control of the Type I error rates, regardless of how the number of studies was distributed across the categories of the moderator. Under heteroscedastic variances across categories, both $F_{(P)}$ and $F_{(S)}$ achieved adequate performance as long as the number of studies was distributed equally across categories. However, under an unbalanced distribution of the number of studies, the practical consequences of allowing for heteroscedastic residual between-studies variances were more evident. Concretely, the $F_{(S)}$ showed good adjustment when the value of the smallest residual between-studies variance was associated with the category with the smallest number of studies (see Figure 5.2), whereas when the value of the largest residual between-studies variance was associated with the category with the smallest number of studies, both tests showed a poor adjustment to the nominal level (see Figure 5.3).

These results allow us to recommend the use of the $F_{(P)}$ test in most conditions, except when the meta-analyst suspects that the true value of τ^2_{res} may vary across categories and the number of studies across categories is unbalanced. In that case, the $F_{(S)}$ test showed the best performance. Note that using a pooled estimate would be expected to provide more accurate results for most scenarios, as the estimate is then based on a larger number of studies. This can be particularly important if the total number of studies is small (e.g., $k < 20$), which has been found to be the case for most Cochrane Reviews (Davey et al., 2011).

The statistical power of all methods was lower than .80 in most conditions, unless the magnitude of the difference between the mean effects across categories was equal to 0.6. As expected, statistical power rates increased with a larger number of studies, yielding rates close to .80 with at least 60 studies (see Figures 5.4, 5.5, and 5.6). Note that the differences in the statistical power rates for the methods may also be caused by either inflated or overly conservative Type I error rates.

In summary, results of our simulation study suggest that out of the different alternatives considered in the present study, the improved *F* test computed using a pooled estimate is the most suitable option to test the statistical association between a categorical moderator and the effect sizes in most conditions. Nevertheless, if the meta-analyst suspects that the residual between-studies variances are heteroscedastic across categories of the moderator and the number of studies is unbalanced across categories, then the *F* test using separate estimations of the residual between-studies variance may be preferable.

The present simulation study was conducted with standardized mean differences, but its results may be generalized to other effect size measures with (asymptotically) normal sampling distributions. Our results are limited to the manipulated conditions. Although the values for the parameters were chosen following a systematic review of 50 meta-analyses on the efficacy of psychological interventions (Rubio-Aparicio et al., in press) to represent realistic conditions, additional simulation studies are needed to assess the performance of the methods under more adverse conditions, such as a non-normal distribution for the true effects within each category of the moderator.

Lastly, an important limitation in this field is that the meta-analyst cannot determine whether the residual between-studies variances are homoscedastic or heteroscedastic across categories, as the parameters are unknown. In the absence of a formal statistic to test the homoscedasticity of the residual between-studies variances across categories, it is possible to compare the model fit using separate or pooled estimates.

Chapter 6

Conclusions

Meta-analytic methodology allows quantitative integration of the results from a set of primary studies focused on a common topic, by the application of statistical methods (Borenstein et al., 2009; Botella & Gambara, 2006; Cooper et al., 2009; Lipsey & Wilson, 2001). The advantages of meta-analysis are numerous. Nowadays, most conclusions about cumulative knowledge in psychology and in other research areas are based on meta-analysis. For instance, applied professionals can make decisions based on the results extracted from a meta-analysis about which therapy is the most effective to treat a certain psychological disorder. Due to the broad scope of meta-analysis, it is really important to achieve valid results for the scientific community, applying the most optimal inferential methods in meta-analysis. Several Monte Carlo simulation studies have been developed in order to investigate which techniques and procedures are most adequate given the characteristics of a meta-analytic database.

In this dissertation three Monte Carlo simulation studies comparing techniques and procedures usually applied in meta-analysis were carried out. The first simulation study (Chapter 3) was focused on computing an average effect size (first inferential objective in any meta-analysis) under a random-effects model, and the other two simulation studies (Chapters 4 and 5) were focused on subgroup analyses (third objective in any meta-analysis) under mixed-effects models. As a previous step, Chapter 2 presented a methodological review of meta-analyses about the effectiveness of clinical psychology treatments, intended to help us design the scenarios for our simulation studies. In the following paragraphs, we summarize the principal conclusions extracted from the empirical part of this dissertation.

Chapter 2 presents a methodological review of 54 meta-analyses of the effectiveness of clinical psychology treatments, using standardized mean differences as the effect size index. We analysed the distribution of the number of studies of the meta-analyses, the distribution of the sample sizes in the studies of each meta-analysis, the distribution of the effect sizes in each of the meta-analyses, the distribution of the between-studies variance values, and the Pearson correlations between effect size and sample size in each meta-analysis. Results were presented as a function of the type of standardized mean difference: posttest standardized mean difference, standardized mean change from prettest to postest, and standardized mean change difference between groups. The first interesting finding was that most meta-analyses used the standardized mean difference computed from the posttest scores to compare experimental and control groups, and although the best option to compare two groups is the standardized mean change difference, this index was scarcely used. On the other hand, results suggested the existence of a relatively low number of studies across most meta-analysis, with large heterogeneity among the effect sizes, violation of the normality assumption in the distribution of the effect sizes, and positive and negative correlation values between effect sizes and sample sizes. Finally, we found that the three quartiles of the mean effect size distribution for the meta-analyses using both the standardized mean difference computed from posttest scores and the standardized mean change differences were similar to those advocated by Cohen (1988) as indicating low, medium and large effect magnitudes; conversely, the three quartiles of the mean effect size distribution for the meta-analyses using the standardized mean changes from pretest to postest were larger than those in Cohen´s (1988) proposal. In sum, the analysis of the distribution of the mean effect sizes through the meta-analyses provides a specific and contextualized guide for the interpretation of the clinical significance of the different types of standardized mean differences within the field of evaluation of clinical psychological interventions.

The first simulation study, described in Chapter 3, aimed to compare the performance of various random-effects meta-analytic methods (standard method, Hartung´s method, profile likelihood method and non-parametric bootstrapping) for computing an average effect size and a confidence interval (CI) around it when the normality assumption of the parametric effects distribution is not met. Three estimators of the heterogeneity variance (DerSimonian and Laird, restricted maximum likelihood, and empirical Bayes) were considered. For comparison purposes, estimates from the

fixed-effect model were also included. Performance of the methods was evaluated in terms of bias and mean squared error of the average effect estimators, empirical coverage probability and width of the CIs, and bias of the standard error. Results suggested that random-effects methods are robust to departures from the normality, with the Hartung´s method and profile likelihood method yielding the best performance under suboptimal conditions.

The second simulation study, presented in Chapter 4, compared the impact of pooled versus separate estimates of the residual between-studies variance on the statistical performance the of $Q_{B(S)}$ and $Q_{B(P)}$ assuming a mixed-effects model. The residual between-studies variance was estimated using DerSimonian and Laird method. The performance of the methods was evaluated in terms of their Type I error and statistical power rates. Results suggested that similar performance can be expected as long as there are at least 20 studies and these are approximately balanced across categories. Conversely, when subgroups were unbalanced, the practical consequences of having heterogeneous residual between-studies variances were more evident, with both tests leading to the wrong statistical conclusion more often than in the conditions with balanced subgroups. A pooled estimate should be preferred for most scenarios, unless the residual between-studies variances are clearly different and there are enough studies in each category to obtain precise separate estimates.

The last simulation study, explained in Chapter 5, examined both approaches, pooled versus separate estimations for the residual between-studies variance, in combination with two methods to test the statistical significance of the moderator, namely the routinely used the Q_B test and an improved F test, each combined with three different estimators of the residual between-studies variance (DerSimonian and Laird, restricted maximum likelihood, and Paule and Mandel). The performance of the twelve different alternatives was evaluated in terms of their Type I error and statistical power rates. Results suggested that the F test computed using a pooled estimate of the residual between-studies variance across categories was the most suitable option in most conditions, although the F test using separate estimates of the residual between-studies variance might be preferable if the residual heterogeneity variances are heteroscedastic, especially when the number of studies is unbalanced across categories. Results showed the same trends for all estimators of the residual between-studies variance. Our findings provide evidence that the F test has clear advantages over the typically implemented Q_B test, and that choice of

pooled versus separate estimation of the residual between-studies variance should be made after examining the characteristics of the meta-analytic database.

To sum up, taking into account the results of this thesis, the principal recommendations for researchers are:

- The summaries of methodological characteristics across 54 meta-analyses provided in Chapter 2, can be useful to design future Monte Carlo and theoretical studies in clinical psychology and related fields.
- The application of the guidelines provided in Chapter 2 will help interpret correctly and in a contextualized way the magnitude of different types of standardized mean differences within the field of clinical psychology.
- The standardized mean change difference should be used more frequently in meta-analyses that of studies with a two-group design.
- The number of studies has an important impact on the performance of the meta-analytic methods compared. About 20 studies are required to get accurate results.
- The heterogeneity variance estimator does not exert an important influence on the results of the statistical methods examined in this dissertation.
- In general, random-effects methods are robust to violations of the normality assumption of the parametric effects distribution. However, Hartung´s and profile likelihood methods are preferable under suboptimal conditions.
- The F method for testing the statistical significance of a categorical moderator outperforms the standard Q_B test in subgroup analyses.
- A pooled between-studies variance estimate is preferred for most scenarios for subgroup analyses, unless the variances are clearly different across categories and there are enough studies in each category to obtain precise separate estimates

References

(References preceded with an asterisk were those included in the methodological review of meta-analyses from Chapter 2)

*Abramowitz, J. S., Tolin, D. F., & Street, G. P. (2001). Paradoxical effects of thought suppression: A meta-analysis of controlled studies. *Clinical Psychology Review, 21*, 683-703.

*Acarturk, C., Cuijpers, P., Van Straten, A., & De Graaf, R. (2009). Psychological treatment of social anxiety disorder: a meta-analysis. *Psychological Medicine, 39*, 241-254.

Adams, D.C., Gurevitch, J., & Rosenberg, M.S. (1997). Resampling tests for meta-analysis of ecological data. *Ecology, 78*, 1277–1283.

*Aderka, I. M., Nickerson, A., Bøe, H. J., & Hofmann, S. G. (2012). Sudden gains during psychological treatments of anxiety and depression: A meta-analysis. *Journal of Consulting and Clinical Psychology, 80*, 93-101.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*, 257-278.

*Bell, A. C., & D'Zurilla, T. J. (2009). Problem-solving therapy for depression: A meta-analysis. *Clinical Psychology Review, 29*, 348-353.

*Benish, S. G., Imel, Z. E., & Wampold, B. E. (2008). The relative efficacy of bona fide psychotherapies for treating post-traumatic stress disorder: A meta-analysis of direct comparisons. *Clinical Psychology Review, 28*, 746-758.

Berkey, C.S., Hoaglin, D.C., Mosteller, F., Colditz, G.A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine, 14*, 395–411.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005). *Comprehensive meta-analysis* (Version 2.0) [Computer software]. Englewood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Borenstein, M., Hedges. L. V., Higgins, J. P. T., & Rothstein, H.R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1,* 97–111.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2014). *Comprehensive meta-analysis* (Version 3.0) [Computer software]. Englewood, NJ: Biostat

Borenstein, M., & Higgins, J. P. T. (2013). Meta-analysis and subgroups. *Prevention Science*, *14*, 134– 143.

Botella, J., & Gambara, H. (2002). *¿Qué es el meta-análisis? [Meta-analysis: what is it?]* Madrid, Spain: Biblioteca Nueva.

Botella, J., & Gambara, H. (2006). Doing and reporting a meta-analysis. *International Journal of Clinical and Health Psychology*, *6*, 425-440

Botella, J., & Sánchez-Meca, J. (2015). *Meta-análisis en ciencias sociales y de la salud [Meta-analysis in social and health sciences]*. Madrid, Spain: Síntesis.

Brockwell, S.E., & Gordon, I.R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine, 20,* 825-840.

Brockwell, S.E., & Gordon, I.R. (2007). A simple method for inference on an overall effect in meta-analysis. *Statistics in Medicine, 26,* 4531-4543.

*Burke, B. L., Arkowitz, H., & Menchola, M. (2003). The efficacy of motivational interviewing: a meta-analysis of controlled clinical trials. *Journal of Consulting and Clinical Psychology, 71*, 843-861.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25,* 4279-4292.

Canty, A., & Ripley, B.D. (2012). boot: Bootstrap R (S-PLUS) Functions, URL http://CRAN.R-project.org/package=boot, R package version 1.3- 7.

*Casement, M. D., & Swanson, L. M. (2012). A meta-analysis of imagery rehearsal for post-trauma nightmares: Effects on nightmare frequency, sleep quality, and posttraumatic stress. *Clinical Psychology Review*, *32*, 566-574.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10,* 101–129.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

*Cuijpers, P., Clignet, F., van Meijel, B., van Straten, A., Li, J., & Andersson, G. (2011). Psychological treatment of depression in inpatients: a systematic review and meta-analysis. *Clinical Psychology Review*, *31*, 353-360.

*Cuijpers, P., Donker, T., van Straten, A., Li, J., & Andersson, G. (2010). Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. *Psychological Medicine*, *40*, 1943-1957.

*Cuijpers, P., Driessen, E., Hollon, S. D., van Oppen, P., Barth, J., & Andersson, G. (2012). The efficacy of non-directive supportive therapy for adult depression: a meta-analysis. *Clinical Psychology Review*, *32*, 280-291.

*Cuijpers, P., Li, J., Hofmann, S. G., & Andersson, G. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clinical Psychology Review*, *30*, 768-778.

*Cuijpers, P., van Straten, A., Warmerdam, L., & Andersson, G. (2009). Psychotherapy versus the combination of psychotherapy and pharmacotherapy in the treatment of depression: a meta-analysis. *Depression and Anxiety*, *26*, 279-288.

Davey, J., Turner, R.M., Clarke, M.J., Higgins, J.P.T. (2011). Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: a cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, *11*, 160.

DerSimonian, R., & Laird, N. (1986). Meta-analysis of clinical trials. *Controlled Clinical Trials, 7*, 177–188.

*Dixon, K. E., Keefe, F. J., Scipio, C. D., Perri, L. M., & Abernethy, A. P. (2007). Psychological interventions for arthritis pain management in adults: a meta-analysis. *Health Psychology*, *26*, 241-250.

*Driessen, E., Cuijpers, P., de Maat, S. C., Abbass, A. A., de Jonghe, F., & Dekker, J. J. (2010). The efficacy of short-term psychodynamic psychotherapy for depression: a meta-analysis. *Clinical Psychology Review*, *30*, 25-36.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*, 171-200.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In S. Kotz, N. L. Johnson (Eds), *Breakthroughs in Statistics* (pp. 569-593). New York: Springer.

Efron, B., & Hastie, T. (2016). *Computer Age Statistical Inference*. New York: Cambridge University Press.

*Ekers, D., Richards, D., & Gilbody, S. (2008). A meta-analysis of randomized trials of behavioural treatment of depression. *Psychological Medicine*, *38*, 611-623.

Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I., & Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine*, *19*, 1707-1728.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*, 532-538.

Fleishman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational and Behavioral Statistics*, *18*, 271-279.

Glass, G.V. (1976). Primary, secondary and meta-analysis of research. *Educational Research*, *5*, 3-8.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd ed.). Boston, MA: Allyn and Bacon.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed-effects analysis of variance and covariance. *Review of Educational Research, 42*, 237–288.

Gonzalez-Mulé, E., & Aguinis, H. (in press). Advancing Theory by Assessing Boundary Conditions With Metaregression: A Critical Review and Best-Practice Recommendations. *Journal of Management*.

*Gooding, P., & Tarrier, N. (2009). A systematic review and meta-analysis of cognitive-behavioural interventions to reduce problem gambling: Hedging our bets? *Behaviour Research and Therapy*, *47*, 592-607.

Guolo, A. (2012). Higher-order likelihood inference in meta-analysis and meta-regression. *Statistics in Medicine, 31*, 313-327.

Guolo, A., & Varin, C. (2012). The R package metaLik for likelihood inference in meta-analysis. *Journal of Statistical Software, 50*, 1-14.

*Hanrahan, F., Field, A. P., Jones, F. W., & Davey, G. C. (2013). A meta-analysis of cognitive therapy for worry in generalized anxiety disorder. *Clinical Psychology Review*, *33*, 120-132.

*Hansen, K., Höfling, V., Kröner-Borowik, T., Stangier, U., & Steil, R. (2013). Efficacy of psychological interventions aiming to reduce chronic nightmares: A meta-analysis. *Clinical Psychology Review*, *33*, 146-155.

Harbord, R. M., & Higgins, J.P.T. (2008). Meta-regression in Stata. *The Stata Journal, 8*, 493-519.

Hardy, R.J., & Thompson, S.G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine, 15*, 619-629.

*Harris, A. H. (2006). Does expressive writing reduce health care utilization? A meta-analysis of randomized trials. *Journal of Consulting and Clinical Psychology*, *74*, 243-252.

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*, *41*, 901-916.

Hartung, J., Argaç, D., & Makambi, K. H. (2002). Small sample properties of tests on homogeneity in one-way Anova and Meta-analysis. *Statistical Papers*, *43*, 197-235.

Hartung, J., & Knapp, G. (2001). On tests of the overall treatment effect in the meta-analysis with normally distributed responses. *Statistics in Medicine, 20*, 1771-1782.

Hartung, J., Makambi, K.H., & Argac, D. (2001). An extended ANOVA F-test with application to the h.eterogeneity problem in meta-analysis. *Biometrical Journal, 43*, 135-146.

*Haug, T., Nordgreen, T., Öst, L. G., & Havik, O. E. (2012). Self-help treatment of anxiety disorders: a meta-analysis and meta-regression of effects and potential moderators. *Clinical Psychology Review*, *32*, 425-445.

*Hausenblas, H. A., Campbell, A., Menzel, J. E., Doughty, J., Levine, M., & Thompson, J. K. (2013). Media effects of experimental presentation of the ideal physique on eating disorder symptoms: A meta-analysis of laboratory studies. *Clinical Psychology Review*, *33*, 168-181.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Henmi, M., & Copas, J.B. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine, 29,* 2969-2983.

*Hesser, H., Weise, C., Westin, V. Z., & Andersson, G. (2011). A systematic review and meta-analysis of randomized controlled trials of cognitive–behavioral therapy for tinnitus distress. *Clinical Psychology Review*, *31*, 545-553.

Higgins, J., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*, 557-560.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172-177.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin.

Hoaglin, D.C. (2015). We know less than we should about methods of meta-analysis. *Research Synthesis Methods, 6*, 287-289.

Hoaglin, D.C. (2016). Misunderstandings about *Q* and 'Cochran's *Q* test' in meta-analysis. *Statistics in Medicine, 35*, 485-495.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11*, 193-206.

Huizenga, H.M., Visser, I., & Dolan, C.V. (2011). Testing overall and moderator effects in random effects meta-regression. *British Journal of Mathematical and Statistical Psychology, 64*, 1–19.

IntHout, J., Ioannidis, J.P.A., Borm, G.F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology, 14*, 25. d

*Kalu, U. G., Sexton, C. E., Loo, C. K., & Ebmeier, K. P. (2012). Transcranial direct current stimulation in the treatment of major depression: a meta-analysis. *Psychological Medicine, 42*, 1791-1800.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

*Kleinstäuber, M., Witthöft, M., & Hiller, W. (2011). Efficacy of short-term psychotherapy for multiple medically unexplained physical symptoms: a meta-analysis. *Clinical Psychology Review, 31*, 146-160.

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*, 2693–2710.

Konstantopoulos, S., & Hedges, L. V. (2009). Analyzing effect sizes: Fixed-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 279-293). New York: Russell Sage Foundation.

Kontopantelis, E., & Reeves, D. (2012a). Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A comparison between DerSimonian–Laird and restricted maximum likelihood. *Statistical Methods in Medical Research*, *21*, 657-659.

Kontopantelis, E., & Reeves, D. (2012b). Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Statistical Methods in Medical Research*, *21*, 409–426.

Kulinskaya, E., Dollinger, M. B., & Bjørkestøl, K. (2011). On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta-analysis of risk difference. *Research Synthesis Methods*, *2*, 254-270.

*Lackner, J. M., Mesmer, C., Morley, S., Dowzer, C., & Hamilton, S. (2004). Psychological treatments for irritable bowel syndrome: a systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *72*, 1100-1113.

*Lansbergen, M. M., Kenemans, J. L., & van Engeland, H. (2007). Stroop interference and attention-deficit/hyperactivity disorder: a review and meta-analysis. *Neuropsychology*, *21*, 251-262.

Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs*, *76*, 286-302.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis. *American Psychologist*, *48*, 1181-1209.

Lipsey, M. W, & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

*Lissek, S., Powers, A. S., McClure, E. B., Phelps, E. A., Woldehawariat, G., Grillon, C., & Pine, D. S. (2005). Classical fear conditioning in the anxiety disorders: a meta-analysis. *Behaviour Research and Therapy*, *43*, 1391-1424.

López-López, J. A., Botella, J., Sánchez-Meca, J., & Marín-Martínez, F. (2013). Alternatives for mixed-effects meta-regression models in the reliability generalization approach: A simulation study. *Journal of Educational and Behavioral Statistics*, *38*, 443-469.

López-López, J.A, Marín-Martínez, F., Sánchez-Meca, J., van den Noortgate, W., & 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.

*Lundahl, B., Risser, H. J., & Lovejoy, M. C. (2006). A meta-analysis of parent training: Moderators and follow-up effects. *Clinical Psychology Review*, *26*, 86-104.

*Malouff, J. M., Thorsteinsson, E. B., Rooke, S. E., Bhullar, N., & Schutte, N. S. (2008). Efficacy of cognitive behavioral therapy for chronic fatigue syndrome: a meta-analysis. *Clinical Psychology Review*, *28*, 736-745.

*Malouff, J. M., Thorsteinsson, E. B., & Schutte, N. S. (2007). The efficacy of problem solving therapy in reducing mental and physical health problems: a meta-analysis. *Clinical Psychology Review*, *27*, 46-57.

Marín-Martínez, F., & Sánchez-Meca, J. (2010). Weighting by inverse variance or by sample size in random-effects meta-analysis. *Educational and Psychological Measurement*, *70*, 56-73.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS medicine*, *6*, e1000097.

Morris, C.N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association, 78*, 47–55.

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, *53*, 17-29.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*, 364-386.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105-125.

National Research Council (1992). *Combining information: Statistical issues and opportunities for research.* Washington, DC: National Academy Press.

*Nestoriuc, Y., Rief, W., & Martin, A. (2008). Meta-analysis of biofeedback for tension-type headache: efficacy, specificity, and treatment moderators. *Journal of Consulting and Clinical Psychology*, *76*, 379-396.

*Oldham, M., Kellett, S., Miles, E., & Sheeran, P. (2012). Interventions to increase attendance at psychotherapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology*, *80*, 928-939.

*Opriş, D., Pintea, S., García-Palacios, A., Botella, C., Szamosközi, Ş., & David, D. (2012). Virtual reality exposure therapy in anxiety disorders: a quantitative meta-analysis. *Depression and Anxiety, 29*, 85-93.

Paule, R.C. & Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, *87*, 377–385.

*Pérez-Mañá, C., Castells, X., Vidal, X., Casas, M., & Capellà, D. (2011). Efficacy of indirect dopamine agonists for psychostimulant dependence: a systematic review and meta-analysis of randomized controlled trials. *Journal of Substance Abuse Treatment*, *40*, 109-122.

*Prendergast, M. L., Urada, D., & Podus, D. (2001). Meta-analysis of HIV risk-reduction interventions within drug abuse treatment programs. *Journal of Consulting and Clinical Psychology*, *69*, 389-405.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295-315). New York: Russell Sage Foundation.

Review Manager (2014). *RevMan* (Version 5.3) [Computer software]. Copenhagen, Denmark: The Nordic Cochrane Centre, The Cochrane Collaboration.

*Richards, D., & Richardson, T. (2012). Computer-based psychological treatments for depression: a systematic review and meta-analysis. *Clinical Psychology Review, 32*, 329-342.

*Roberts, M. E., Tchanturia, K., Stahl, D., Southgate, L., & Treasure, J. (2007). A systematic review and meta-analysis of set-shifting ability in eating disorders. *Psychological Medicine, 37*, 1075-1084.

*Rodenburg, R., Benjamin, A., de Roos, C., Meijer, A. M., & Stams, G. J. (2009). Efficacy of EMDR in children: A meta-analysis. *Clinical Psychology Review, 29*, 599-606.

*Rosa-Alcázar, A. I., Sánchez-Meca, J., Gómez-Conesa, A., & Marín-Martínez, F. (2008). Psychological treatment of obsessive–compulsive disorder: A meta-analysis. *Clinical Psychology Review, 28*, 1310-1325.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.) Newbury Park, CA: Sage.

Rubio-Aparicio, M., Marín-Martínez, F., Sánchez-Meca, J., & López-López, J.A. (in press). A methodological review of meta-analyses about the effectiveness of clinical psychology treatments. *Behavior Research Methods*.

Rubio-Aparicio, M., Sánchez-Meca, J., López-López, J.A., Marín-Martínez, F., & Botella, J. (2017). Analysis of categorical moderators in mixed-effects meta-analysis: Consequences of using pooled vs. separate estimates of the residual between-studies variances. *British Journal of Mathematical and Statistical Psychology, 70*, 439–456.

Sánchez-Meca, J., & Botella, J. (2010). Revisiones sistemáticas y meta-análisis: Herramientas para la práctica profesional [Systematic reviews and meta-analysis: Tools for practitioners]. *Papeles del Psicólogo, 31*, 7-17.

Sánchez-Meca, J., López-López, J.A., & 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.

Sánchez-Meca, J., & Marín-Martínez, F. (1997). Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type I error. *Quality and Quantity*, *31*, 385-399.

Sánchez-Meca J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods*, *13*, 31–48.

Sánchez-Meca, J., & Marín-Martínez, F. (2010). Meta-analysis. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., Vol. 7, pp. 274–282). Oxford, UK: Elsevier.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods, 8*, 448-467.

*Sánchez-Meca, J., Rosa-Alcázar, A. I., Marín-Martínez, F., & Gómez-Conesa, A. (2010). Psychological treatment of panic disorder with or without agoraphobia: a meta-analysis. *Clinical Psychology Review, 30*, 37-50.

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology, 62*, 97-128.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge, MA: Hogrefe & Huber.

Senn, S. (2008). The t-test tool. *Significance*, *5*, 40–41.

*Shadish, W. R., & Baldwin, S. A. (2005). Effects of behavioral marital therapy: a meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology, 73*, 6-14.

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

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., ... & Bouter, L. M. (2007). Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, *7*, 10.

Sidik, K., & Jonkman, J.N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine, 21*, 3153-3159.

Sidik, K., & Jonkman, J.N. (2005). A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics, 15,* 823-838.

Sidik, K., & Jonkman, J.N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine, 26*, 1964-1981.

*Smit, Y., Huibers, M. J., Ioannidis, J. P., van Dyck, R., van Tilburg, W., & Arntz, A. (2012). The effectiveness of long-term psychoanalytic psychotherapy—A meta-analysis of randomized controlled trials. *Clinical Psychology Review, 32,* 81-92.

*Sockol, L. E., Epperson, C. N., & Barber, J. P. (2011). A meta-analysis of treatments for perinatal depression. *Clinical Psychology Review*, *31*, 839-849.

*Spek, V., Cuijpers, P., Nyklícek, I., Riper, H., Keyzer, J., & Pop, V. (2007). Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. *Psychological Medicine, 37*, 319-328.

*Sprenger, L., Gerhards, F., & Goldbeck, L. (2011). Effects of psychological treatment on recurrent abdominal pain in children - A meta-analysis. *Clinical Psychology Review, 31,* 1192-1197.

Thompson, S.G., & Higgins, J.P.T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*, 1559–1573.

Valentine, J. C., & Cooper, H. M. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinhouse.

Van den Noortgate, W., & Onghena, P. (2005). Parametric and nonparametric bootstrap methods for meta-analysis. *Behavior Research Methods, 37*, 11–22.

Veroniki, A.A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J.P.T., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*, 55–79.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*, 261–293.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48.

Viechtbauer, W., López-López, J.A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*, *20*, 360-374.

*Viués-Ortega, J. (2010). Applied behavior analytic intervention for autism in early childhood: Meta-analysis, meta-regression and dose–response meta-analysis of multiple outcomes. *Clinical Psychology Review*, *30*, 387-399.

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, *34*, 28–35.

*Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: an empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *69*, 875.

*Williams, J., Hadjistavropoulos, T., & Sharpe, D. (2006). A meta-analysis of psychological and pharmacological treatments for body dysmorphic disorder. *Behaviour Research and Therapy*, *44*, 99-111.

*Wittouck, C., Van Autreve, S., De Jaegere, E., Portzky, G., & van Heeringen, K. (2011). The prevention and treatment of complicated grief: A meta-analysis. *Clinical Psychology Review*, *31*, 69-78.

*Young, K. M., Northern, J. J., Lister, K. M., Drummond, J. A., & O'Brien, W. H. (2007). A meta-analysis of family-behavioral weight-loss treatments for children. *Clinical Psychology Review*, *27*, 240-249.

Appendices

Appendix 2A

R code for posttest standardized mean difference in Chapter 2

```
install.packages("metafor")
install.packages("QRM")
library (metafor)
library (QRM)

meta = 41
vd = function (Ntratado,Ncontrol,d) {
  var =
(Ntratado+Ncontrol)/(Ntratado*Ncontrol)+d^2/(2*(Ntratado+Ncontrol))

}
Is= rep(NA,meta)
Qs = rep(NA,meta)
pQs= rep (NA,meta)
ks = rep(NA,meta)
d_pnorms = rep(NA,meta)
d_asims = rep(NA,meta)
d_curts = rep(NA,meta)
d_medias = rep(NA,meta)
d_ICLi = rep(NA,meta)
d_ICLs = rep(NA,meta)
d_tau2s = rep(NA,meta)
d_tau2sREML = rep(NA,meta)
d_tau2sPM = rep(NA,meta)
N_asims = rep(NA,meta)
N_curts = rep(NA,meta)
N_mins = rep(NA,meta)
N_maxs = rep(NA,meta)
N_medias = rep(NA,meta)
N_mdns = rep(NA,meta)
ratiovars = rep(NA,meta)
corsNd = rep(NA,meta)
corsNd_p = rep(NA,meta)
cor_neg = 0
cor_nula = 0
cor_pos = 0
for (i in (1:meta)) {
 a = read.csv(paste("BASEd",i,".csv",sep=""), header = T, sep=";",
dec=",")
 vi =vd(a$Ntratado,a$Ncontrol,a$d)
 ks[i] = length(a$d)
  b = shapiro.test(a$d)
  d_pnorms[i] = b$p.value
  d_asims[i] = skewness(a$d)
  d_curts[i] = kurtosis(a$d)

  c = rma(a$d,vi,method="DL")
  c_REML= rma(a$d,vi,method="REML")
  c_PM= rma(a$d,vi,method="PM")

  Is[i]=c$I2[1]
  Qs[i]=c$QE[1]
  pQs[i]=c$QEp[1]
```

```
  d_medias[i] = abs(c$b[1])
  d_ICLi[i]= c$ci.lb
  d_ICLs[i]= c$ci.ub
  d_tau2s[i] = c$tau2
  d_tau2sREML[i] = c_REML$tau2
  d_tau2sPM[i] = c_PM$tau2
  N_asims[i] = skewness(a$Ntotal)
  N_curts[i] = kurtosis(a$Ntotal)
  N_mins[i] = min(a$Ntotal)
  N_maxs[i] = max(a$Ntotal)
  N_mdns[i] = median(a$Ntotal)
  rationvars[i] = mean(vi)/c$tau2
  if(c$tau2==0){
    rationvars[i]=NA}
  e=cor.test(a$Ntotal,a$d)
  corsNd[i] = e$estimate
  corsNd_p[i] = e$p.value

  if(e$estimate<0 & e$p.value<= .05) {
    cor_neg = cor_neg+1 }
  if(e$p.value> .05) {
    cor_nula = cor_nula+1}
  if(e$estimate>0 & e$p.value<= .05) {
    cor_pos = cor_pos+1}
  }
############################################################
summary(ks)
summary(d_pnorms)
mean(d_pnorms <= .05,na.rm=T)
summary(d_asims)
summary(d_curts)
summary(d_medias)
summary(d_ICLi)
summary(d_ICLs)
summary(d_tau2s)
summary(d_tau2sREML)
summary(d_tau2sPM)
summary(Is)
summary(pQs)
summary(N_asims)
summary(N_curts)
summary(N_mdns)
data.frame(N_mins,N_maxs)
summary(rationvars,na.rm=T)
summary(corsNd)
mean(corsNd_p <= .05,na.rm=T)
summary(corsNd_p)
cor_neg
cor_nula
cor_pos
```

R code for standardized mean change in Chapter 2

```
install.packages("metafor")
install.packages("QRM")
library (metafor)
library (QRM)

meta = 11
vdintra1 = function(Ntotal,d) {
  var = (Ntotal-1)/(Ntotal*(Ntotal-3))*(1+Ntotal*d^2)-d^2/((1-
3/(4*Ntotal-5)))^2
  var
}

vdintra2=function(Ntotal,d){

  var=(0.7/Ntotal)*((Ntotal-1)/(Ntotal-3))*(1+((Ntotal*d^2)/(0.7)))-
d^2/((1-3/(4*Ntotal-5)))^2
  var

}

Is= rep(NA,meta)
Qs = rep(NA,meta)
pQs= rep (NA,meta)
ks = rep(NA,meta)
d_pnorms = rep(NA,meta)
d_asims = rep(NA,meta)
d_curts = rep(NA,meta)
d_medias = rep(NA,meta)
d_ICLi = rep(NA,meta)
d_ICLs = rep(NA,meta)
d_tau2s = rep(NA,meta)
d_tau2sREML = rep(NA,meta)
d_tau2sPM = rep(NA,meta)
N_asims = rep(NA,meta)
N_curts = rep(NA,meta)
N_mins = rep(NA,meta)
N_maxs = rep(NA,meta)
N_medias = rep(NA,meta)
N_mdns = rep(NA,meta)
ratiovars = rep(NA,meta)
corsNd = rep(NA,meta)
corsNd_p = rep(NA,meta)

cor_neg = 0
cor_nula = 0
cor_pos = 0

for (i in (1:meta)) {


  a = read.csv(paste("BASEdMR",i,".csv",sep=""), header = T, sep=";",
dec=",")


if(i==4 || i==7 || i==9) {vi= vdintra1(a$Ntotal,a$d)}  else{
vi=vdintra2(a$Ntotal,a$d)}


  ks[i] = length(a$d)
  b = shapiro.test(a$d)
  d_pnorms[i] = b$p.value
  d_asims[i] = skewness(a$d)
  d_curts[i] = kurtosis(a$d)
```

```
  c = rma(a$d,vi,method="DL")
  c_REML= rma(a$d,vi,method="REML")
  c_PM= rma(a$d,vi,method="PM")

  Is[i]=c$I2[1]
  Qs[i]=c$QE[1]
  pQs[i]=c$QEp[1]

  d_medias[i] = abs(c$b[1])
  d_ICLi[i]= c$ci.lb
  d_ICLs[i]= c$ci.ub
  d_tau2s[i] = c$tau2
  d_tau2sREML[i] = c_REML$tau2
  d_tau2sPM[i] = c_PM$tau2
  N_asims[i] = skewness(a$Ntotal)
  N_curts[i] = kurtosis(a$Ntotal)
  N_mins[i] = min(a$Ntotal)
  N_maxs[i] = max(a$Ntotal)
  N_mdns[i] = median(a$Ntotal)
  rationvars[i] = mean(vi)/c$tau2
  if(c$tau2==0){
    rationvars[i]=NA}
  e=cor.test(a$Ntotal,a$d)
  corsNd[i] = e$estimate
  corsNd_p[i] = e$p.value

  if(e$estimate<0 & e$p.value<= .05) {
    cor_neg = cor_neg+1 }
  if(e$p.value> .05) {
    cor_nula = cor_nula+1}
  if(e$estimate>0 & e$p.value<= .05) {
    cor_pos = cor_pos+1}

}



############################################################


summary(ks)
summary(d_pnorms)
mean(d_pnorms <= .05,na.rm=T)
summary(d_asims)
summary(d_curts)
summary(d_medias)
summary(d_ICLi)
summary(d_ICLs)
summary(d_tau2s)
summary(d_tau2sREML)
summary(d_tau2sPM)
summary(Is)
summary(pQs)
summary(N_asims)
summary(N_curts)
summary(N_mdns)
data.frame(N_mins,N_maxs)
summary(rationvars,na.rm=T)
summary(corsNd)
mean(corsNd_p <= .05,na.rm=T)
summary(corsNd_p)
cor_neg
cor_nula
cor_pos
```

R code for standardized mean change difference in Chapter 2

```
install.packages("metafor")
install.packages("QRM")
library (metafor)
library (QRM)

meta = 2

vdelta = function(Ntratado,Ncontrol,d) {
  c1 = 0.6*((Ntratado+Ncontrol)/(Ntratado*Ncontrol))
  c2 = ((Ntratado+Ncontrol-2)/(Ntratado+Ncontrol-4))
  c3 = 1+(Ntratado*Ncontrol)/(0.6*(Ntratado+Ncontrol))*d^2
  cm = 1-3/(4*(Ntratado+Ncontrol)-9)
  c4 = d^2/cm^2
  var = c1*c2*c3-c4
  var
}

Is= rep(NA,meta)
Qs = rep(NA,meta)
pQs= rep (NA,meta)

ks = rep(NA,meta)
d_pnorms = rep(NA,meta)
d_asims = rep(NA,meta)
d_curts = rep(NA,meta)
d_medias = rep(NA,meta)
d_tau2s = rep(NA,meta)
d_tau2sREML = rep(NA,meta)
d_tau2sPM = rep(NA,meta)
N_asims = rep(NA,meta)
N_curts = rep(NA,meta)
N_mins = rep(NA,meta)
N_maxs = rep(NA,meta)
N_medias = rep(NA,meta)
N_mdns = rep(NA,meta)
rationvars = rep(NA,meta)
corsNd = rep(NA,meta)
corsNd_p = rep(NA,meta)

cor_neg = 0
cor_nula = 0
cor_pos = 0

for (i in (1:meta)) {

a = read.csv(paste("BASEdc",i,".csv",sep=""), header = T, sep=";",
dec=",")

  vi =vdelta(a$Ntratado,a$Ncontrol,a$d)


  ks[i] = length(a$d)
  b = shapiro.test(a$d)
  d_pnorms[i] = b$p.value
  d_asims[i] = skewness(a$d)
  d_curts[i] = kurtosis(a$d)
  c = rma(a$d,vi,method="DL")
  c_REML= rma(a$d,vi,method="REML")
  c_PM= rma(a$d,vi,method="PM")
  Is[i]=c$I2[1]
  Qs[i]=c$QE[1]
  pQs[i]=c$QEp[1]
```

```
  d_medias[i] = abs(c$b[1])
  d_tau2s[i] = c$tau2
  d_tau2sREML[i] = c_REML$tau2
  d_tau2sPM[i] = c_PM$tau2
  N_asims[i] = skewness(a$Ntotal)
  N_curts[i] = kurtosis(a$Ntotal)
  N_mins[i] = min(a$Ntotal)
  N_maxs[i] = max(a$Ntotal)
  N_mdns[i] = median(a$Ntotal)
  rationvars[i] = mean(vi)/c$tau2
  if(c$tau2==0){
    rationvars[i]=NA}
  e=cor.test(a$Ntotal,a$d)
  corsNd[i] = e$estimate
  corsNd_p[i] = e$p.value
  if(e$estimate<0 & e$p.value<= .05) {
    cor_neg = cor_neg+1 }
  if(e$p.value> .05) {
    cor_nula = cor_nula+1}
  if(e$estimate>0 & e$p.value<= .05) {
    cor_pos = cor_pos+1}

}


############################################################

summary(ks)
summary(d_pnorms)
mean(d_pnorms <= .05,na.rm=T)
summary(d_asims)
summary(d_curts)
summary(d_medias)
summary(d_ICLi)
summary(d_ICLs)
summary(d_tau2s)
summary(d_tau2sREML)
summary(d_tau2sPM)
summary(Is)
summary(pQs)
summary(N_asims)
summary(N_curts)
summary(N_mdns)
data.frame(N_mins,N_maxs)
summary(rationvars,na.rm=T)
summary(corsNd)
mean(corsNd_p <= .05,na.rm=T)
summary(corsNd_p)
cor_neg
cor_nula
cor_pos
```

Appendix 2B

Characteristics of the meta-analyses included in the systematic review

Meta-analysis	*d* index (Equation)	Formula for the sampling variance	*k*	$\bar{d}$	$\hat{\tau}^2_{DL}$	$\hat{\tau}^2_{REML}$	$\hat{\tau}^2_{PM}$	*p*	I^2
Abramowitz et al. (2001)	*d* (2.1)	Eq. 2.3	54	0.250	0.522	0.625	0.719	< .0001	81.3
Acarturk et al. (2009)	*d* (2.1)	Eq. 2.3	45	0.740	0.111	0.104	0.088	.0001	50.3
Aderka et al. (2012)	d_{g3} (2.12)	Eq. 2.13	19	0.630	0.109	0.083	0.066	< .0001	68.5
Bell & D'Zurilla (2009)	*d* (2.1)	Eq. 2.3	21	0.694	0.391	0.560	0.567	< .0001	83.6
Benish et al. (2008)	*d* (2.1)	Eq. 2.3	15	0.187	0.000	0.000	0.000	.9808	0
Burke et al. (2003)	*d* (2.1)	Eq. 2.3	13	0.291	0.022	0.019	0.020	.0824	37.7
Casement & Swanson (2012)	d_{c2} (2.6)	Eq. 2.7	13	0.696	0.090	0.070	0.062	< .0001	77.9
Cuijpers et al. (2009)	*d* (2.1)	Eq. 2.3	19	0.307	0.002	0.012	0.001	.4098	3.8
Cuijpers, Li et al. (2010)	*d* (2.1)	Eq. 2.3	70	0.195	0.058	0.056	0.070	.0021	35.8
Cuijpers, Donker et al. (2010)	*d* (2.1)	Eq. 2.3	24	0.067	0.098	0.101	0.099	.0093	45.1
Cuijpers et al. (2011)	*d* (2.1)	Eq. 2.3	15	0.289	0.000	0.000	0.000	.8662	0
Cuijpers et al. (2012)	*d* (2.1)	Eq. 2.3	18	0.589	0.008	0.007	0.008	.3501	8.8
Dixon et al. (2007)	*d* (2.1)	Eq. 2.3	20	0.205	0.008	0.000	0.014	.2493	16.4
Driessen et al. (2010)	d_{c2} (2.6)	Eq. 2.7	21	1.266	0.152	0.177	0.173	< .0001	72.7
Ekers et al. (2008)	*d* (2.1)	Eq. 2.3	14	0.072	0.051	0.038	0.061	.1607	27.4
Gooding & Tarrier (2009)	*d* (2.1)	Eq. 2.3	18	0.726	0.163	0.179	0.164	< .0001	67.8
Hanrahan et al. (2012)	*d* (2.1)	Eq. 2.3	19	0.928	0.433	0.689	0.789	< .0001	81.8
Hansen et al. (2012)	d_{c2} (2.6)	Eq. 2.7	11	0.563	0.084	0.106	0.103	< .0001	83.1
Harris (2006)	*d* (2.1)	Eq. 2.3	14	0.240	0.081	0.092	0.109	.0004	65.3
Haug et al. (2012)	*d* (2.1)	Eq. 2.3	54	0.799	0.132	0.130	0.123	< .0001	71.8
Hausenblas et al. (2013)	d_{c1} (2.4)	Eq. 2.5	54	0.038	0.056	0.064	0.069	< .0001	62.9
Hesser et al. (2011)a	*d* (2.1)	Eq. 2.3	25	0.600	0.046	0.043	0.047	.0123	43.1
Hesser et al. (2011)b	d_{c2} (2.6)	Eq. 2.7	10	0.584	0.100	0.104	0.102	< .0001	90.2
Kalu et al. (2012)	*d* (2.1)	Eq. 2.3	7	0.738	0.272	0.311	0.389	.0266	58.0
Kleinstäuber et al. (2011)	*d* (2.1)	Eq. 2.3	18	0.399	0.142	0.169	0.192	< .0001	75.1
Lackner et al. (2004)	*d* (2.1)	Eq. 2.3	12	0.766	0.106	0.109	0.165	.0354	47.1

Meta-analysis	*d* index (Equation)	Formula for the sampling variance	*k*	$\bar{d}$	$\hat{\tau}^2_{DL}$	$\hat{\tau}^2_{REML}$	$\hat{\tau}^2_{PM}$	*p*	I^2
Lansbergen et al. (2007)	*d* (2.1)	Eq. 2.3	18	0.220	0.384	0.636	0.685	< .0001	86.1
Lissek et al. (2005)	*d* (2.1)	Eq. 2.3	22	0.219	0.130	0.140	0.162	.0002	59.9
Lundahl et al. (2006)	*d* (2.1)	Eq. 2.3	70	0.463	0.055	0.058	0.046	.0004	39.9
Malouff et al. (2007)	*d* (2.1)	Eq. 2.3	38	0.546	0.292	0.538	0.680	< .0001	83.6
Malouff et al. (2008)	*d* (2.1)	Eq. 2.3	15	0.476	0.125	0.150	0.163	< .0001	74.8
Nestoriuc et al. (2008)a	*d* (2.1)	Eq. 2.3	18	0.298	0.018	0.015	0.021	.2970	13.1
Nestoriuc et al. (2008)b	d_{c1} (2.4)	Eq. 2.5	70	0.747	0.124	0.112	0.161	< .0001	44.9
Oldham et al. (2012)	*d* (2.1)	Eq. 2.3	33	0.378	0.062	0.065	0.066	< .0001	65.0
Opris et al. (2012)	*d* (2.1)	Eq. 2.3	23	0.490	0.255	0.294	0.352	< .0001	67.4
Pérez-Mañá et al. (2011)	*d* (2.1)	Eq. 2.3	21	0.203	0.081	0.071	0.059	< .0001	63.4
Prendergast et al. (2001)	*d* (2.1)	Eq. 2.3	11	0.393	0.157	0.108	0.083	< .0001	75.0
Richards& Richardson (2012)	*d* (2.1)	Eq. 2.3	33	0.565	0.141	0.139	0.129	< .0001	81.1
Roberts et al. (2007)	*d* (2.1)	Eq. 2.3	14	0.363	0.023	0.011	0.027	.1971	23.8
Rodenburg et al. (2009)	*d* (2.1)	Eq. 2.3	7	0.560	0.065	0.068	0.064	.1831	32.1
Rosa-Alcázar et al (2008)	*d* (2.1)	Eq. 2.3	24	1.075	0.173	0.172	0.378	.0002	57.9
Sánchez-Meca et al. (2010)	*d* (2.1)	Eq. 2.3	61	1.012	0.261	0.317	0.363	< .0001	71.0
Shadish & Baldwin (2005)	*d* (2.1)	Eq. 2.3	30	0.708	0.160	0.035	0.431	.0014	49.3
Smit et al. (2012)	*d* (2.1)	Eq. 2.3	10	0.331	1.024	0.816	0.789	< .0001	93.6
Sockoll et al. (2011)a	*d* (2.1)	Eq. 2.3	14	0.764	0.171	0.189	0.194	< .0001	70.4
Sockoll et al. (2011)b	d_{c1} (2.4)	Eq. 2.5	24	1.662	0.467	0.521	0.510	< .0001	80.1
Spek et al. (2007)	*d* (2.1)	Eq. 2.3	11	0.409	0.077	0.134	0.145	< .0001	78.6
Sprenger et al. (2011)	*d* (2.1)	Eq. 2.3	10	0.627	0.114	0.114	0.147	.0174	55.2
Virués-Ortega (2010)a	d_{c2} (2.6)	Eq. 2.7	8	0.984	0.159	0.179	0.294	.0037	66.8
Virués-Ortega (2010)b	d_{g3} (2.12)	Eq. 213	9	1.307	0.242	0.213	0.190	.0010	69.5
Western & Morrison (2001)	d_{c2} (2.6)	Eq. 2.7	8	2.220	0.512	0.588	0.589	< .0001	93.5
Williams et al. (2006)	d_{c2} (2.6)	Eq. 27	10	1.250	0.117	0.136	0.073	.0131	56.9
Wittouck et al. (2011)	*d* (2.1)	Eq. 2.3	14	0.163	0.095	0.105	0.190	.0005	64.3
Young et al. (2007)	d_{c2} (2.6)	Eq. 2.7	36	0.725	0.167	0.261	0.387	< .0001	71.0

Note. a and b labels after a study indicate separate analyses of two sets of reported effect sizes. *d* = posttest standardized mean difference; d_{c1} = standardized mean change calculated using in the denominator the standard deviation of the pretest-posttest change scores; d_{c2} = standardized mean change calculated using in the denominator the standard deviation of the pretest scores; d_{g3} = standardized mean change difference calculated using in the denominator an average of the pretest standard deviations in the experimental and control groups; *k* = number of studies; $\overline{d}$ = mean effect size applying DL to estimate the between-studies variance; $\hat{\tau}^2_{DL}$ = between-studies variance estimated using the DerSimonian and Laird (1986) method; $\hat{\tau}^2_{REML}$ = between-studies variance estimated using restricted maximum likelihood; $\hat{\tau}^2_{PM}$ = between-studies variance estimated using Paule and Mandel's (1982) method; *p* = *p*-value associated to the heterogeneity *Q* statistic; I^2 = index to quantify the amount of heterogeneity (%).

Appendix 3A

R code of simulation study in Chapter 3

```
#install.packages("metafor")
#install.packages("metaLik")
#install.packages("boot")

### need to install metaLik version 0.41.0 to make this code work
#install.packages("https://cran.r-
project.org/src/contrib/Archive/metaLik/metaLik_0.41.0.tar.gz")

library(metafor)
library(metaLik)
library(boot)
library(parallel)

### load adjusted profile.metaLik() function that doesn't print output
source("profile.metaLik.r")

### Set of conditions
v1.mus = c(0,.2,.5,.8)
v2.tau2s = c(.03,.06,.11,.18,.39)
v3.shapes = c(0,1,2,3,4,5)
v4.ks = c(10,20,40,60)
v5.averNs = c(20,30,50,100)

### to split up by mu, set value of mu here
v1.mus = 0

tabla.condiciones0 = expand.grid(v1.mus, 0, 0, v4.ks, v5.averNs)
tabla.condiciones1 = expand.grid(v1.mus, v2.tau2s, v3.shapes, v4.ks,
v5.averNs)
tabla.condiciones = rbind(tabla.condiciones0,tabla.condiciones1)
colnames(tabla.condiciones) <- c("mus", "tau2s", "shapes", "ks",
"averNs")

iters = 10000

cores <- 60
cl <- makePSOCKcluster(cores)

################################################ Fleishman`s Algorithm

sandk <- function(x){
  ##
  #Calculates the mean, variance, skew, and kurtosis#
  # for a data set and returns them in that order.#
  #The formulas for skew and kurtosis are from page 85#
  # of Kendall and Steward 1969, vol.1#
  ##
  n <- length(x)
  m1p <- mean(x)
  m2 <- sum((x-m1p)^2)/n
  m3 <- sum((x-m1p)^3)/n
  m4 <- sum((x-m1p)^4)/n
  k1 <- m1p
  k2 <- n*m2/(n-1)
  k3 <- ((n^2)/((n-1)*(n-2)))*m3
  k4 <- (((n^2)/((n-1)*(n-2)))/(n-3))*((n+1)*m4 - 3*(n-1)*m2^2)
  g1 <- k3/(k2^(3/2))
  g2 <- k4/(k2^2)
  return(c(k1,k2,g1,g2))
}
```

```
fleishtarget <- function(x,a){
  ##
  #The target function for solving equations 18, 19, and 20#
  # from page 523 of Fleishman.#
  #It does this by changing the system of three equations into a #
  # minimization problem.  Set the equations equal to zero, square,#
  # and sum up. The b, c, and d that minimize the set of equations #
  # at zero must also solve the three individually.#
  ##
  b<-x[1]
  cc<-x[2]
  d<-x[3]
  g1<-a[1]
  g2<-a[2]
  (2 - ( 2*b^2 + 12*b*d + g1^2/(b^2+24*b*d+105*d^2+2)^2 + 30*d^2 ) )^2
+
    (g2 - (
24*(b*d+cc^2*(1+b^2+28*b*d)+d^2*(12+48*b*d+141*cc^2+225*d^2)) ) )^2+
    (cc - (g1/(2*(b^2+24*b*d+105*d^2+2)) ) )^2
}

findbcd <- function(skew,kurtosis){
  ##
  #Uses the built in minimization function to solve for b, c, and d#
  # if the skew and kurtosis are given. Try findbcd(1.75,3.75) and#
  # compare to Table 1 on page 524 of Fleishman.
  ##
  optim(c(1,0,0),fleishtarget,a=c(skew,kurtosis),method="BFGS",
        control=list(ndeps=rep(1e-10,3),reltol=1e-10,maxit=1e8))
}

rfleish <- function(n,mean=0,var=1,skew=0,kurtosis=0){
  ##
  #Generates n random variables with specified first four moments#
  # using Fleishman's power method. Note that not all combinations#
  # of skew and kurtosis are possible (see Figure 1 on page 527).#
  #Must satisfy skew^2 < 0.0629576*kurtosis + 0.0717247.#
  ##
  Z<-rnorm(n,0,1)
  bcd<-findbcd(skew,kurtosis)$par
  b<-bcd[1]
  cc<-bcd[2]
  d<-bcd[3]
  a<--1*cc
  Y<-a+b*Z+cc*Z^2+d*Z^3
  X<-mean+sqrt(var)*Y
  return(X)
}

##############################################

cmm.approx=function(mi){1-3/(4*mi-1)}

### no longer needed
#ghedges=function(ge,gc){num=mean(ge)-mean(gc); ne=length(ge);
nc=length(gc)
#den=sqrt(((ne-1)*(sd(ge)^2)+(nc-1)*(sd(gc)^2))/(ne+nc-2));num/den}
#dhedges=function(ge,gc){gl=length(ge)+length(gc)-2;
cmm.approx(gl)*ghedges(ge,gc)}
#dhedvar=function(n,d){2/n+d^2/(4*n)} ### where n=ne=nc

### function to directly simulate bias-corrected d-values and sampling
variances without having to simulate raw data
sim.d <- function(k, deltai, n1i, n2i) {
  mi <- n1i + n2i - 2
  yi <- rnorm(k, mean=deltai, sd=sqrt(1/n1i + 1/n2i)) / sqrt(rchisq(k,
df=mi) / mi)
  yi <- cmm.approx(mi) * yi
  vi <- 1/n1i + 1/n2i + yi^2 / (2*(n1i + n2i))
  return(data.frame(yi, vi))
```

```
}

# bootstrap function

boot.func = function(data, indices) {

  #library(metafor)
  #res = try(rma(yi, vi, data=data, subset=indices, method="DL"),
silent=TRUE)
  #if (is.element("try-error", class(res))) {
  #  NA
  #} else {
  #  c(coef(res), vcov(res), res$tau2, res$se.tau2^2)
  #}

  dat <- data[indices,]

  k <- nrow(dat)
  yi <- dat$yi
  vi <- dat$vi
  wi <- 1/vi
  theta.hat <- sum(wi*yi)/sum(wi)
  Q <- sum(wi * (yi - theta.hat)^2)
  sumwi <- sum(wi)
  cval <- sumwi - sum(wi^2)/sumwi
  tau2 <- max(0, (Q - (k-1)) / cval)
  var.tau2 <- 2/cval^2 * (sum(wi^2 * (vi+tau2)^2) - 2*sum(wi^3 *
(vi+tau2)^2)/sumwi + (sum(wi^2 * (vi+tau2)))^2 / sumwi^2)
  wi <- 1/(vi + tau2)
  sumwi <- sum(wi)
  mu <- sum(wi*yi)/sumwi
  var.mu <- 1/sumwi

  return(c(mu, var.mu, tau2, var.tau2))

}

for (condicion in 1:nrow(tabla.condiciones)) {

  mu    <- tabla.condiciones[condicion,1]
  tau2  <- tabla.condiciones[condicion,2]
  shape <- tabla.condiciones[condicion,3]
  k     <- tabla.condiciones[condicion,4]
  averN <- tabla.condiciones[condicion,5]

  if (shape==0) { # conditionals to set skewness and kurtosis values
    asym = 0
    curt = 0
  }
  if (shape==1) {
    asym = -2
    curt = 3.65
  }
  if (shape==2) {
    asym = -1
    curt = 0.47
  }
  if (shape==3) {
    asym = 0
    curt = -0.58
  }
  if (shape==4) {
    asym = 1
    curt = 0.51
  }
  if (shape==5) {
    asym = 2
    curt = 3.74
  }
```

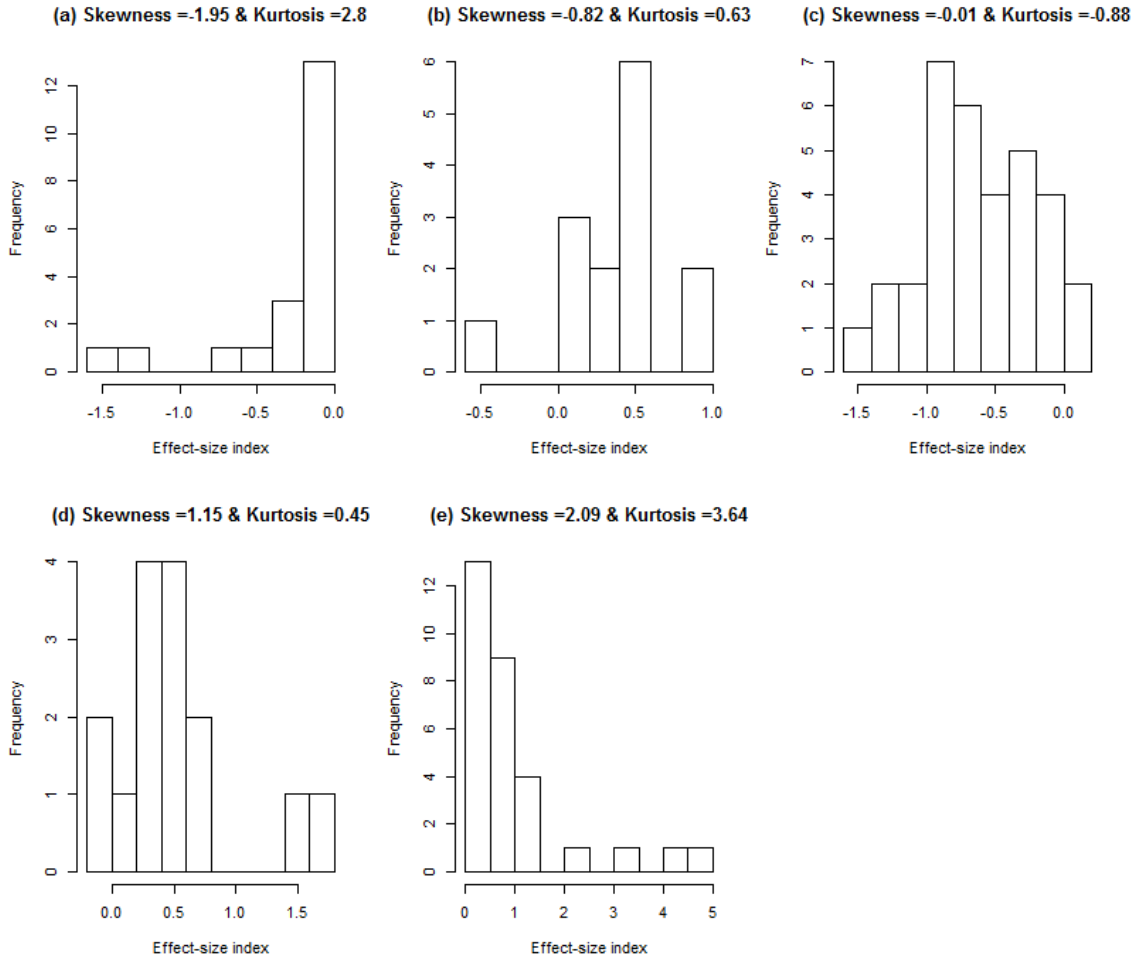
```
  name <- paste(mu, " ", tau2, " ", shape," ", k, " ", averN)

  mDL=rep(NA,iters)
  mREML=rep(NA,iters)
  mEB=rep(NA,iters)
  mFE=rep(NA,iters)
  mBoots=rep(NA,iters)

  cobDL=rep(NA,iters)
  cobREML=rep(NA,iters)
  cobEB=rep(NA,iters)
  cobKHDL=rep(NA,iters)
  cobKHREML=rep(NA,iters)
  cobKHEB=rep(NA,iters)
  cobFE=rep(NA,iters)
  cobBootsP=rep(NA,iters)
  cobBootsBCa=rep(NA,iters)
  cobPL=rep(NA,iters)

  ampDL=rep(NA,iters)
  ampREML=rep(NA,iters)
  ampEB=rep(NA,iters)
  ampKHDL=rep(NA,iters)
  ampKHREML=rep(NA,iters)
  ampKHEB=rep(NA,iters)
  ampFE=rep(NA,iters)
  ampBootsP=rep(NA,iters)
  ampBootsBCa=rep(NA,iters)
  ampPL=rep(NA,iters)

  etDL=rep(NA,iters)
  etREML=rep(NA,iters)
  etEB=rep(NA,iters)
  etKHDL=rep(NA,iters)
  etKHREML=rep(NA,iters)
  etKHEB=rep(NA,iters)
  etFE=rep(NA,iters)
  etBoots=rep(NA,iters)

  i = 1

  while (i <= iters) {

    ## simulate the data

    deltas = rfleish(k,mu,tau2,asym,curt)
    Ns = rchisq(k,4)+averN-4 # variable follows a chi-square
distribution with mean averN and asymmetry around +1.4
    ns = round(Ns/2)
    data <- sim.d(k, deltas, ns, ns)

    ##1 METHOD: ESTANDAR_DL

    resDL = rma(yi, vi, data=data, method="DL")
    mDL[i] = resDL$b
    ampDL[i] = resDL$ci.ub - resDL$ci.lb
    cobDL[i] = ifelse(mu <= resDL$ci.ub & mu >= resDL$ci.lb,1,0)
    etDL[i] = resDL$se

    ##2 METHOD: ESTANDAR_REML

    resREML = try(rma(yi, vi, data=data, method="REML"), silent=TRUE)
    if (inherits(resREML, "try-error")) ### skip iteration if REML
doesn't converge
      next
    mREML[i] = resREML$b
    ampREML[i] = resREML$ci.ub - resREML$ci.lb
    cobREML[i] = ifelse(mu <= resREML$ci.ub & mu >= resREML$ci.lb,1,0)
    etREML[i] = resREML$se
```

```
    ##3 METHOD: ESTANDAR_EB

    resEB = rma(yi, vi, data=data, method="EB")
    if (inherits(resEB, "try-error")) ### skip iteration if EB doesn't
converge
      next
    mEB[i] = resEB$b
    ampEB[i] = resEB$ci.ub - resEB$ci.lb
    cobEB[i] = ifelse(mu <= resEB$ci.ub & mu >= resEB$ci.lb,1,0)
    etEB[i] = resEB$se

    ##4 METHOD: KNAPPHARTUNG_DL

    resKHDL = rma(yi, vi, data=data, method="DL", knha=TRUE)
    ampKHDL[i] = resKHDL$ci.ub - resKHDL$ci.lb
    cobKHDL[i] = ifelse(mu <= resKHDL$ci.ub & mu >= resKHDL$ci.lb,1,0)
    etKHDL[i] = resKHDL$se

    ##5 METHOD: KNAPPHARTUNG_REML

    resKHREML = rma(yi, vi, data=data, method="REML", knha=TRUE)
    ampKHREML[i] = resKHREML$ci.ub - resKHREML$ci.lb
    cobKHREML[i] = ifelse(mu <= resKHREML$ci.ub & mu >=
resKHREML$ci.lb,1,0)
    etKHREML[i] = resKHREML$se

    ##6 METHOD: KNAPPHARTUNG_EB

    resKHEB = rma(yi, vi, data=data, method="EB", knha=TRUE)
    ampKHEB[i] = resKHEB$ci.ub - resKHEB$ci.lb
    cobKHEB[i] = ifelse(mu <= resKHEB$ci.ub & mu >= resKHEB$ci.lb,1,0)
    etKHEB[i] = resKHEB$se

    ##7 METHOD: FIXED-EFFECT

    resFE = rma(yi, vi, data=data, method="FE")
    mFE[i] = resFE$b
    ampFE[i] = resFE$ci.ub - resFE$ci.lb
    cobFE[i] = ifelse(mu <= resFE$ci.ub & mu >= resFE$ci.lb,1,0)
    etFE[i] = resFE$se

    ##8 METHOD: NON PARAMETRIC BOOTSTRAPPING_DL

    #res.boot = boot(data, boot.func, R=1000)
    res.boot = boot(data, boot.func, R=1000, parallel="snow", cl=cl,
ncpus=cores)
    ciB = boot.ci(res.boot)
    mBoots[i] = (ciB$normal[1,2] + ciB$normal[1,3]) / 2
    ampBootsP[i] = ciB$percent[1,5] - ciB$percent[1,4]
    ampBootsBCa[i] = ciB$bca[1,5] - ciB$bca[1,4]
    cobBootsP[i] = ifelse(mu <= ciB$percent[1,5] & mu >=
ciB$percent[1,4],1,0)
    cobBootsBCa[i] = ifelse(mu <= ciB$bca[1,5] & mu >=
ciB$bca[1,4],1,0)
    etBoots[i] = (ciB$normal[1,3] - (ciB$normal[1,2] +
ciB$normal[1,3]) / 2) / 1.96

    ##9 METHOD: PROFILE LIKELIHOOD

    mlik = metaLik(yi~1, sigma2=vi, data=data)
    ciPL = profile.metaLik(mlik, param=1, display=FALSE)
    ampPL[i] = ciPL$upper.rskov - ciPL$lower.rskov
    cobPL[i] = ifelse(mu <= ciPL$upper.rskov  & mu >=
ciPL$lower.rskov,1,0)

    i = i + 1

  }

  bias_DL   = round(mean(mDL)-mu,4)
```

```
  bias_REML = round(mean(mREML)-mu,4)
  bias_EB   = round(mean(mEB)-mu,4)
  bias_FE   = round(mean(mFE)-mu,4)
  bias_BOOT = round(mean(mBoots)-mu,4)

  mse_DL   = round(mean((mDL-mu)^2),4)
  mse_REML = round(mean((mREML-mu)^2),4)
  mse_EB   = round(mean((mEB-mu)^2),4)
  mse_FE   = round(mean((mFE-mu)^2),4)
  mse_BOOT = round(mean((mBoots-mu)^2),4)

  ajnc_DL       = round(mean(cobDL),4)
  ajnc_REML     = round(mean(cobREML),4)
  ajnc_EB       = round(mean(cobEB),4)
  ajnc_KHDL     = round(mean(cobKHDL),4)
  ajnc_KHREML   = round(mean(cobKHREML),4)
  ajnc_KHEB     = round(mean(cobKHEB),4)
  ajnc_FE       = round(mean(cobFE),4)
  ajnc_BOOT_P   = round(mean(cobBootsP),4)
  ajnc_BOOT_BCa = round(mean(cobBootsBCa),4)
  ajnc_PL       = round(mean(cobPL),4)

  ac_DL       = round(mean(ampDL),4)
  ac_REML     = round(mean(ampREML),4)
  ac_EB       = round(mean(ampEB),4)
  ac_KHDL     = round(mean(ampKHDL),4)
  ac_KHREML   = round(mean(ampKHREML),4)
  ac_KHEB     = round(mean(ampKHEB),4)
  ac_FE       = round(mean(ampFE),4)
  ac_BOOT_P   = round(mean(ampBootsP),4)
  ac_BOOT_BCa = round(mean(ampBootsBCa),4)
  ac_PL       = round(mean(ampPL),4)

  ajse_DL     = round(((sd(mDL)-median(etDL))/sd(mDL))*100,2)
  ajse_REML   = round(((sd(mREML)-median(etREML))/sd(mREML))*100,2)
  ajse_EB     = round(((sd(mEB)-median(etEB))/sd(mEB))*100,2)
  ajse_KHDL   = round(((sd(mDL)-median(etKHDL))/sd(mDL))*100,2)
  ajse_KHREML = round(((sd(mREML)-median(etKHREML))/sd(mREML))*100,2)
  ajse_KHEB   = round(((sd(mEB)-median(etKHEB))/sd(mEB))*100,2)
  ajse_FE     = round(((sd(mFE)-median(etFE))/sd(mFE))*100,2)
  ajse_BOOT   = round(((sd(mBoots)-median(etBoots))/sd(mBoots))*100,2)

  resultados <-
cbind(bias_DL,bias_REML,bias_EB,bias_FE,bias_BOOT,mse_DL,mse_REML,mse_
EB,mse_FE,mse_BOOT,ajnc_DL,ajnc_REML,ajnc_EB,ajnc_KHDL,ajnc_KHREML,ajn
c_KHEB,ajnc_FE,

ajnc_BOOT_P,ajnc_BOOT_BCa,ajnc_PL,ac_DL,ac_REML,ac_EB,ac_KHDL,ac_KHREM
L,ac_KHEB,ac_FE,ac_BOOT_P,ac_BOOT_BCa,ac_PL,ajse_DL,ajse_REML,ajse_EB,
ajse_KHDL,ajse_KHREML,ajse_KHEB,ajse_FE,ajse_BOOT)

  if (length(v1.mus) > 1) {
    write.table(cbind(name, resultados), "results.dat", append = TRUE,
sep =" ", col.names = FALSE, row.names = FALSE, quote = FALSE)
  } else {
    write.table(cbind(name, resultados), paste0("results_mu=", v1.mus,
".dat"), append = TRUE, sep =" ", col.names = FALSE, row.names =
FALSE, quote = FALSE)
  }

}

stopCluster(cl)
```

Appendix 3B

Five examples of real meta-analyses selected from Rubio-Aparicio et al. (in press) systematic review



The histogram represents five examples selected from the Rubio-Aparicio et al. (in press) systematic review. Each histogram refers to the estimated effect sizes distribution, in all cases being the standardized mean difference the effect size index (defined in our Eq. 3.23). These examples have been selected to illustrate the skewness and kurtosis combinations defined in our simulation study. Thus, histogram (a) shows the effect sizes distribution of the Dixon, Keefe, Scipio, Perri, and Abernethy (2009) meta-analysis, which exhibited skewness = -1.95 and kurtosis = 2.8. These values are similar to the pair (-2, 3.65) included in our simulation study. Histogram (b) shows the effect distribution of the Roberts, Tchanturia, Stahl, Southgate, and Treasure (2007) meta-

analysis, which exhibited skewness = -0.82 and kurtosis = 0.63. These values are similar to the pair (-1, 0.47) included in our simulation study. Histogram (c) shows the effect distribution of the Richards and Richardson (2012) meta-analysis, which exhibited skewness = -0.01 and kurtosis = -0.88. These values are similar to the pair (0, -0.58) included in our simulation study. Histogram (d) shows the effect distribution of the Malouff, Thorsteinsson, Rooke, Bhullar, and Schutte (2008) meta-analysis, which exhibited skewness = 1.15 and kurtosis = 0.45. These values are similar to the pair (1, 0.51) included in our simulation study. Finally, histogram (e) shows the effect distribution of the Shadish and Baldwin (2005) meta-analysis, which exhibited skewness = 2.09 and kurtosis = 3.64. These values are similar to the pair (2, 3.74) included in our simulation study.

Appendix 4A

Database for the example

Study	*d*	S_d	Random assignment
1	1.341	0.369	1
2	0.581	0.340	1
3	0.757	0.351	1
4	0.508	0.479	1
5	-0.023	0.558	1
6	0.044	0.277	1
7	0.428	0.270	1
8	0.819	0.521	1
9	-0.086	0.245	2
10	0.602	0.258	2
11	1.282	0.447	2
12	1.023	0.388	2
13	0.927	0.378	2
14	0.483	0.236	2
15	0.807	0.246	2
16	0.692	0.246	2
17	0.594	0.330	2
18	0.582	0.320	2
19	0.697	0.291	2
20	0.833	0.326	2
21	2.651	0.485	2
22	1.232	0.386	2
23	1.896	0.455	2
24	1.837	0.451	2
25	0.281	0.361	2
26	0.410	0.377	2

Study	*d*	S_d	Random assignment
27	0.797	0.402	2
28	0.431	0.377	2
29	0.623	0.394	2
30	0.650	0.365	2
31	1.702	0.498	2
32	1.073	0.480	2
33	0.403	0.404	2
34	3.468	0.520	2
35	3.263	0.496	2
36	3.023	0.488	2
37	1.040	0.389	2
38	1.473	0.460	2
39	1.164	0.441	2
40	0.993	0.427	2
41	-0.344	0.381	2
42	-0.098	0.361	2
43	0.905	0.276	2
44	0.665	0.264	2
45	0.982	0.280	2
46	0.727	0.252	2
47	0.879	0.218	2
48	0.681	0.439	2
49	1.193	0.478	2
50	1.131	0.466	2

Note. *d* = standardized mean difference for each study; S_d = standard error for the *d* index in each study. Random assignment = 1, no; 2, yes

Source: Sánchez-Meca et al. (2010).

Appendix 4B

R code of simulation study in Chapter 4

```
#install.packages("metafor")
#library(metafor)

ks = c(12,20,40,60)
Eqs = c(0,1)
EqLabel = c("EQUILIBRADO", "DESEQUILIBRADO")
tau2as = c(.08,.16)
f1s = c(0,1)
f1sLabel = c("tau2 iguales", "tau2 desiguales")
deltas = c(0.5,0.3,0.1)
f2s = c(0,1)
f2sLabel=c("Grupo1 delta=0.5","Grupo2 delta=0.5")

iters=10000

################################################ Algoritmo de Fleishman.

sandk<-function(x){
##
#Calculates the mean, variance, skew, and kurtosis#
# for a data set and returns them in that order.#
#The formulas for skew and kurtosis are from page 85#
# of Kendall and Steward 1969, vol.1#
##
n <- length(x)
m1p <- mean(x)
m2 <- sum((x-m1p)^2)/n
m3 <- sum((x-m1p)^3)/n
m4 <- sum((x-m1p)^4)/n
k1 <- m1p
k2 <- n*m2/(n-1)
k3 <- ((n^2)/((n-1)*(n-2)))*m3
k4 <- (((n^2)/((n-1)*(n-2)))/(n-3))*((n+1)*m4 - 3*(n-1)*m2^2)
g1 <- k3/(k2^(3/2))
g2 <- k4/(k2^2)
return(c(k1,k2,g1,g2))
}


fleishtarget<-function(x,a){
##
#The target function for solving equations 18, 19, and 20#
# from page 523 of Fleishman.#
#It does this by changing the system of three equations into a #
# minimization problem.  Set the equations equal to zero, square,#
# and sum up.  The b, c, and d that minimize the set of equations #
# at zero must also solve the three individually.#
##
b<-x[1]
cc<-x[2]
d<-x[3]
g1<-a[1]
g2<-a[2]
(2 - ( 2*b^2 + 12*b*d + g1^2/(b^2+24*b*d+105*d^2+2)^2 + 30*d^2 ) )^2 +
(g2 - ( 24*(b*d+cc^2*(1+b^2+28*b*d)+d^2*(12+48*b*d+141*cc^2+225*d^2))
) )^2+
(cc - (g1/(2*(b^2+24*b*d+105*d^2+2)) ) )^2
}


findbcd<-function(skew,kurtosis){
```

```
##
#Uses the built in minimization function to solve for b, c, and d#
# if the skew and kurtosis are given. Try findbcd(1.75,3.75) and#
# compare to Table 1 on page 524 of Fleishman.
##
optim(c(1,0,0),fleishtarget,a=c(skew,kurtosis),method="BFGS",
control=list(ndeps=rep(1e-10,3),reltol=1e-10,maxit=1e8))
}


rfleish<-function(n,mean=0,var=1,skew=0,kurtosis=0){
##
#Generates n random variables with specified first four moments#
# using Fleishman's power method. Note that not all combinations#
# of skew and kurtosis are possible (see Figure 1 on page 527).#
#Must satisfy skew^2 < 0.0629576*kurtosis + 0.0717247.#
##
Z<-rnorm(n,0,1)
bcd<-findbcd(skew,kurtosis)$par
b<-bcd[1]
cc<-bcd[2]
d<-bcd[3]
a<--1*cc
Y<-a+b*Z+cc*Z^2+d*Z^3
X<-mean+sqrt(var)*Y
return(X)}

##############################################



ghedges=function(ge,gc){ne=length(ge); nc=length(gc); num=mean(ge)-
mean(gc)
den=sqrt(((ne-1)*(sd(ge)^2)+(nc-1)*(sd(gc)^2))/(ne+nc-2)); num/den}

dhedges=function(ge,gc){ne=length(ge); nc=length(gc); cm=1-
3/(4*(ne+nc)-9); cm*ghedges(ge,gc)}

dhedvar=function(n,d){2/n+d^2/(4*n)}

qtest=function(d,w){m=sum(d*w)/sum(w); sum(w*(d-m)^2)}

tasarechazo=array(NA,dim=c(2,length(deltas),length(ks),length(Eqs),len
gth(tau2as),length(f1s),length(f2s)),dimnames=list(c("Qpooled","Qsepar
ate"),deltas,ks,EqLabel,tau2as,f1sLabel,f2sLabel))


   for (k in ks){
   for (Eq in Eqs){
   for (tau2a in tau2as){
   for (f1 in f1s){
   for (delta in deltas){
   for (f2 in f2s){

ka = ifelse(Eq==0,k/2,k/4)
kb = ifelse(Eq==0,k/2,3*k/4)
tau2b0 = ifelse(f1==0,tau2a,0.08)
tau2b = ifelse((tau2a==0.08 & f1==1),0.16,tau2b0)
DELTAA = ifelse(f2==0,0.5,delta)
DELTAB = ifelse(f2==0,delta,0.5)

             p_Qs=rep(0,iters)
             p_Qp=rep(0,iters)

            cat("k =", k, "\tGrupos Equilibrados? =", Eq, "\ttau2a =",
tau2a, "\ttau2 iguales? =", f1,"\tdelta =", delta,"\tPosición delta =
", f2, "\n")

             i=1
             while(i<=iters){
```

```
                deltasa=rnorm(ka,DELTAA,sqrt(tau2a))
                nsa=round(rfleish(ka,30,15,1.386,1.427))

                d1=rep(NA,ka)
                mod1=rep(1,ka)


                j=1
                while (j<=ka){

                deltaa=deltasa[j]
                na=nsa[j]

                ge=rnorm(na,deltaa,1); gc=rnorm(na,0,1)

                dh1=dhedges(ge,gc)
                d1[j]=dh1

                j=j+1

                }

                deltasb=rnorm(kb,DELTAB,sqrt(tau2b))
                nsb=round(rfleish(kb,30,15,1.386,1.427))

                d2=rep(NA,kb)
                mod2=rep(0,kb)

                j=1
                while (j<=kb){

                deltab=deltasb[j]
                nb=nsb[j]

                ge=rnorm(nb,deltab,1); gc=rnorm(nb,0,1)

                dh2=dhedges(ge,gc)
                d2[j]=dh2


                j=j+1


                }

var1=dhedvar(na,d1)
a1=rma(d1,var1,method="DL")
tau21=a1$tau2
w1=1/(var1+tau21)

var2=dhedvar(nb,d2)
a2=rma(d2,var2,method="DL")
tau22=a2$tau2
w2=1/(var2+tau22)

dt=c(d1,d2)
vt=c(var1,var2)
wt=c(w1,w2)
mod=c(mod1,mod2)

###Qb separate

qtest1=qtest(d1,w1) ; qtest2=qtest(d2,w2) ; qw= qtest1+qtest2 ;
qtestt=qtest(dt,wt)

Qb=qtestt-qw
P_Qb=1-pchisq(Qb,1)

if(P_Qb<=.05)
```

```
p_Qs[i]=1

###Qb pooled

m=rma(dt,vt,method="DL",mods=~factor(mod))

if(m$QMp<=.05)
p_Qp[i]=1

      i=i+1

    }



tasarechazo["Qpooled",as.character(delta),as.character(k),EqLabel[Eq+1
],as.character(tau2a),f1sLabel[f1+1],f2sLabel[f2+1]]=mean(p_Qp)

tasarechazo["Qseparate",as.character(delta),as.character(k),EqLabel[Eq
+1],as.character(tau2a),f1sLabel[f1+1],f2sLabel[f2+1]]=mean(p_Qs)


       dump("tasarechazo", file="error tipo I y
potencia.txt",append=FALSE)
       }



   }

}
}
}
}
```

Appendix 5A

Equivalence between the *F* statistic for subgroup analysis and meta-regression

In the context of a simple meta-regression, the association of the moderator with the effect sizes can be tested with the Q_R statistic, which is analogous to Q_B (defined in Eq. 4.3). The Q_R statistic is computed with

$$Q_R = Z^2 = \left(\frac{B_1}{\sqrt{V(B_1)}}\right)^2,$$

where B_1 represents the slope estimate indicating how the size of the effect changes as X_i increases by one unit, and is obtained with

$$B_1 = \frac{\sum_{i=1}^{k} \widetilde{w}_i X_i T_i - \sum_{i=1}^{k} \widetilde{w}_i X_i \sum_{i=1}^{k} \widetilde{w}_i T_i}{\sum_{i=1}^{k} \widetilde{w}_i X_i^2 - \left(\sum_{i=1}^{k} \widetilde{w}_i X_i\right)^2},$$

with $\widetilde{w}_i = w_i / \sum_{i=1}^{k} w_i$ and $w_i = 1/(\hat{\sigma}_i^2 + \hat{\tau}_{res}^2)$. The variance of B_1 can be estimated with

$$Var[B_1] = \left(\frac{\sum_{i=1}^{k} w_i X_i^2 - \left(\sum_{i=1}^{k} w_i X_i\right)^2}{\sum_{i=1}^{k} w_i}\right)^2.$$

Similarly, model misspecification can be examined with Q_E, using the same principle as Q_W in subgroup analysis. The Q_E statistic is obtained with

$$Q_E = \sum_{i=1}^{k} w_i (T_i - B_0 - B_1 X_i)^2,$$

where B_0 stands for the model intercept and is given by

$$B_0 = \sum_{i=1}^{k} \widetilde{w}_i T_i - B_1 \sum_{i=1}^{k} \widetilde{w}_i X_i.$$

Then, if the moderator is dichotomous, an *F* statistic for meta-regression, with the same form as that presented in Eq. 5.1 for subgroup analysis, is given by

$$F = \frac{\frac{Q_R}{(m-1)}}{\frac{Q_E}{(k-m)}},$$

where *m* takes a value of 2 and *k* represents the number of studies.

Note that the denominator of the *F* statistic just presented corresponds to the adjustment factor proposed in Knapp and Hartung (2003)

$$s_w^2 = \frac{\sum_{i=1}^{k} w_i(T_i - B_0 - B_1 X_i)^2}{k - m} = \frac{Q_E}{k - m}.$$

Consequently, the *t*-test proposed by Knapp and Hartung to test for a moderator can be computed as

$$t = \frac{Z}{\sqrt{s_w^2}}, \text{ or equivalently } F = t^2 = \frac{Z^2}{s_w^2}.$$

Appendix 5B

R code of simulation study in Chapter 5

```
#install.packages("metafor")
#library(metafor)

v1.ks <- c(12,20,40,60)
v2.Eqs <- c(0,1) ### 0 note balanced distribution of k and 1
unbalanced distribution
v3.tau2as <- c(0.08,0.16,0.32)
v4.tau2bs <- c(0.08,0.16,0.32)
v5.deltas <- c(1,2,3,4)
v6.averNs <- c(20,40,60,80)

tabla.condiciones <- expand.grid(v1.ks, v2.Eqs, v3.tau2as, v4.tau2bs,
v5.deltas, v6.averNs)
colnames(tabla.condiciones) <- c("ks", "Eqs", "tau2as", "tau2bs",
"deltas","averNs")

iters <- 10000

ghedges <- function(ge,gc) {
   ne  <- length(ge)
   nc  <- length(gc)
   num <- mean(ge)-mean(gc)
   den <- sqrt(((ne-1)*(sd(ge)^2)+(nc-1)*(sd(gc)^2))/(ne+nc-2))
   num/den
}

dhedges <- function(ge,gc) {
   ne <- length(ge)
   nc <- length(gc)
   cm <- 1-3/(4*(ne+nc)-9)
   cm*ghedges(ge,gc)
}

dhedvar <- function(n,d)
   2/n+d^2/(4*n)

qtest <- function(d,w) {
   m <- sum(d*w)/sum(w)
   sum(w*(d-m)^2)
}

for (condicion in 1:nrow(tabla.condiciones)) {

   k <- tabla.condiciones[condicion,1]
   Eq <- tabla.condiciones[condicion,2]
   tau2a <- tabla.condiciones[condicion,3]
   tau2b <- tabla.condiciones[condicion,4]
   delta <- tabla.condiciones[condicion,5]
   averN <- tabla.condiciones[condicion,6]

   name <- paste(k, " ", Eq, " ", tau2a," ", tau2b," ", delta, " ",
averN)

      ka <- ifelse(Eq==0, k/2, k/4)   ### with unbalanced
distribution, group/category 1 will be the smallest (K/4)
      kb <- ifelse(Eq==0, k/2, 3*k/4) ### with unbalanced
distribution, group/category 1 will be the largest  (3K/4)

      DELTAA <- ifelse(delta==4, 0.7, 0.5)
      DELTAB <- ifelse(delta==1, 0.5, ifelse(delta==2, 0.3, 0.1))

      p_Qs_DL   <- rep(0, iters)
```

```
p_Qp_DL   <- rep(0, iters)
p_Qs_REML <- rep(0, iters)
p_Qp_REML <- rep(0, iters)
p_Qs_PM   <- rep(0, iters)
p_Qp_PM   <- rep(0, iters)
p_Fs_DL   <- rep(0, iters)
p_Fp_DL   <- rep(0, iters)
p_Fs_REML <- rep(0, iters)
p_Fp_REML <- rep(0, iters)
p_Fs_PM   <- rep(0, iters)
p_Fp_PM   <- rep(0, iters)

i <- 1

while (i <= iters) {

   deltasa <- rnorm(ka,DELTAA,sqrt(tau2a))
   nsa <- rchisq(ka,4)+averN-4

   d1 <- rep(NA,ka)
   mod1 <- rep(1,ka)

   j <- 1

   while (j <= ka) {

      deltaa <- deltasa[j]
      na <- nsa[j]

      ge <- rnorm(na,deltaa,1)
      gc <- rnorm(na,0,1)

      dh1 <- dhedges(ge,gc)
      d1[j] <- dh1

      j <- j + 1

   }

   deltasb <- rnorm(kb,DELTAB,sqrt(tau2b))
   nsb <- rchisq(kb,4)+averN-4

   d2 <- rep(NA,kb)
   mod2 <- rep(0,kb)

   j <- 1

   while (j <= kb) {

      deltab <- deltasb[j]
      nb <- nsb[j]

      ge <- rnorm(nb,deltab,1)
      gc <- rnorm(nb,0,1)

      dh2 <- dhedges(ge,gc)
      d2[j] <- dh2

      j <- j + 1

   }

   ###FIRST CATEGORY
   var1 <- dhedvar(na,d1)

   ##"DL" Estimator
   a1DL    <- rma(d1, var1, method="DL")
   tau21DL <- a1DL$tau2
   w1DL    <- 1/(var1+tau21DL)
```

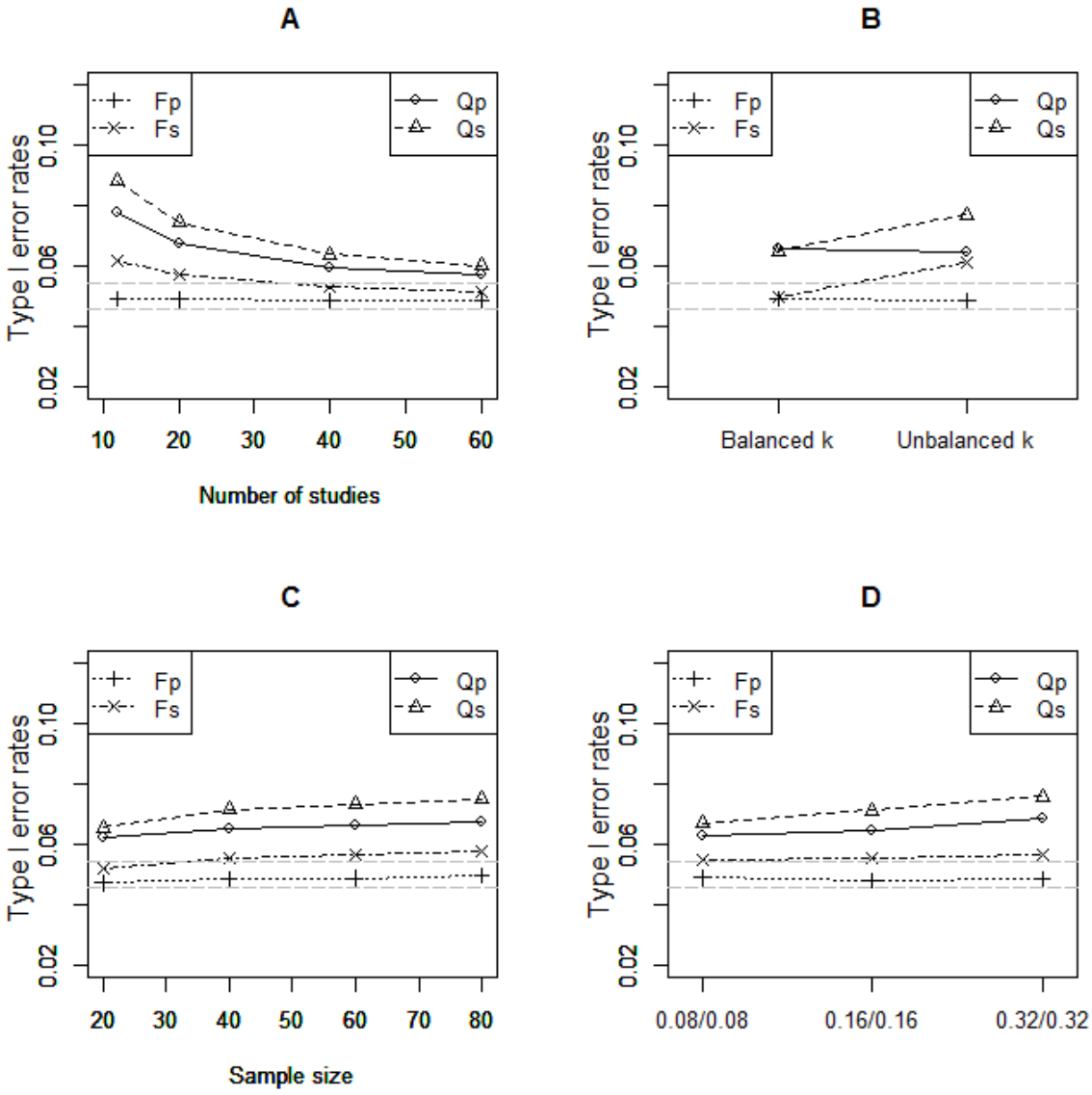
```
          ##"REML" Estimator
          a1REML    <- rma(d1, var1, method="REML")
          tau21REML <- a1REML$tau2
          w1REML    <- 1/(var1+tau21REML)

          ##"PM" Estimator
          a1PM    <- rma(d1, var1, method="PM")
          tau21PM <- a1PM$tau2
          w1PM    <- 1/(var1+tau21PM)

          ###SECOND CATEGORY
          var2 <- dhedvar(nb,d2)

          ##"DL" Estimator
          a2DL    <- rma(d2, var2,method="DL")
          tau22DL <- a2DL$tau2
          w2DL    <- 1/(var2+tau22DL)

          ##"REML" Estimator
          a2REML    <- rma(d2, var2, method="REML")
          tau22REML <- a2REML$tau2
          w2REML    <- 1/(var2+tau22REML)

          ##"PM" Estimator
          a2PM    <- rma(d2, var2, method="PM")
          tau22PM <- a2PM$tau2
          w2PM    <- 1/(var2+tau22PM)

          dt     <- c(d1,d2)
          vt     <- c(var1,var2)
          wtDL   <- c(w1DL,w2DL)
          wtREML <- c(w1REML,w2REML)
          wtPM   <- c(w1PM,w2PM)
          mod    <- c(mod1,mod2)

          ###Qb pooled with "DL"
          mDL <- rma(dt, vt, method="DL", mods=~factor(mod))

          if (mDL$QMp <= .05)
             p_Qp_DL[i] <- 1

          ###Qb pooled with "REML"
          mREML <- rma(dt, vt, method="REML", mods=~factor(mod))

          if (mREML$QMp <= .05)
             p_Qp_REML[i] <- 1

          ###Qb pooled with "PM"
          mPM <- rma(dt, vt, method="PM", mods=~factor(mod))

          if (mPM$QMp <=.05)
             p_Qp_PM[i] <- 1

          ###F pooled with "DL"
          nDL <- rma(dt, vt, method="DL", knha=T, mods=~factor(mod))

          if (nDL$QMp <=.05)
             p_Fp_DL[i] <- 1

          ###F pooled with "REML"
          nREML <- rma(dt, vt, method="REML", knha=T,
mods=~factor(mod))

          if (nREML$QMp <=.05)
             p_Fp_REML[i] <- 1

          ###F pooled with "PM"
          nPM <- rma(dt, vt, method="PM", knha=T, mods=~factor(mod))

          if (nPM$QMp <= .05)
```

```
      p_Fp_PM[i] <- 1

   ##Qb separate with "DL"
   qtest1sDL <- qtest(d1,w1DL)
   qtest2sDL <- qtest(d2,w2DL)
   qwsDL     <- qtest1sDL+qtest2sDL
   qtesttsDL <- qtest(dt,wtDL)
   QbsDL     <- qtesttsDL-qwsDL

   if ((1-pchisq(QbsDL,1)) <=.05)
      p_Qs_DL[i] <- 1

   ##Qb separate with "REML"
   qtest1sREML <- qtest(d1,w1REML)
   qtest2sREML <- qtest(d2,w2REML)
   qwsREML     <- qtest1sREML+qtest2sREML
   qtesttsREML <- qtest(dt,wtREML)

   QbsREML <- qtesttsREML-qwsREML

   if ((1-pchisq(QbsREML,1)) <= .05)
      p_Qs_REML[i] <- 1

   ##Qb separate with "PM"
   qtest1sPM <- qtest(d1,w1PM)
   qtest2sPM <- qtest(d2,w2PM)
   qwsPM     <- qtest1sPM+qtest2sPM
   qtesttsPM <- qtest(dt,wtPM)

   QbsPM <- qtesttsPM-qwsPM

   if ((1-pchisq(QbsPM,1)) <= .05)
      p_Qs_PM[i] <- 1

   ##F separate with "DL"
   FsDL <- QbsDL/(qwsDL/(k-2))

   if ((1-pf(FsDL,1,k-2)) <= .05)
      p_Fs_DL[i] <- 1

   ##F separate with "REML"
   FsREML <- QbsREML/(qwsREML/(k-2))

   if ((1-pf(FsREML,1,k-2)) <= .05)
      p_Fs_REML[i] <- 1

   ##F separate with "PM"
   FsPM <- QbsPM/(qwsPM/(k-2))

   if ((1-pf(FsPM,1,k-2)) <= .05)
      p_Fs_PM[i] <- 1

   i <- i + 1

}

Qpooled_DL     <- mean(p_Qp_DL)
Qseparate_DL   <- mean(p_Qs_DL)
Qpooled_REML   <- mean(p_Qp_REML)
Qseparate_REML <- mean(p_Qs_REML)
Qpooled_PM     <- mean(p_Qp_PM)
Qseparate_PM   <- mean(p_Qs_PM)
Fpooled_DL     <- mean(p_Fp_DL)
Fseparate_DL   <- mean(p_Fs_DL)
Fpooled_REML   <- mean(p_Fp_REML)
Fseparate_REML <- mean(p_Fs_REML)
Fpooled_PM     <- mean(p_Fp_PM)
Fseparate_PM   <- mean(p_Fs_PM)
```

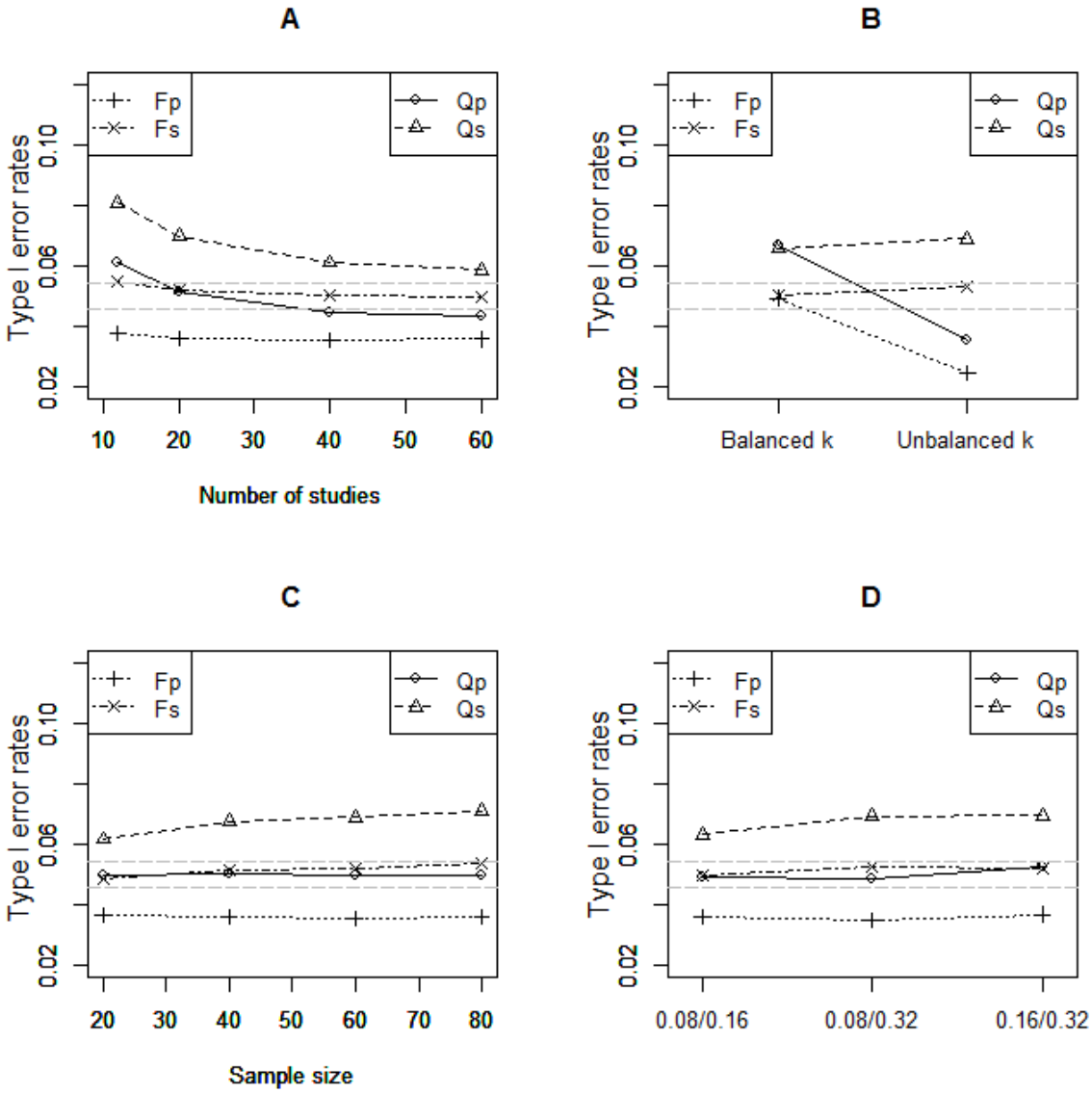
```
      resultados <- cbind(Qpooled_DL, Qseparate_DL, Qpooled_REML,
Qseparate_REML, Qpooled_PM, Qseparate_PM, Fpooled_DL, Fseparate_DL,
Fpooled_REML, Fseparate_REML, Fpooled_PM, Fseparate_PM)

      write.table(cbind(name, resultados), "Resultados_KH.dat", append
= TRUE, sep = " ", col.names = FALSE, row.names = FALSE, quote =
FALSE)

}
```

Appendix 5C

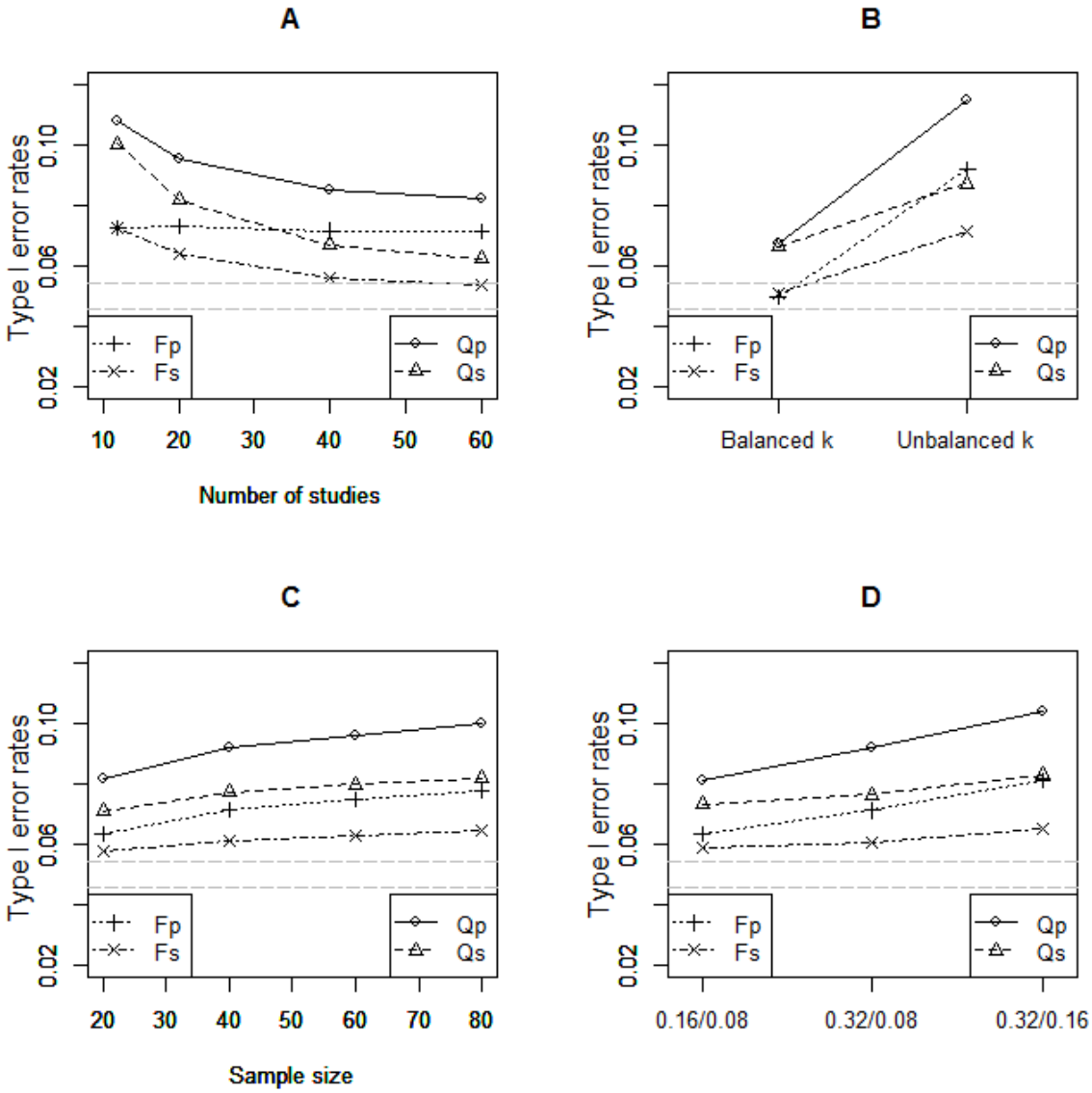
Supplementary figures using the DL and REML estimators



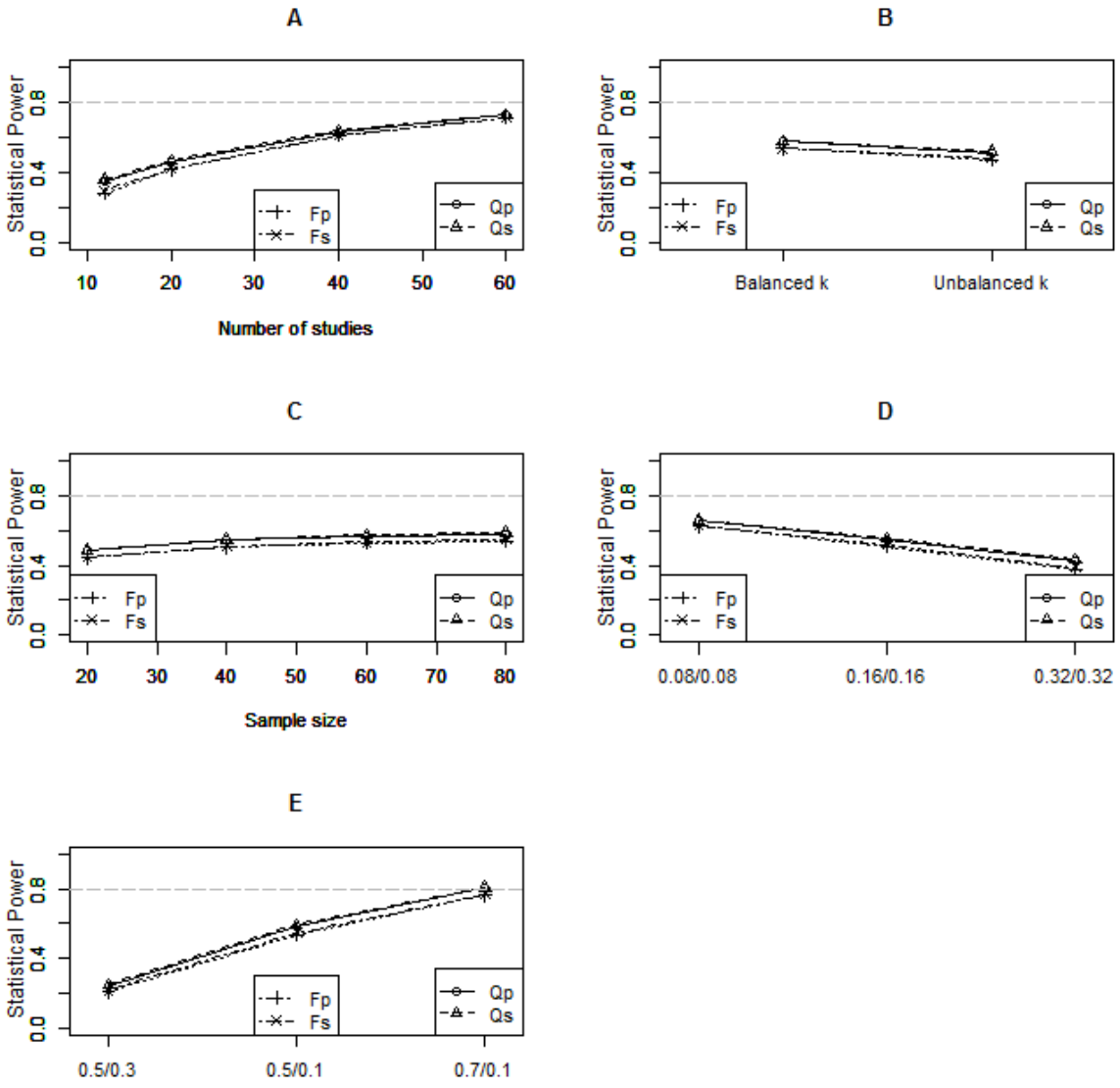
Supplementary Fig. 5C.1. Average Type I error rates in scenarios with homoscedastic residual between-studies variances across categories of the moderator using the DerSimonian and Laird estimator.



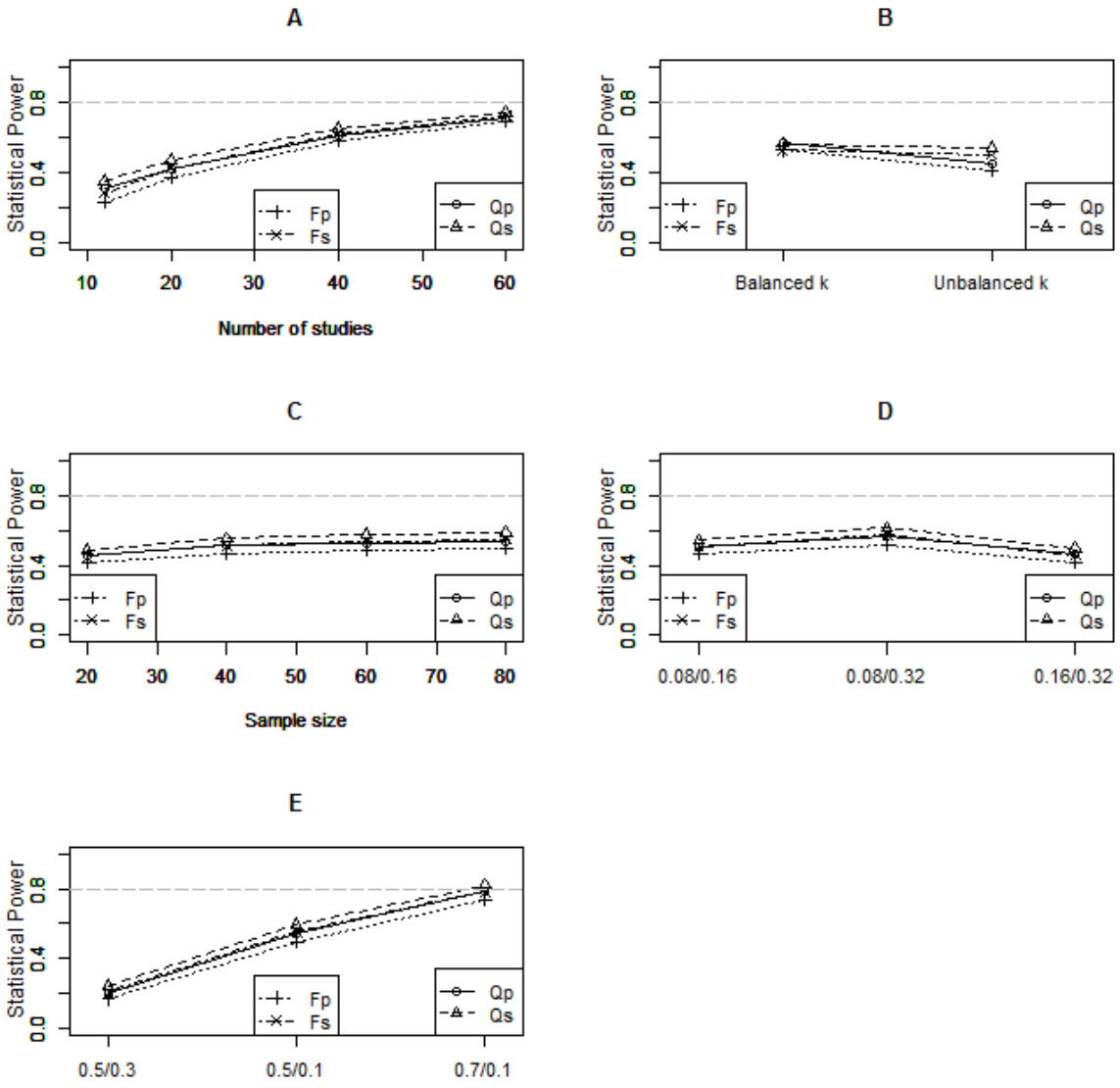
Supplementary Fig. 5C.2. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the categories of the moderator and smaller variance in the smaller category using the DerSimonian and Laird estimator.



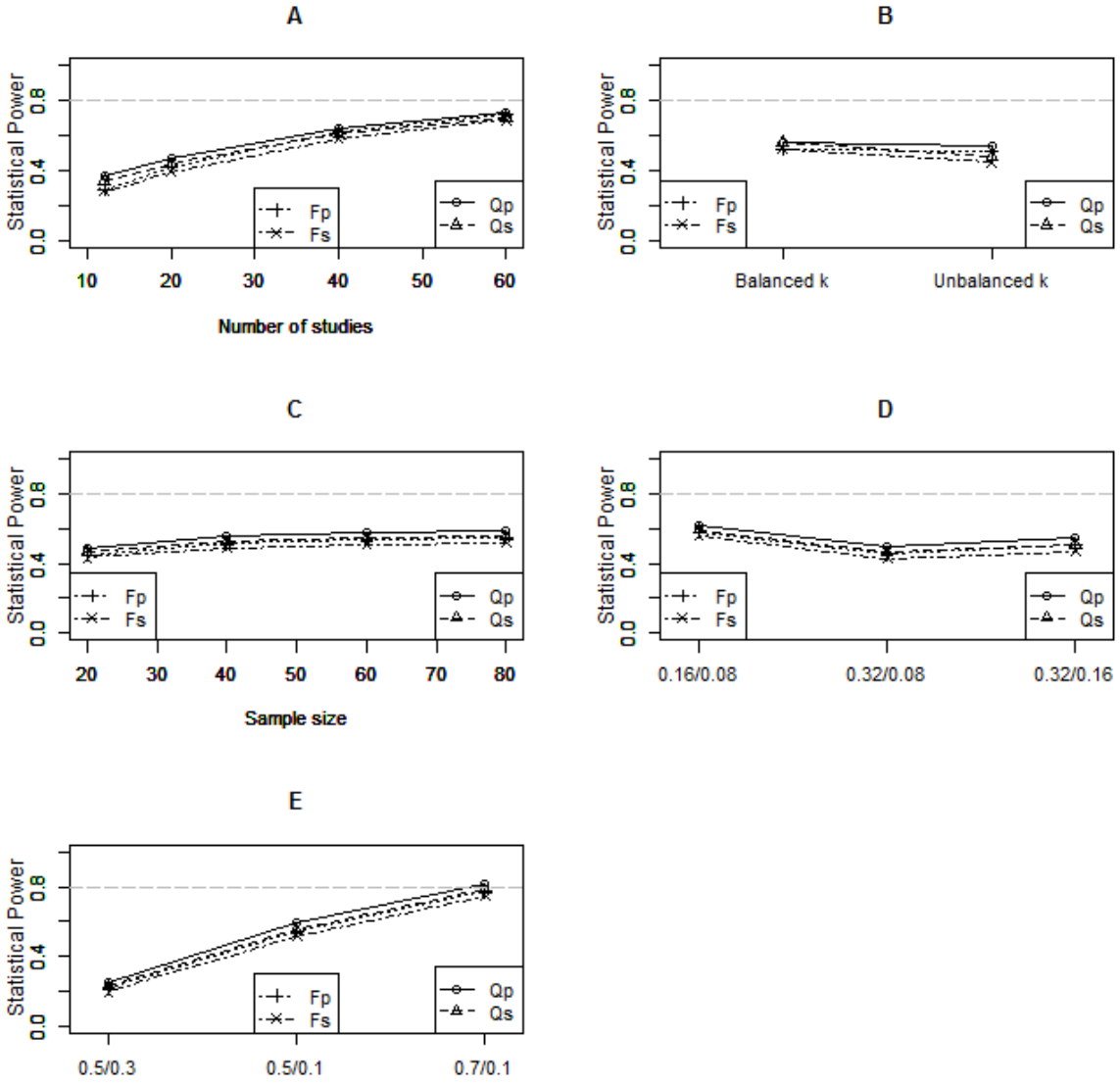
Supplementary Fig. 5C.3. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances in each category of the moderator and larger variance in the smaller category using the DerSimonian and Laird estimator.



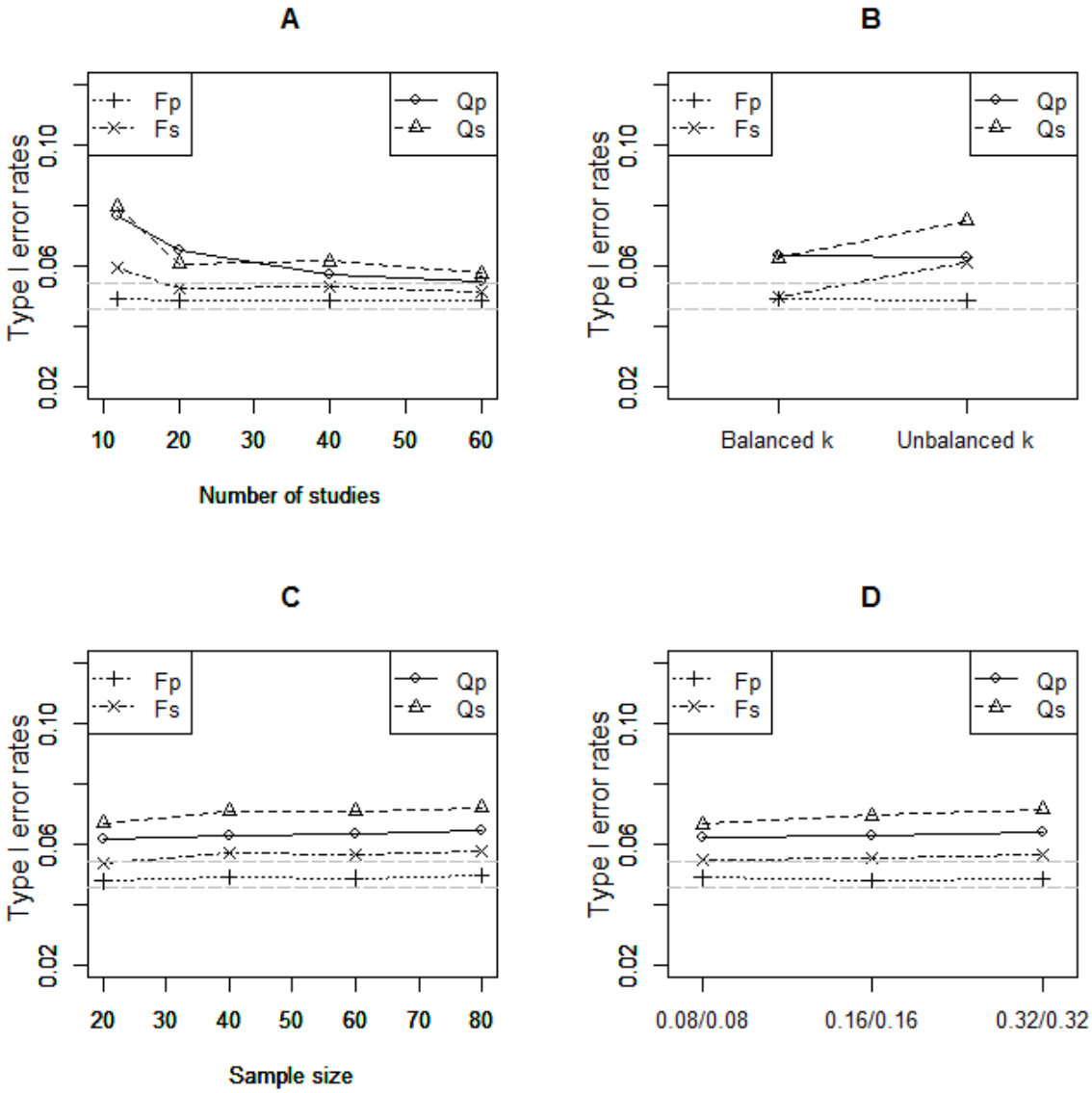
Supplementary Fig. 5C.4. Average power rates in scenarios with homoscedastic residual between-studies variances across categories of the moderator using the DerSimonian and Laird estimator.



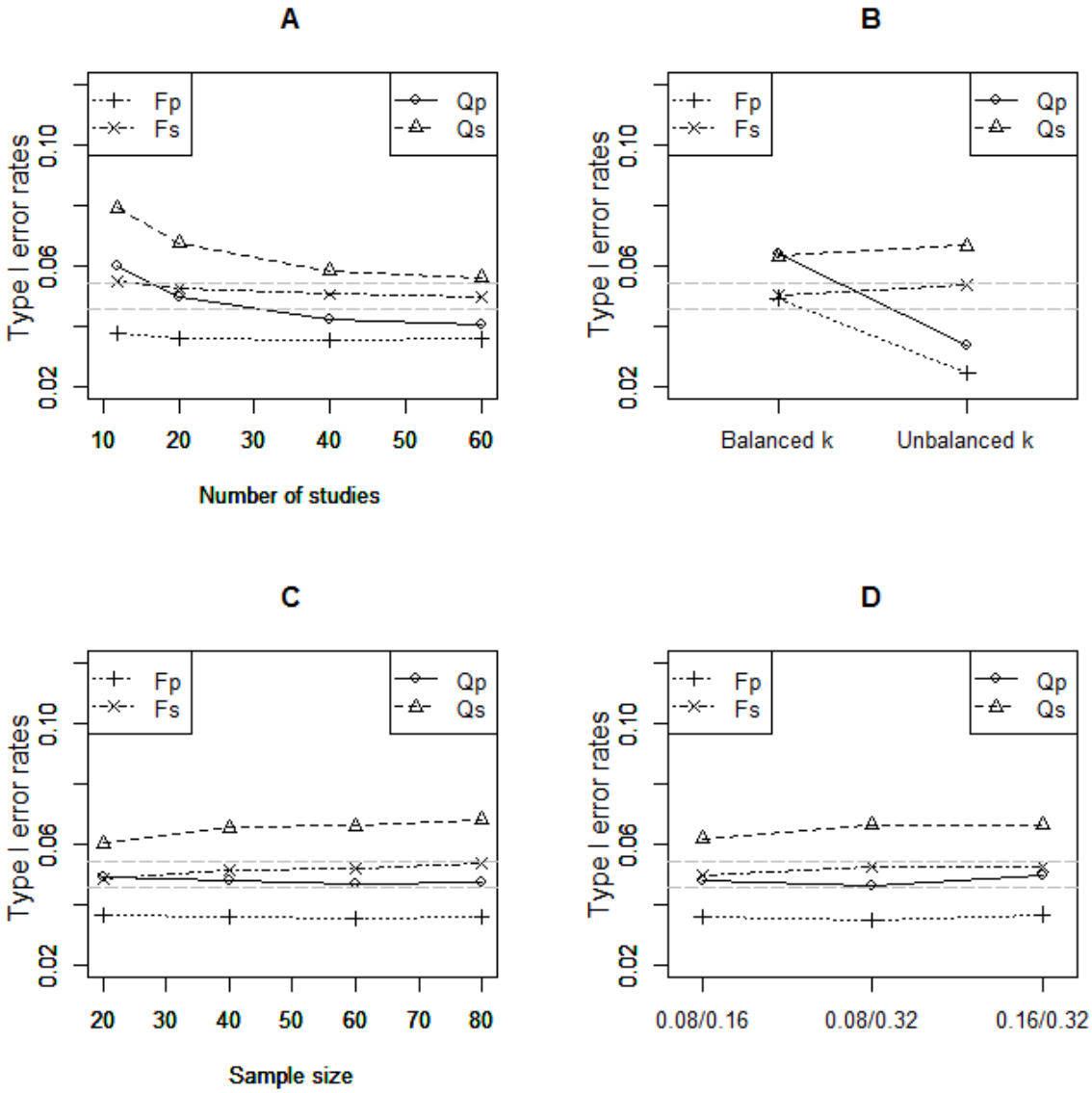
Supplementary Fig. 5C.5. Average power rates in scenarios with heteroscedastic residual between-studies variances across categories of the moderator and smaller variance in the smaller category using the DerSimonian and Laird estimator.



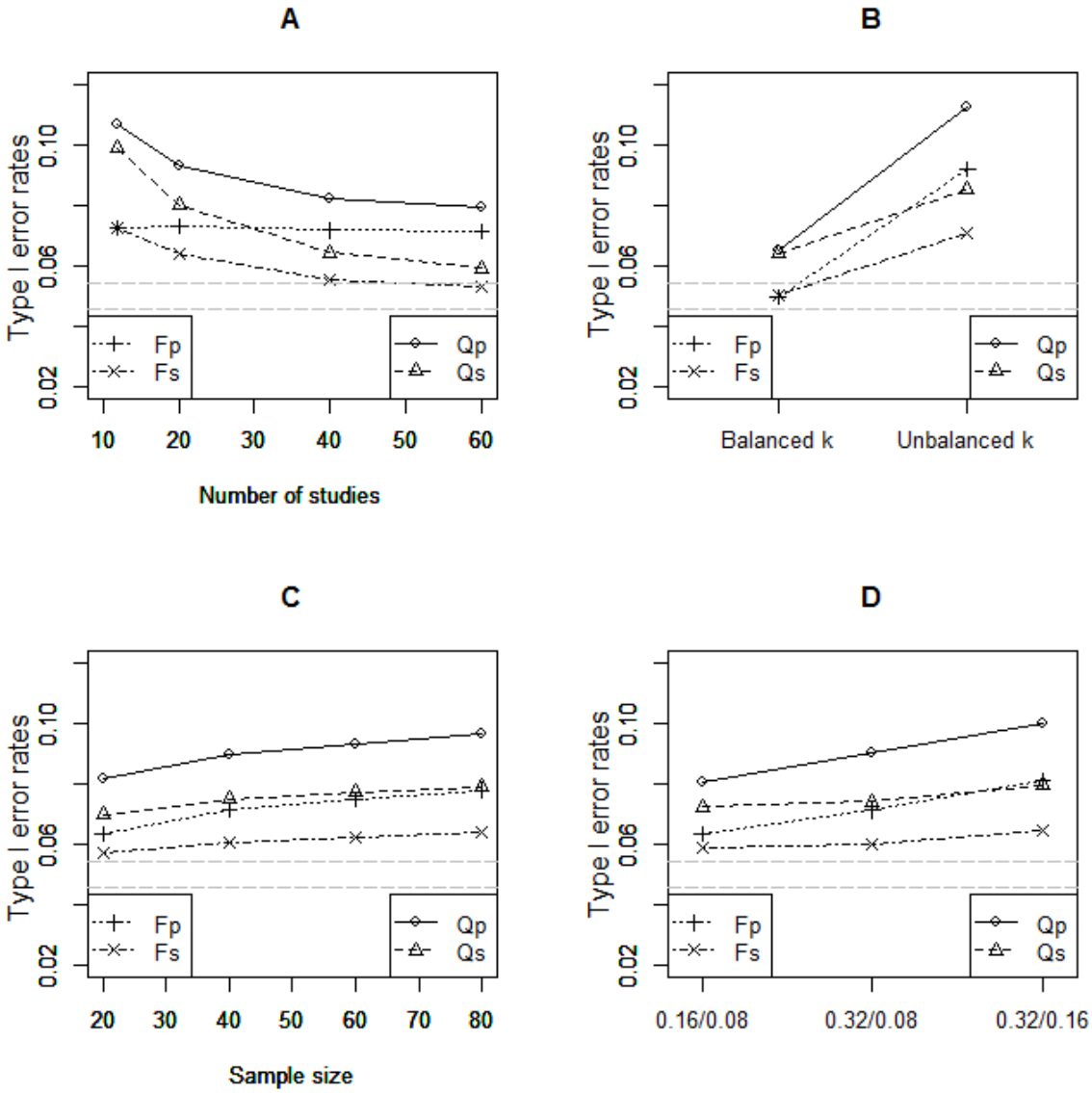
Supplementary Fig. 5C.6. Average power rates in scenarios with heteroscedastic residual between-studies variances across categories of the moderator and larger variance in the smaller category using the DerSimonian and Laird estimator.



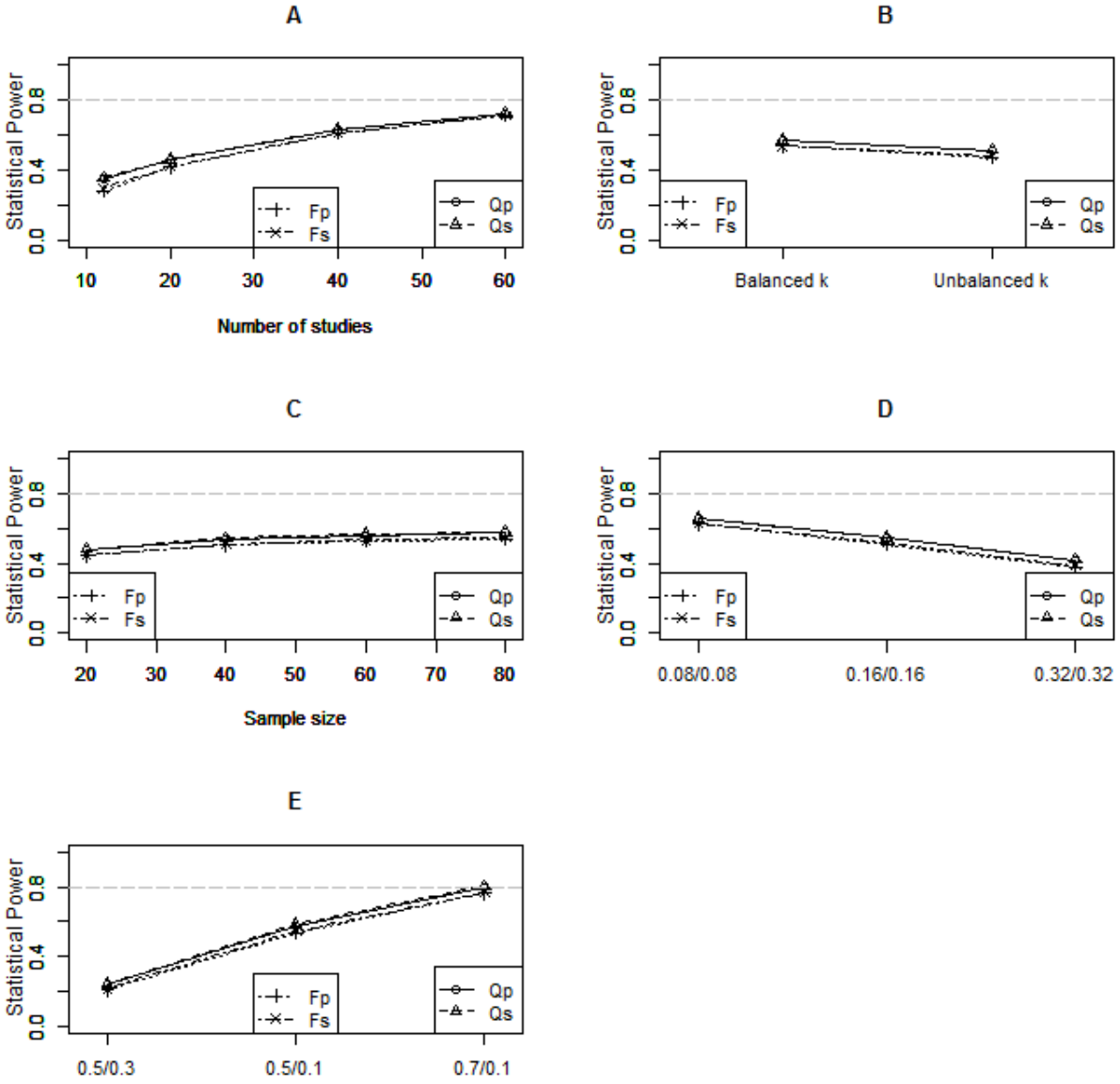
Supplementary Fig. 5C.7. Average Type I error rates in scenarios with homoscedastic residual between-studies variances across categories of the moderator using the restricted maximum likelihood estimator.



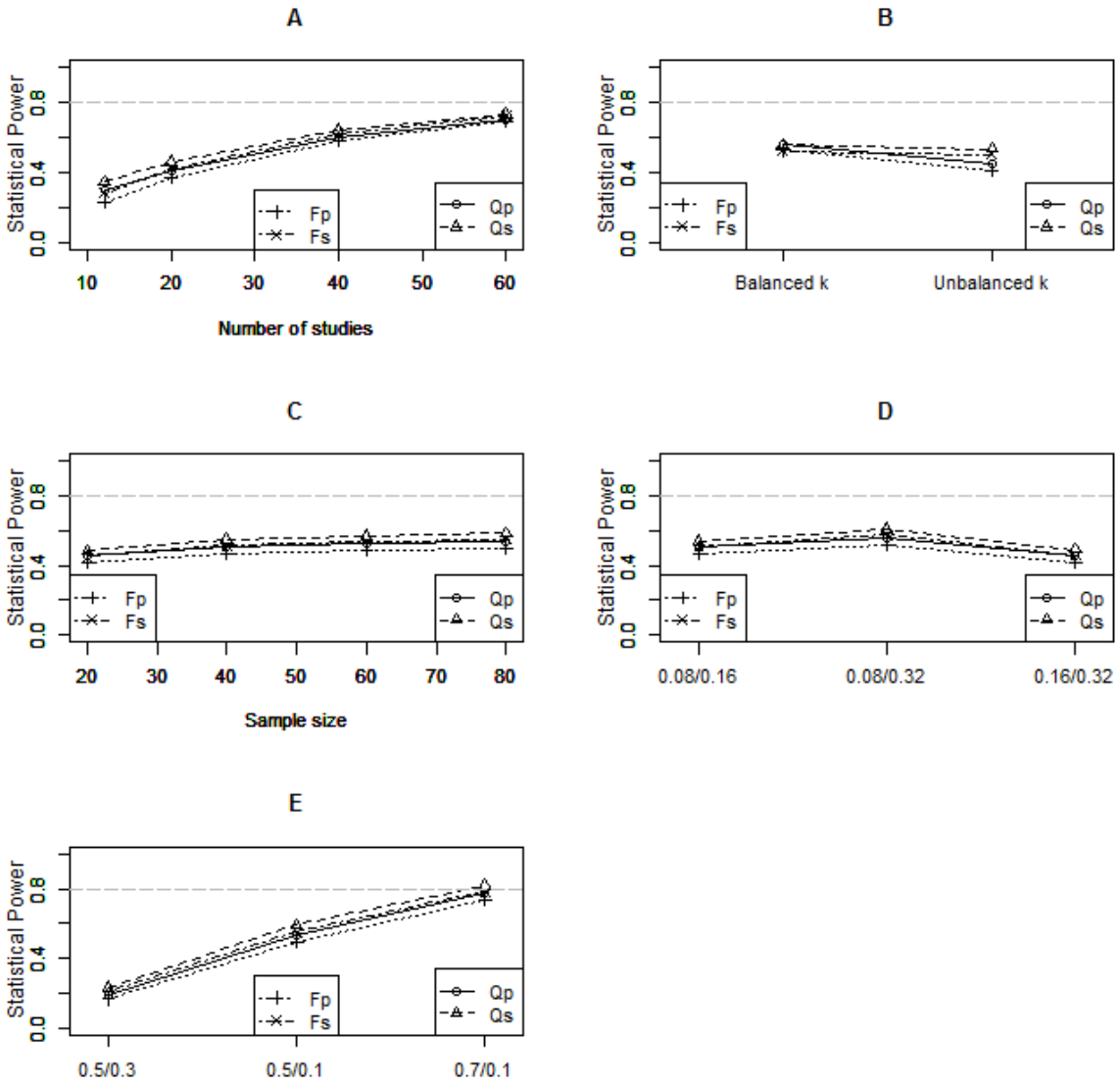
Supplementary Fig. 5C.8. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the categories of the moderator and smaller variance in the smaller category using the restricted maximum likelihood estimator.



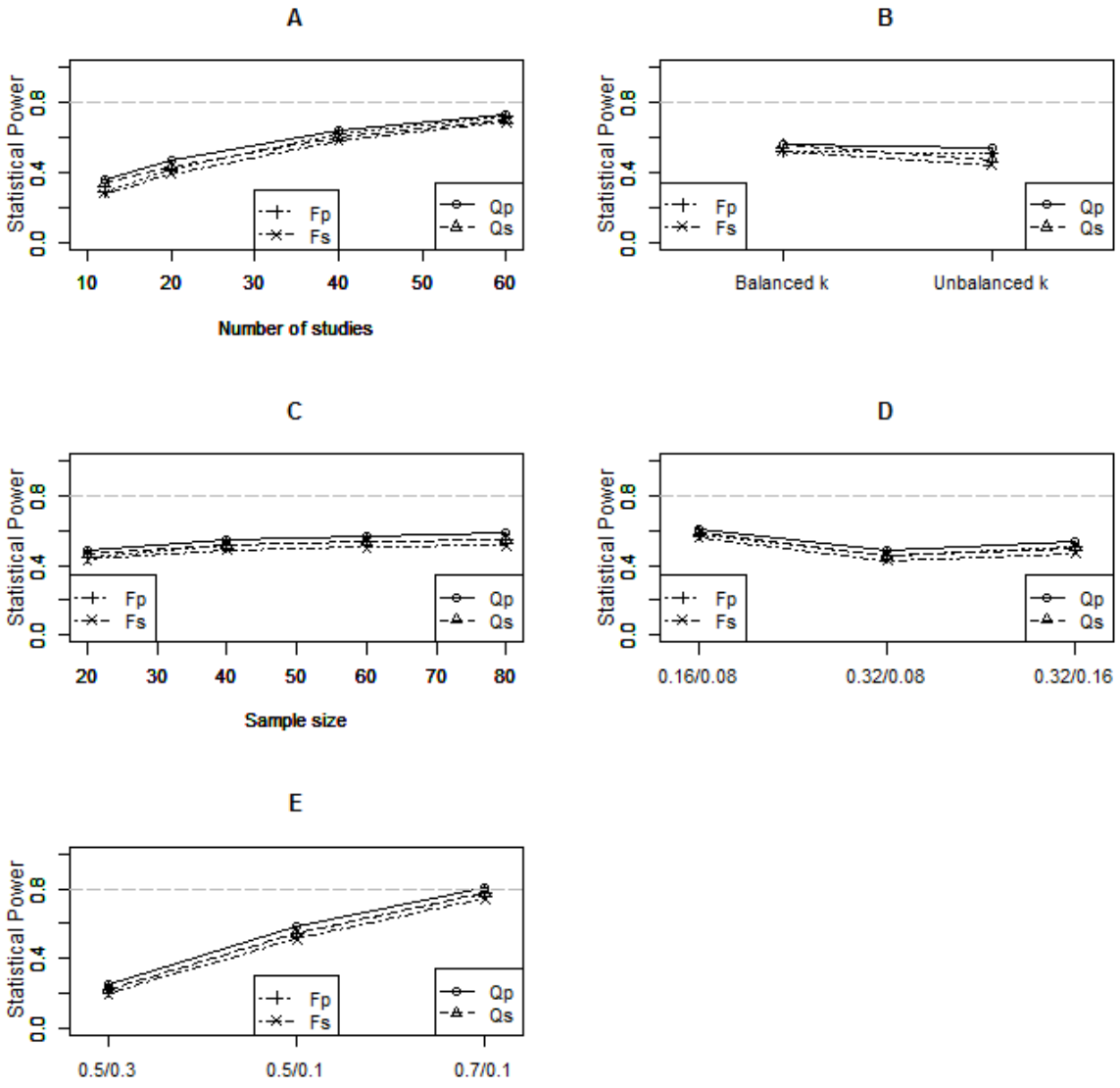
Supplementary Fig. 5C.9. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances in each category of the moderator and larger variance in the smaller category using the restricted maximum likelihood estimator.



Supplementary Fig. 5C.10. Average power rates in scenarios with homoscedastic residual between-studies variances across categories of the moderator using the restricted maximum likelihood estimator.



Supplementary Fig. 5C.11. Average power rates in scenarios with heteroscedastic residual between-studies variances across categories of the moderator and smaller variance in the smaller category using the restricted maximum likelihood estimator.



Supplementary Fig. 5C.12. Average power rates in scenarios with heteroscedastic residual between-studies variances across categories of the moderator and larger variance in the smaller category using the restricted maximum likelihood estimator.