UNIVERSIDAD DE MURCIA

FACULTAD DE PSICOLOGÍA

The Mixed-Effects Model for the Detection of Moderator Variables in Meta-analysis: A Comparison of Procedures through the Monte Carlo Method

---

El Modelo de Efectos Mixtos para la Detección de Variables Moderadoras en Meta-análisis: una Comparación de Procedimientos a través del Método Monte Carlo

José Antonio López López

2012

UNIVERSITY OF MURCIA

FACULTY OF PSYCHOLOGY

The Mixed-Effects Model for the Detection of Moderator Variables in Meta-analysis: A Comparison of Procedures through the Monte Carlo Method

---

El Modelo de Efectos Mixtos para la Detección de Variables Moderadoras en Meta-análisis: una Comparación de Procedimientos a través del Método Monte Carlo

Doctoral Thesis

Author: José Antonio López López

Supervisors: Dr. Julio Sánchez-Meca and Dr. Fulgencio Marín-Martínez

Murcia, 2012

Trabajo financiado con una Beca-Contrato Predoctoral

de la Fundación Séneca, Agencia de Ciencia

y Tecnología de la Región de Murcia

"Before what has been found can be used, before it can persuade skeptics, influence policy, affect practice, it must be known. Someone must organize it, integrate it, extract the message. A hundred dissertations are mute. Someone must read them and discover what they say" (Gene V. Glass, 1976, p. 4)

Agradecimientos

A mis directores de tesis, los doctores **Julio Sánchez** y **Fulgencio Marín**, por el apoyo, dedicación y entusiasmo mostrados a lo largo de este trabajo. Me considero muy afortunado por haber tenido la oportunidad de trabajar y, sobre todo, aprender con ellos en lo profesional y en lo personal durante estos años.

A mis compañeros de la Facultad de Psicología, los doctores **Pedro Fernández**, **Óscar Hernández**, **Federica Sassi**, **Vanesa Valero**, **Victoria Plaza**, **Guillermo Campoy** y **Luis Fuentes**; y también a los que, como yo, aún no son doctores, **Violeta Pina**, **Lucía Colodro**, **Aurora Orenes**, **Violeta Provencio** y **Rubén Rodríguez**. Ellos (entre otros) han llenado de alegría y complicidad mi lugar de trabajo durante estos cuatro años.

A mis compañeros de área y/o de la Unidad de Meta-análisis, los doctores **José Antonio López Pina**, **Manuel Ato**, **Juan José López**, **Mª Dolores Hidalgo**, **Ana Benavente**, **Ana Isabel Rosa**, **Antonia Gómez**, **Rafael Rabadán** y **Antonio Velandrino**, por la confianza y cariño que me han demostrado desde el primer día en que comencé a trabajar aquí.

Por último, y muy especialmente, por la comprensión y el apoyo incondicionales sin los que nunca habría sido capaz de culminar este proyecto, ***a mi familia***.

Index

Preface

Describing what mixed-effects meta-regression models are, as well as the implications of the empirical work conducted for this dissertation, are not easy tasks, as I needed seven chapters to address those issues. A previous step, however, is to define what a dissertation is, and that did not prove to be simpler to me. After working on this project for several years, I cannot summarize such a long process in a short sentence. Therefore, I will detail several considerations based on my own experience along this section, with the aim to provide a general picture of how this dissertation was planned and carried out.

A first possibility is to conceive this dissertation as the product of several decisions. The first important decision that I made concerning this project was to become a bachelor student in Psychology in 2003. One of the first courses that I took was *Methodology for Psychological Research*, in which Dr. Juan José López García provided a first approximation to the systematic measurement and analysis of the human behavior. In the second semester, along the subject *Data Analysis in Psychology*, Dr. Fulgencio Marín Martínez presented many statistical tools for the descriptive analysis of data from a sample of

subjects. By then, I was already interested in this quantitative approach to the psychological field, guided by objective measurement and replicability.

In the second year, one of the most interesting courses for me was *Statistical Models in Psychology*, along which Dr. Julio Sánchez Meca showed how sample information can be generalized to broader groups of subjects through statistical inference techniques. One year later, I took the courses of *Psychometrics*, where Dr. María Dolores Hidalgo Montesinos and Dr. José Antonio López Pina introduced us to tests and their properties and underlying models, and *Research Designs in Psychology*, along which Dr. Manuel Ato García provided us with a wider perspective of how psychological experiments can be conducted. After my first three years, I had already decided to get a PhD in the knowledge area of Methodology of the Behavioral Sciences, and I have kept in the pursuit of that goal until these days.

This dissertation can also be regarded as the product of the effort done not only by the author, but also by several scholars who collaborated with me along the process. I started to get familiar with meta-analysis, the main topic of my dissertation, when I became an internal student with Dr. Fulgencio Marín Martínez in 2004. Along the next three years, we read and discussed several papers and books on meta-analysis, and I was gradually involved in a meta-analytic review about the efficacy of psychological treatments for patients with obsessive-compulsive disorder, headed by Dr. Julio Sánchez Meca, and published afterwards in a prestigious scientific journal such as Clinical Psychology Review (Rosa-Alcázar, Sánchez-Meca, Gómez-Conesa, & Marín-Martínez, 2008). My collaboration with both of them increased in the last year of my degree, with several meetings every week, and one of the first consequences of this joint effort was my Degree Thesis (López-López, 2008).

Apart from the co-promotors of this dissertation, with whom I have always kept in touch along these years, I also had the chance to go abroad several times and to work under the supervision of different experts on meta-analysis. In 2010, I did my first

internship at the Maastricht University (The Netherlands) to work with Dr. Wolfgang Viechtbauer, and this collaboration helped me to improve my technical skills regarding software for simulation studies and was very important to develop one of the empirical studies of this dissertation, presented in Chapter 5. My second internship took place in 2011 at the Peabody Research Institute in Nashville, Tennessee (United States of America) under the supervision of Dr. Mark W. Lipsey and some members of his team such as Dr. Sandra Jo Wilson and Dr. Emily Tanner-Smith. The fact that this research group is more focused on applied than methodological work helped me to adapt my simulated scenarios to more realistic conditions, and to be more aware of which methodological advances are currently needed in Psychology. Finally, in 2012 I went to the Catholic University of Leuven (Belgium) to work under the supervision of Dr. Eva Ceulemans, and this experience allowed me to improve my technical skills regarding presentation of results and to get familiar with multilevel models, which are becoming very important in my research field.

Last, but not least, this dissertation can be regarded as the product of a specific context. Two elements can be remarked due to the crucial influence that they exerted on the topic and on the feasibility of the project itself. The first of them is The Meta-analysis Unit, headed by Dr. Julio Sánchez Meca, who has been doing research on meta-analysis since the early 80s, and whose expertise and orientation helped me to find a relevant and useful topic for current science. The second one is the Fundación Séneca, Agencia de Ciencia y Tecnología de la Región de Murcia, which sponsored this project and allowed me to work full time on the development of this dissertation since 2009. They also sponsored all of my internships (eleven months in total), giving me the chance to collaborate with other research groups and to enrich my education and the quality of my research.

I think that a dissertation is a research project that can be conducted in many different ways. I hope that the previous paragraphs can illustrate how this dissertation was carried out, and I hope that they can reflect how fascinating this process was to me, and how grateful I am to everyone that made it possible.

The first chapter of this dissertation is an introduction, which is not intended to be exhaustive. The reader interested can easily find a vast amount of books and papers focusing on the conceptual and technical issues of meta-analysis, many of them cited along Section 1.1. Therefore, the purpose of this first chapter is to present the main ideas concerning this methodology, with special attention given to the topics that are direct and indirectly addressed in the empirical part of the dissertation. In Chapter 2, several outcome variables in meta-analysis will be described, focusing on the ones that were employed in the empirical part of this dissertation. In Chapter 3, mixed-effects meta-regression models are detailed, along with some alternative methods available for estimating and testing the most relevant parameters. The fact that different methods are available when fitting such models poses a problem to the meta-analyst, since the method choice might have an influence on the results.

The empirical part of this dissertation includes three simulation studies comparing different methodological alternatives when fitting mixed-effects meta-regression models, and spans Chapters 4 to 7. In Chapter 4, seven methods for the estimation of the heterogeneity variances and the model predictive power are compared. In Chapter 5, the influence of seven heterogeneity variance estimators and six methods to test the model coefficients is assessed. Chapter 6 constitutes an application of some of the methods compared before to the reliability generalization approach, which entails working with reliability coefficients as outcome variables. Finally, some general conclusions, limitations of the studies here presented, and implications for future research are provided in Chapter 7.

Resumen

La producción científica ha crecido exponencialmente en las últimas décadas en prácticamente todos los campos y disciplinas. Esta situación ha hecho necesario que los investigadores diseñen métodos para la síntesis eficiente del conocimiento. De entre ellos destaca el meta-análisis, que surge en el ámbito de la Psicología (Glass, 1976). El meta-análisis es una metodología de revisión sistemática de la investigación basada en criterios objetivos y caracterizada por la aplicación de métodos cuantitativos.

El meta-análisis constituye un gran avance respecto a las revisiones narrativas tradicionales en términos de precisión, fiabilidad y validez (Cooper y Hedges, 2009b). Desde que fue inicialmente propuesto por Gene V. Glass (1976), esta metodología ha sido desarrollada y ampliamente aplicada en multitud de disciplinas como las Ciencias del Comportamiento, Biológicas y de la Salud (e.g., Cooper, Hedges y Valentine, 2009; Marín-Martínez, Sánchez-Meca y López-López, 2009). En un meta-análisis, el primer reto en los análisis estadísticos suele consistir en escoger y calcular un índice del tamaño del efecto que permita presentar los resultados de los estudios individuales en una métrica común. Posteriormente, lo habitual es que cada tamaño del efecto se pondere por una función de su precisión (e.g., Pigott, 2001), de manera que los valores más precisos tengan una mayor

influencia en los resultados globales. El Capítulo 2 de esta Tesis Doctoral está dedicado a los diferentes índices del tamaño del efecto, con especial atención a los que se emplearon en los capítulos empíricos posteriores.

Una vez que los resultados individuales son directamente comparables, el meta-analista puede calcular un promedio de los efectos, que suele complementarse con una evaluación de la heterogeneidad entre los resultados integrados. En caso de que aparezcan discrepancias entre los resultados de los estudios individuales, el meta-análisis permite al investigador efectuar una búsqueda de variables moderadoras que puedan explicar al menos parte de esta variabilidad. Los análisis de moderadores han cobrado una gran importancia a lo largo de las últimas décadas, ya que en la práctica es muy habitual encontrar inconsistencias entre los efectos estimados en los diferentes estudios. Una de las alternativas más empleadas en la actualidad para llevar a cabo estos análisis de moderadores son los llamados modelos de meta-regresión de efectos mixtos.

En la actualidad, existen diferentes métodos para la estimación y el contraste de la significación de los parámetros de un modelo de meta-regresión de efectos mixtos. Esta situación puede resultar problemática, ya que la elección del método estadístico podría afectar a los resultados y conclusiones de un meta-análisis. La presente Tesis Doctoral incluye un total de tres estudios de simulación Monte Carlo donde se compararon diferentes métodos para el ajuste de modelos de meta-regresión de efectos mixtos con un moderador, con la finalidad de guiar las decisiones de los investigadores en función de las condiciones concretas de aplicación (número de estudios del meta-análisis, tamaño muestral medio de los estudios y características de la distribución de los efectos paramétricos y de los tamaños muestrales).

En el tercer capítulo de esta Tesis Doctoral se presentarán los métodos comparados a lo largo de los estudios empíricos para la estimación y contraste de la significación de los parámetros más relevantes en los modelos de meta-regresión de efectos mixtos. Uno de estos parámetros es la varianza inter-estudios residual, que

representa la cantidad de heterogeneidad entre los resultados individuales (distinta del error de muestreo aleatorio) no explicada tras incorporar uno o más moderadores al modelo (Viechtbauer, 2008). En la Sección 3.2 de este trabajo se describirán siete estimadores de este parámetro. Dado que la varianza inter-estudios residual es uno de los elementos del factor de ponderación (de los tamaños del efecto) en un modelo de efectos mixtos, obtener estimaciones precisas de este parámetro supone un aspecto importante. Otro análisis es el contraste de la significación de los moderadores incluidos en el modelo, para el cual se presentarán seis alternativas metodológicas en la Sección 3.3 de la presente Tesis Doctoral. Por último, en cuanto a la estimación de la potencia predictiva del modelo, la Sección 3.4 se centrará en la propuesta de Raudenbush (1994) para modelos meta-analíticos, basada en la re-estimación de la varianza inter-estudios tras la inclusión de uno o más predictores en el modelo. La existencia de hasta siete estimadores de la varianza inter-estudios supone que existen (al menos) siete métodos alternativos para calcular la potencia predictiva en los modelos que se estudian en este trabajo.

Dado el gran número de alternativas disponibles para el ajuste de modelos de meta-regresión de efectos mixtos, un primer objetivo general de esta Tesis Doctoral fue el de analizar hasta qué punto difieren los resultados en función del método empleado, con el fin de determinar qué alternativas son preferibles dadas unas condiciones determinadas. Para ello, se llevaron a cabo tres estudios de simulación Monte Carlo, y cada uno de ellos incorporó un amplio espectro de condiciones realistas en Psicología y otros ámbitos relacionados. Un segundo objetivo general de este trabajo consistía en comprobar si existen condiciones bajo las cuales el método estadístico seleccionado no afecta a los resultados. Por una parte, se esperaba que ninguno de los métodos comparados mostrase un funcionamiento apropiado bajo las condiciones más adversas. Por otra parte, se esperaba que todos los métodos tenderían a ofrecer resultados convergentes (y precisos) cuando las condiciones de aplicación fuesen óptimas o con un número suficiente de estudios y de unidades por estudio en el meta-análisis.

En el primer estudio de simulación, presentado en el Capítulo 4, se encontraron algunas diferencias en el funcionamiento de los diferentes estimadores de la varianza inter-estudios, con tendencias similares para los diferentes métodos tanto en la estimación de la varianza inter-estudios total como en la residual (es decir, tras la adición de uno o más moderadores al modelo). En un extremo, los métodos de Hunter y Schmidt (HS), máxima verosimilitud (ML) y de Sidik y Jonkman (SJ) proporcionaron estimaciones negativamente sesgadas para la varianza inter-estudios (total y residual), mientras que el método de Hedges (HE) se mostró insesgado aunque con una baja eficiencia relativa en comparación con los demás estimadores. En el otro extremo, los estimadores de DerSimonian y Laird (DL), máxima verosimilitud restringida (REML) y el estimador empírico de Bayes (EB) mostraron mejores resultados, aunque se observó un sesgo negativo en el primero de ellos para los valores más altos del parámetro. Estos resultados sugieren que los métodos REML y EB constituyen opciones adecuadas para la estimación de la varianza inter-estudios (total y residual) en modelos meta-analíticos. El número de estudios ejerció una clara influencia en los resultados, y ningún método alcanzó estimaciones precisas con menos de 20 estudios. En contraste con lo anterior, se obtuvieron estimaciones precisas con 80 estudios para todos los métodos y sin importar los restantes factores manipulados.

Un objetivo adicional en el estudio presentado en el Capítulo 4 era el de analizar el rendimiento de los diferentes métodos para la estimación de la potencia predictiva en modelos de meta-regresión de efectos mixtos, siguiendo la propuesta de Raudenbush (1994). De nuevo, los métodos HS, ML, SJ y HE proporcionaron los resultados menos precisos, mientras que los estimadores DL, REML y EB se mostraron como los más apropiados. Dentro de este grupo, el estimador EB alcanzó los mejores resultados al combinar los criterios de sesgo, tasa de truncamientos a 0 y 1 y eficiencia. El número de estudios emergió de nuevo como el factor más influyente para todos los métodos, y al menos 40 estudios fueron necesarios para que las estimaciones obtenidas con los diferentes métodos fuesen precisas.

El segundo estudio de simulación, descrito en el Capítulo 5, comparó el funcionamiento de diferentes métodos para el contraste de moderadores en modelos de meta-regresión de efectos mixtos. La elección del estimador de la varianza inter-estudios residual apenas alteró los resultados, pero sí que se encontraron discrepancias importantes en función del método aplicado para el contraste de la significación estadística de los coeficientes de regresión. En algunos trabajos anteriores se argumentó que el método tradicional para el contraste de los coeficientes de estos modelos, que asume una distribución normal para los coeficientes paramétricos, no incorpora la incertidumbre derivada del proceso de estimación de las varianzas muestrales, lo cual podría dar lugar a la obtención de resultados erróneos (e.g., Hardy y Thompson, 1996; Henmi y Copas, 2010). Cuando se examinó su rendimiento en este estudio, el método tradicional mostró un inadecuado control de la tasa de error Tipo I, dando lugar a rechazos incorrectos de la hipótesis nula.

De entre las distintas alternativas al método tradicional examinadas en el Capítulo 5, el procedimiento propuesto por Knapp y Hartung (2003) se mostró como una opción idónea, debido a su simplicidad de cálculo y a las adecuadas tasas empíricas de error Tipo I encontradas al aplicarlo. Hay que destacar, no obstante, que este método mostró un mejor rendimiento sin el truncamiento propuesto por los autores, el cual condujo a una pérdida de potencia estadística. El método de Huber-White y el test de razón de verosimilitudes, que también fueron incluidos en esta comparación, no mostraron un control apropiado de la tasa de error Tipo I. Finalmente, el test de permutaciones funcionó de manera similar al método de Knapp y Hartung no truncado. Aunque este último sería preferible en la mayoría de situaciones, el test de permutaciones representa una alternativa apropiada cuando no sea posible asumir que los estudios del meta-análisis han sido seleccionados mediante un proceso de muestreo aleatorio (Manly, 1997). Por otra parte, fueron necesarios en torno a 40 estudios para que los diferentes métodos alcanzasen tasas de potencia cercanas a 0.80, tal y como recomendó Jacob Cohen (1988).

Los estudios presentados en los Capítulos 4 y 5 se centraron en una variable dependiente normalmente distribuida, la diferencia entre medias estandarizada. Por el contrario, el último estudio de simulación de esta Tesis Doctoral, presentado en el Capítulo 6, exploró algunas variables dependientes en meta-análisis dentro del enfoque de generalización de la fiabilidad. En este estudio, se llevó a cabo una comparación de métodos para la estimación de los coeficientes del modelo de meta-regresión y el contraste de la significación de moderadores. En cuanto a las variables dependientes, el coeficiente alfa, que tiene una distribución muestral asimétrica, fue comparado con tres transformaciones normalizadoras. Los resultados sólo mostraron ligeras discrepancias para las diferentes variables dependientes. Por lo que respecta a los métodos estadísticos para el contraste de moderadores, las tendencias fueron similares a las descritas en el Capítulo 5: los resultados fueron casi idénticos con los diferentes estimadores de la varianza inter-estudios residual, mientras que el método de Knapp y Hartung no truncado mejoró el rendimiento del método tradicional en términos de tasa empírica de error Tipo I y tasa de potencia estadística. De nuevo, más de 30 estudios fueron necesarios para que los métodos alcanzasen tasas de potencia estadística satisfactorias.

La interpretación conjunta de los hallazgos de los tres estudios de simulación permite desgranar varias conclusiones concernientes a los modelos de meta-regresión de efectos mixtos. En primer lugar, el método escogido para la estimación de la varianza inter-estudios residual no mostró un influjo en los resultados del contraste de la significación estadística de los coeficientes de regresión (tampoco para variables dependientes con distribución muestral asimétrica), pero sí en la estimación de la potencia predictiva de estos modelos utilizando la propuesta de Raudenbush (1994); en este apartado, los estimadores DL, REML y (especialmente) EB proporcionaron los resultados más precisos.

Otra de las conclusiones alcanzadas a la luz de los resultados de las simulaciones de este trabajo está relacionada con el método para el contraste de los coeficientes en un

modelo de meta-regresión de efectos mixtos. La prueba *z* tradicionalmente empleada para el contraste de la significación estadística de moderadores en estos modelos mostró resultados poco precisos, mientras que la aplicación del método de Knapp y Hartung (2003) no truncado mejoró los resultados de manera consistente. Estas tendencias se mantuvieron cuando la variable dependiente empleada en los análisis tenía una distribución muestral asimétrica. Según estos resultados, el uso del método de Knapp y Hartung no truncado debería generalizarse cuando se contraste estadísticamente la asociación de un moderador con los tamaños del efecto mediante modelos de meta-regresión de efectos mixtos.

Por último, en cada una de las simulaciones se manipularon varios factores. De ellos, el número de estudios se mostró invariablemente como un factor crucial para obtener resultados precisos en modelos de meta-regresión de efectos mixtos; en concreto, los resultados de los estudios de esta Tesis Doctoral sugieren que se requieren alrededor de 40 estudios para poder llevar a cabo estos análisis con ciertas garantías, mientras que la interpretación de los resultados debería ser muy prudente en las situaciones donde el número de estudios incluidos en el meta-análisis esté por debajo de esta cifra. En cuanto a los demás factores, un mayor número de participantes por estudio en promedio conllevó la obtención de resultados más precisos, mientras que el grado de heterogeneidad entre los efectos paramétricos de cada meta-análisis mostró una influencia desigual en función del objetivo concreto: una mayor heterogeneidad entre efectos paramétricos afectó negativamente a la precisión de los contrastes de la significación de moderadores, al tiempo que mejoró los resultados en la estimación de la potencia predictiva de los modelos objeto de estudio en esta Tesis Doctoral.

Chapter 1

Introduction

1.1 Meta-analysis

Research production has exponentially grown along the last decades. Nowadays, it is common to find a great amount of studies analyzing the same phenomena in most scientific fields (e.g., Hedges, 2007). The reasons for this fact are diverse. As Cooper and Hedges (2009a) stated, "multiple studies on the same problem or hypothesis arise because investigators are unaware of what others are doing, because they are skeptical about the results of past investigations, or because they wish to extend (that is, generalize or search for influences on) previous findings. Experience has shown that even when considerable effort is made to achieve strict replication, results across studies are rarely identical at any high level of precision" (p. 4). The vast amount of scientific work published poses a problem of how to organize and summarize findings from different studies on the same topic. Given the need for accumulating scientific evidence, *research syntheses* have become an essential tool for researchers and practitioners interested on the most recent developments in their fields.

A research synthesis is carried out with the aim to clarify the state of the art in a given topic, by integrating information from multiple studies conducted to date (Marín-Martínez, 1996). The synthesist will have to face different challenges along the reviewing process. Although the studies analyzed the same research question, the methodological focus, measurement instruments, context, and sample characteristics will typically fluctuate from one to another. Also, contradictory results are likely to be found among studies, due to sampling error, study characteristics, or both (Hedges & Olkin, 1985). A research synthesis is, in sum, a complex process which requires systematization at each of its stages.

Research syntheses firstly appeared in Psychology and Education, but they have spread through many other disciplines, especially Medical Sciences and Social Policy Analysis. When a research synthesis is conducted, conclusions are addressed not only to scholars, but also to practitioners, policy makers, and the general public (Cooper & Hedges, 2009a). Meta-analysis is a methodology for research synthesis whose characteristics will be detailed along the next sections.

1.1.1 Meta-analysis and other forms of research

A way to delimit the aim and implications of meta-analysis is by comparing this methodology with some other forms of research. In this section, meta-analysis is firstly compared with primary and secondary researches, which also make use of quantitative methods. Later, differences between meta-analysis and other forms of research synthesis are detailed.

1.1.1.1 Primary research, secondary research, and meta-analysis

Data analysis can be conducted with different research goals (Glass, 1976). The most conventional one is accounted for by *primary analysis*, which refers to the original analysis of data previously collected for an individual study. Another possibility is to carry out a *secondary analysis*, that is, to re-analyze data to answer different research questions, or to address the original question employing different statistical techniques. Both primary and secondary analyses are considered as empirical studies (American Psychological Association, 2010).

The concept *meta-analysis*, or "analysis of analyses", was coined by Gene V. Glass (1976). Glass proposed this term to label "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976, p. 3). This methodology can be applied to many different disciplines in a very wide range of situations (e.g., Cooper et al., 2009). It is sensible to carry out a meta-analysis when enough studies are available on the same research question. In that case, a meta-analysis allows to integrate results from the individual studies, as well as to explain possible inconsistencies between findings from different studies (Borenstein, Hedges, Higgins, & Rothstein, 2009; Botella & Gambara, 2002; Cooper et al., 2009; Lipsey & Wilson, 2001).

Two main differences between the analyses conducted in individual studies and in meta-analysis can be outlined. Firstly, the analysis unit in an individual study is (usually) the subject, while in a meta-analysis the unit of analysis is (usually) the study. Secondly, while analyses in the individual studies are (usually) conducted by applying ordinary least squares (OLS) techniques, these procedures are inappropriate in a meta-analysis. The reason is that the variance of each unit of analysis in a meta-analysis (e.g., the study) is inversely proportional to the sample size and, since sample sizes widely vary along the set of studies of most meta-analyses, the assumption of homoscedasticity required for OLS analysis does not hold (Aloe, Becker, & Pigott, 2011; Raudenbush, 1994; Sánchez-Meca &

Marín-Martínez, 2008). Thus, instead of OLS procedures, weighted least squares (WLS) techniques are typically employed in meta-analysis.

The estimates obtained in a meta-analysis will have smaller standard errors and narrower confidence intervals than those obtained in the individual studies, leading to a gain of statistical power (Bonett, 2009; Cohn & Becker, 2003; Normand, 1999). Apart from this, an individual study using a very large sample of subjects will yield an accurate, powerful estimate of the effect in that study (Borenstein, Hedges, Higgins, & Rothstein, 2010), but this does not allow for generalizations to other scenarios different to that considered in the study (Raudenbush, 2009). Therefore, the main advantage of a meta-analysis is that conclusions using this methodology can be more broadly generalized than those achieved from primary or secondary analyses (Lau, Ioannidis, & Schmid, 1998; Matt & Cook, 2009).

1.1.1.2 Narrative reviews, systematic reviews, and meta-analysis

Although a few combinations of quantitative results can be found in the first decades of the twentieth century (e.g., Pearson, 1904) and even before (see Stigler, 1986), research syntheses until the 1970s were mostly qualitative and narrative (Glass, McGaw, & Smith, 1981; Sánchez-Meca & Ato-García, 1989). In a *narrative review*, an expert on the field reads and interprets the individual reports before elaborating conclusions that intend to summarize the state of the art. The main characteristic in such narrative research syntheses is the lack of a systematic schedule to make decisions, as well as the absence of any quantitative indicators (Marín-Martínez, 1996).

Due to this lack of systematization, several limitations of narrative reviews can be enumerated (Botella & Gambara, 2002; Marín-Martínez, 1996; Rosenthal & DiMatteo, 2001; Sánchez-Meca, 1986). The main problem in these reviews is that some crucial steps

such as inclusion and weighting of the studies can be affected by the expert's opinion and expectations, posing problems of subjectivity. Also, since most decisions made along the reviewing process are typically not specified in the report, another problem is the lack of replicability. Moreover, results from the individual studies are not quantified, so that it is not possible to assess their magnitude and variability. Furthermore, when narrative reviews quantify the results from a set of empirical studies, they usually count the number of statistically and nonstatistically significant results, a strategy that can lead to misleading results (e.g., Hedges & Olkin, 1985).

As an alternative, *systematic reviews* allow researchers to conduct research syntheses guided by objectivity, systematization, and replicability. The use of quantitative integration methods in a systematic review is known as meta-analysis, which was conceived to overcome the aforementioned limitations affecting narrative reviews (Glass, 1976; Rosenthal & DiMatteo, 2001; Sánchez-Meca, 1986). For this reason, *meta-analysis* can also be referred to as *quantitative review* (American Psychological Association, 2010). The methods employed for the integration of a set of studies have undergone enormous change, and quantitative and objective methods have become more and more implemented to the detriment of qualitative and subjective ones (Chalmers, Hedges, & Cooper, 2002; Shadish, Chacón-Moscoso, & Sánchez-Meca, 2005; Valentine, Cooper, Patall, Tyson, & Robinson, 2010).

In sum, meta-analysis allows researchers to quantitatively integrate the numeric results from a set of studies on the same topic, by applying the same rules and scientific rigor demanded for empirical studies (Botella & Gambara, 2002; Cooper, 1998; Hedges & Olkin, 1985; Hunter & Schmidt, 2004; Sánchez-Meca & Ato, 1989; Schulze, 2004). This scientific rigor leads to more valid conclusions than those achieved through narrative reviews.

As Rosenthal and DiMatteo (2001) stated, "meta-analysis allows researchers to arrive at conclusions that are more accurate and more credible than can be presented in

any one primary study or in a nonquantitative, narrative synthesis" (p. 61). Meta-analysis has, however, some limitations as well, as it will be detailed along this chapter.

1.1.2 Phases of a meta-analysis

A meta-analysis entails several phases (e.g., Botella & Gambara, 2006; Cooper et al., 2009; Lipsey & Wilson, 2001): (1) defining the research question, (2) literature search, (3) coding of studies, (4) statistical analyses and interpretation, and (5) publication. In this section, each of them will be briefly described.

1.1.2.1 Defining the research question

First of all, the constructs whose relationships are intended to be studied in the meta-analytic review must be specified. As pointed out by Cooper (2007), at this stage the meta-analyst must specify the research evidence relevant to those relationships. To reach this goal, all variables implied in the relationships of interest must be identified and described, including not only dependent and independent variables, but also some potential moderator variables. Before all of that can be stated, some previous planning about the synthesis process and further findings may be needed (Valentine, Pigott, & Rothstein, 2010).

Although this is a conceptual phase which does not entail many tasks, it will have a great influence on the remaining stages of the meta-analysis. A clear and precise definition of the research question is crucial before searching for the individual studies (Reed & Baxter, 2009), which constitutes the next phase. Also, the nature of the relationships of interest will affect the computation of effect sizes (Lipsey, 2009), which constitute the main outcome variable in a meta-analysis.

1.1.2.2 Literature search

Once the research question has been established, the next goal consists of locating and retrieving the individual studies that analyzed that question. A set of inclusion and exclusion criteria for the studies must be defined. Typically, the meta-analyst is interested in primary studies, that is, studies that recruited a sample of subjects, employed measurement instruments, and reported quantitative results. Also, the search must be restricted to a range of years. A common practice is to delimit the search from the year when the research question was firstly proposed in the literature to the present. Another issue is the specification of the languages that the research team can read. Moreover, the studies can be required to fulfill some other criteria in order to restrict the search process to studies with a specific design type, a minimum sample size and/or a minimum methodological quality. Depending on the question addressed in the meta-analysis, some selection criteria of the studies will be referred to the population to which the participants in the samples pertain, the kind of experimental manipulations (e.g., types of treatment, interventions, or programs), and the type of outcomes measured in the participants.

Regarding search procedures, a combination of several strategies should be the best option (Reed & Baxter, 2009). Nowadays, electronic sources constitute an indispensable tool (White, 2009), including general databases such as the Web of Knowledge or ProQuest, specific ones like PsycINFO and MEDLINE, or search engines like Scholar Google. Choosing the right terms, or key-words, is a crucial issue if the researcher aims to find the relevant pieces of empirical evidence on the topic (Cooper, 2007). Since the goal of the literature search is completeness, the search terms should include all relevant words to the topic of interest, including synonyms and related terms (Reed & Baxter, 2009).

Other strategies for the retrieval of the individual studies of interest are backward and forward searches. Backward search refers to the identification of publications by checking the citations included in the already retrieved documents. On the other hand, a

forward search involves identifying all items that cited a retrieved publication (Normand, 1999). Informal sources (e.g., conference contributions, master and doctoral theses) and experts' consultation constitute very valuable complements in the search process, especially for the retrieval of fugitive or grey literature (Sutton, 2009), that is, unpublished documents or manuscripts published in journals or books that cannot be found through formal sources (Sánchez-Meca & Marín-Martínez, 2010).

The main threat at this point is *publication bias*, which occurs when, in a given research field, studies providing statistically significant results are more likely to reach publication than the ones with nonsignificant results (Hedges, 1992; Sutton, 2009). Efforts to locate and to retrieve unpublished studies constitute a very important issue in a meta-analysis, as well as to check whether publication bias can produce a bias in the meta-analysis results (Begg & Mazumdar, 1994; Hedges & Vevea, 1996; Rothstein, Sutton, & Borenstein, 2005; Sánchez-Meca & Marín-Martínez, 2010).

1.1.2.3 Coding of studies

At this step, information from the variables considered as potential moderators must be gathered. Moderators refer to those variables that might affect the magnitude of the relationship under study. Although that list of potential moderators will vary from one meta-analytic review to another, three broad categories of moderator variables can be distinguished: methodological, substantive, and extrinsic variables.

Substantive variables are those specific to the phenomenon under study. They are strongly dependent on the research topic and constitute the group of variables that, on a theoretical basis, are expected to be related to the study outcomes. In psychological research, this category includes characteristics of the sample subjects (e.g., age, gender,

ethnicity, severity of the disorder), of the treatment (e.g., duration, theoretical approach, therapists' experience), and of the context (e.g., geographical and cultural environment).

Methodological variables refer to characteristics of the designs and methods of the studies, whose influence should be discarded before interpreting any substantive relationship (Lipsey, 2009). Within the psychological field, they can include aspects such as the design type (e.g., experimental vs. quasi-experimental), the type of control group (e.g., active vs. inactive, psychological vs. pharmacological placebo), attrition, use or not of blinded assessors, or use of intention-to-treat vs. completers analyses. Some of them are of interest irrespective of the field where the meta-analysis is carried out. Studies methodologically flawed can offer biased estimates of the effects. In order to assess the potential risk of bias in the effect estimates from the studies, the meta-analyst must include some quality checklist or scale proposed in the literature.

Lastly, 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 (Lipsey, 2009; Sánchez-Meca & Marín-Martínez, 2010). As methodological moderators, extrinsic variables may appear as confounding variables in a meta-analysis, and ignoring them might lead to a wrong interpretation of the results. This category includes features like the publication year, publication source (published vs. unpublished), the main author's affiliation and sex or the existence of a potential conflict of interests with regards to the funding source of the study.

In practice, the coding of studies is a complex process that usually entails many decisions to be made by the meta-analytic team, because the information about the variables of interest is not always clearly reported in the individual studies. A coding protocol with a list of items must be developed, in order to guarantee the transparency and replicability of the coding process (Wilson, 2009). In such protocols, the important information from each unit of analysis (e.g., study) is gathered and, in some cases, readily computerized for further analytic purposes. Another useful tool for the meta-analytic

team is a coding manual where all decisions and inferences to be made from the incompletely reported information in the individual studies are specified.

The main threat at this stage is the lack of reliability in the coding process (Orwin & Vevea, 2009). Both intrarater and interrater reliabilities should be evaluated (Wilson, 2009). Intrarater reliability refers to the consistency of a single coder when applying the coding protocol to the same studies in different occasions. Although some discrepancies might arise between the coding decisions of a single person, interrater reliability is usually the main concern for the meta-analyst at this stage. Interrater reliability is the degree of agreement between coders. For its evaluation, at least a random sample of the meta-analytic units (e.g., studies) should be independently coded by two or more members of the research team. Interrater reliability can be assessed using indices such as intraclass correlation and Cohen's kappa for continuous and categorical moderators, respectively. When discrepancies are found between different coders, the final decision should be made based on coder consensus (Orwin & Vevea, 2009).

1.1.2.4 Statistical analyses and interpretation

At this stage, the meta-analyst must select and apply the most suitable procedures to combine the results across the individual studies and to analyze possible sources of the variability among the study results (Cooper, 2007). This phase entails several computational stages, from the calculation of an effect size from each unit (e.g., study) to the statistical integration of results.

The first goal for the meta-analyst at this stage consists of choosing a numeric index to summarize the results from each unit of analysis. For the remainder of this dissertation, studies will be assumed to be the unit of analysis, although some other scenarios are feasible in meta-analytic applications (e.g., one study can provide multiple

outcomes). This goal is achieved by computing an effect size index for each study. Effect sizes constitute the main outcome variable in a meta-analysis (e.g., Hedges, 1992, 2007; Hedges & Olkin, 1985; Sánchez-Meca & Marín-Martínez, 2010). Different indices can be computed depending on the purpose of the meta-analysis and the metric of the variable/s implied in the relationship of interest. This issue will be addressed in Chapter 2.

Once the information from the studies has been summarized, statistical analyses can be conducted. The first step will be a descriptive analysis of the variables coded from each study, which will take part in the inferential analyses to be conducted afterwards. Descriptive analyses of all data collected so far can provide a picture of the "typical" study (Lipsey, 2009). That information is achieved by computing indexes such as the mean or median, standard deviation or range, percentage, and some asymmetry index. Charts such as stem-and-leaf displays and box-and-whisker plots (Tukey, 1977) are also recommended to illustrate the data distribution.

After descriptive analyses, the first inferential goal for the meta-analyst is the calculation of an overall effect size estimate. When computing this average, effect sizes are usually weighted by some function of their respective sample sizes, with greater weights for the most accurate estimates, that is, for the estimates computed from the largest samples (e.g., Shadish & Haddock, 2009). Choice of weights will require the assumption of an underlying statistical model, typically a fixed-effect or a random-effects model (this matter will be considered further in this chapter). The overall effect estimate is usually complemented with a confidence interval, which allows the synthesist to test the hypothesis that the overall effect is null.

The estimation of the mean effect allows the meta-analyst to answer questions such as: does the intervention program work on average? What is the mean precision of the instrument? However, the overall effect size in a meta-analysis may not be very informative in situations where several studies have conflicting results. In this case, a

somewhat better alternative is to average a subset of studies providing similar statistical conclusions (Normand, 1999).

In addition to the overall effect estimate, it is also interesting to evaluate the heterogeneity across the effect sizes of the meta-analytic data set. For this purpose, the Q statistic (Cochran, 1954; Hedges & Olkin, 1985) is often employed to test the null hypothesis that variability among the effect sizes is only due to random sampling error. Nonetheless, the power of the Q statistic is strongly dependent on the number of studies (Aguinis, Gottfredson, & Wright, 2011; Baker, White, Cappelleri, Kluger, & Coleman, 2009; Hardy & Thompson, 1998; Sagie & Koslowsky, 1993; Sánchez-Meca & Marín-Martínez, 1997; Schmidt, Oh, & Hayes, 2009; Viechtbauer, 2007c). For that reason, the I^2 index (Higgins & Thompson, 2002), which quantifies the percentage of heterogeneity among effect sizes different to sampling error, has been recommended as a complement to the statistical conclusion of the Q statistic (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006; Shadish & Haddock, 2009).

Finally, the influence of moderators on the variability of the effect sizes is analyzed. A moderator variable is a categorical or continuous variable that exerts an influence on the direction and/or strength of the relationship of interest (Baron & Kenny, 1986). Since moderator analyses constitute the main focus for all of the empirical studies developed along this dissertation, such analyses will be described with more detail later in this dissertation.

Several computerized alternatives are available to the researcher when conducting a meta-analysis, including specific software (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2005; Review Manager, 2011; Rosenberg, Adams, & Gurevitch, 1999) or macros developed for their implementation in generic statistical packages (e.g., Harbord & Higgins, 2008; Viechtbauer, 2010). In addition to the statistical computation, these tools also allow the researcher to elaborate some graphical displays specifically designed for

meta-analysis, such as the funnel plot and the forest plot (Anzures-Cabrera & Higgins, 2010; Borman & Grigg, 2009).

1.1.2.5 Publication

Once the data analysis and result interpretation is finished, it is time to write the meta-analytic report (Sánchez-Meca & Botella, 2010). Some guidelines have been published to help meta-analysts to correctly report and write a meta-analysis, such as the PRISMA statement (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*; Moher, Liberatti, Tetzlaff, Altman, and The PRISMA Group, 2009) and the AMSTAR statement (*Assessment of Multiple SysTemAtic Reviews*; Shea et al., 2007).

The schedule in a meta-analytic report is similar to that of primary research (American Psychological Association, 2010): introduction, method, results, and discussion. The introduction must justify the need for carrying out the meta-analysis (Sánchez-Meca & Marín-Martínez, 2010), providing some theoretical background that will allow readers to understand the relevance of the findings reported in the meta-analysis. In the introduction, the objectives of the meta-analysis must be made explicit. In the method section, several steps already detailed along Section 1.1.2 of this dissertation must be described: search for the studies, coding of the moderator variables, selection of the effect size index, and statistical analyses. It is important to specify all statistical techniques employed and decisions made along the meta-analytic review, in order to guarantee its replicability.

The results section must include, as detailed before, both descriptive and inferential analyses, providing the reader with an overview of the set of studies integrated, as well as an average of the effect size estimates and further analyses to explain at least part of the variability among them. Including tables and charts in this

section will be helpful to the reader (Botella & Gambara, 2006). In the discussion, the implications of the results obtained must be presented, together with the limitations of the meta-analysis. In a meta-analytic study, these limitations will often include aspects such as language restrictions, failure to locate unpublished studies, or failure to retrieve located studies (Clarke, 2009). The meta-analytic report should finish offering some recommendations for future research in the field.

At the end of the report, when listing the references, the American Psychological Association (2010) recommends to include and remark those from the individual studies integrated in the meta-analysis, or to provide these references in an appendix if the number of studies exceeds 50. Lastly, if space restrictions of the journal allow for that, it is recommendable to include an appendix with the whole database, where the main variables employed in the statistical analyses are gathered. This will allow any interested researcher to try to replicate the results or even to conduct complementary analyses using different statistical techniques.

1.1.3 Limitations of meta-analysis

Like in any primary research, multiple threats to the validity of a meta-analysis can limit the scope and the generalizability of the results (Hedges, 1992; Lau et al., 1998; Marín-Martínez, 1996; Matt & Cook, 2009; Rosenthal & DiMatteo, 2001). Some of these limitations can also affect any other form of research, while some others are specifically related to meta-analysis.

One limitation is related to *publication bias*, which was previously defined. It constitutes a threat not only to the meta-analytic conclusions, but also to the statistical techniques employed in a meta-analysis. Publication bias will tend to induce a negative correlation between effect size and sample size in a set of published studies. This is

because studies with small sample sizes have to estimate a large effect size to reach a statistically significant result and be published. Since the weighting factor in meta-analysis is usually a function of sample size, this trend might produce biased results (Begg & Mazumdar, 1994; Henmi & Copas, 2010; Levine, Asada, & Carpenter, 2009; Slavin & Smith, 2009).

Because studies with small sample sizes have a low statistical power (Matt & Cook, 2009) and therefore a low probability to find statistically significant results, they can be specially affected by publication bias. For that reason, some authors have proposed excluding underpowered studies in a meta-analysis as a way to solve the problem of publication bias (Hedges & Pigott, 2001; Kraemer, Gardner, Brooks, & Yesavage, 1998; Muncer, Craigie, & Holmes, 2003). A different approach, which has been recently proposed to deal with this problem (Moreno et al., 2012), is a modification of the weights such that it minimizes the impact of small studies on the pooled results.

Several authors have made great efforts to help researchers to determine the extent to which publication bias might affect the validity of their results and conclusions, by means of different statistical methods and graphical displays (Duval & Tweedie, 2000a, 2000b; Howell & Shields, 2008; Light & Pillemer, 1984; Rothstein et al., 2005; Sutton, Duval, Tweedie, Abrams, & Jones, 2000). Ferguson and Brannick (2012) examined a sample of 91 recent meta-analyses from the psychological field, and found that 70% of them made some effort to analyze publication bias. Finally, it should be pointed out that, despite it constitutes one of the main concerns for a meta-analyst, publication bias is even more problematic in non-quantitative syntheses, because methods for dealing with it are very limited (Sutton, 2009).

Another issue is the influence of *reporting and methodological quality of the individual studies* on the meta-analysis results. Meta-analysis was conceived as a methodology for the integration of information from several individual studies, considering that results from each individual study will provide important data that should

not be discarded (Glass, 1976). However, fluctuations in the rigor with which the individual studies are conducted and their results are reported may affect the results and conclusions of the research synthesis (Valentine, 2009). Consequently, meta-analysis has been criticized due to the inclusion of studies irrespective of their quality, and this criticism has been labeled as "garbage in and garbage out" (Hunt, 1997). Two main approaches have been implemented regarding quality (Lipsey & Wilson, 2001). One of these approaches consists of including only studies that fulfill several quality criteria, while the other implies incorporating quality to the analyses, either as a weighting factor (e.g., Rosenthal, 1995) or as part of moderator analyses, that is, treating quality as an empirical issue (Valentine, 2009).

However, analyzing the quality of the individual studies and its influence on the meta-analysis results is not a trivial issue. Rules for assessing quality and determining its relevance to the relationship of interest remain unclear at present (López-Pina, Sánchez-Meca, & Núñez-Núñez, 2011, July; Normand, 1999). As an attempt to circumvent this problem, dozens of quality scales have been developed, with their items reflecting quality indicators that might have an influence on the results from each individual study. The aim of such scales was to obtain a pooled value for the quality of an individual study. However, the sum of items addressing different aspects related to the methodological quality did not prove to be useful in meta-analysis up to date (Valentine, 2009). Currently, it is more accepted to assess the methodological quality of the studies by applying a list of individual items, but without reporting a total score (Herbison, Hay-Smith, & Gillespie, 2006; Higgins & Green, 2008; Jüni, Altman, & Egger, 2001; Littell, Corcoran, & Pillai, 2008). In sum, there is still much work to be done before reaching consensus about how to assess the quality of the meta-analytic units and which procedure works best to minimize the effect of this threat to the meta-analytic conclusions.

Another problem which is receiving more and more attention in meta-analysis is the *dependency among effect size estimates* (Ahn, Myers, & Jin, 2012; Glesser & Olkin,

2009; Lipsey, 2009). Perhaps the most common type of dependency in meta-analysis arises when multiple effect sizes are extracted from the same participant sample on similar outcome constructs (e.g., effect sizes are clustered within studies). Another situation where dependency can be found occurs when some research teams are responsible for multiple studies included in the meta-analysis (e.g., studies are clustered within research teams). The most commonly used meta-analytic methods do not account for such dependency structures when the meta-analytic data are clustered.

Although multivariate techniques can be employed for handling dependent effect size estimates, such techniques require information about the covariance structure that is rarely available or reported in the individual studies (Gleser & Olkin, 2009; Jackson, Riley, & White, 2011). In order to satisfy the assumption of independent effect sizes, most meta-analysts traditionally either selected one effect size per cluster for the analyses, or created one effect size per cluster by averaging all effect sizes within that cluster (Hedges & Olkin, 1985; Marín-Martínez & Sánchez-Meca, 1999; Rosenthal & Rubin, 1986) or simply choosing one of them based on substantive reasons. Such strategies lead, in all cases, to a loss of information (Becker, 2000), and averaging several effect sizes within a study can even provide misleading results if those effect sizes are negatively correlated, which can only be checked in the unlikely event that the primary report includes all participants' data. Moreover, some other meta-analytic studies analyzed the whole set of effect sizes ignoring dependencies. Nevertheless, failure to recognize dependency and to use appropriate analytic techniques to cope with it can lead to inaccurate estimates of effects and their standard errors, the latter usually being too small (Hedges, 2009; Van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, in press; Van Houwelingen, Arends, & Stijnen, 2002).

Since none of the aforementioned strategies sounds completely satisfactory, it is little surprise that methodologists have begun to propose new methods to deal with dependent effect sizes that are feasible for most meta-analysts to use. One of these

methods is based on robust variance estimation (Hedges, Tipton, & Johnson, 2010). Also, multilevel models, which account for variation at different levels (Gelman & Hill, 2007; Goldstein, Browne, & Rasbash, 2002), have been proposed as an alternative to analyze meta-analytic databases containing dependency structures (Beretvas & Pastor, 2003; Hox & de Leeuw, 2003; Konstantopoulos, 2011; Stevens & Taylor, 2009; Van den Noortgate et al., in press).

Critics of meta-analysis have raised some other limitations in the literature (e.g., Hunt, 1997). One of those is the *overemphasis on the main effects* of each individual study, to the detriment of other interesting findings (e.g., interactive within-study effects). Another famous criticism to meta-analysis is the *apples and oranges* argument, which remarks on the fact that meta-analysis involves summarizing results from studies that might widely vary in several aspects, such as the operationalization and measurement of the variables of interest or their methodological framework. Besides the criticisms, meta-analysis has consolidated as an indispensable methodology accepted by the scientific community in all empirical sciences.

1.1.4 Meta-analysis and Evidence-Based Practice

Decisions affecting professional practice should be endorsed by the best scientific evidences available (e.g., Sánchez-Meca, Marín-Martínez, & López-López, 2011). This is the main goal of the so-called *Evidence-Based Practice* approach, which recently emerged with the aim to support practitioners and policy makers by providing them with the best empirical findings in their fields. When multiple studies are available on the same topic, the best evidence can be produced by a meta-analysis integrating their results.

Probably, the first key date concerning the *Evidence-Based Practice* approach is the establishment of the U.K. Cochrane Center in 1992, with the aim to "facilitate the creation

of an international network to prepare and maintain systematic reviews of the effects of interventions across the spectrum of health care practices" (Cooper & Hedges, 2009a, p. 10). One year later, the Cochrane Collaboration was founded[1], reaching in 2006 the amount of 11,000 contributors, and being considered for many people as the gold standard for assessing treatment effectiveness in Medical Sciences at present.

The Cochrane Collaboration promotes high-quality systematic reviews by helping meta-analysts in different ways (White, 2009). One of these services is the specification of criteria for including studies in a meta-analysis, as well as for reporting results. Also, the Cochrane Library gathers several databases containing useful documents such as systematic reviews already done or methodological improvements involving research syntheses.

The Cochrane Collaboration is considered nowadays as a great support for practitioners and policy makers in Medicine (Baker et al., 2009). This institution has established the following ranking for scientific evidences:

I. Evidence obtained from a meta-analysis of randomized controlled experiments.

II. Evidence obtained from (at least) a randomized controlled experiment.

III. Evidence obtained from (at least) a controlled study (not randomized).

IV. Evidence obtained from (at least) a quasi-experimental study.

V. Evidence obtained from descriptive studies.

VI. Evidence obtained from an experts committee.

[1] www.cochrane.org

In a similar vein, the Campbell Collaboration was established in 2000 with a multidisciplinary focus. The goals of its founders are summarized in this definition[2]: "The Campbell Collaboration (C2) helps people make well-informed decisions by preparing, maintaining and disseminating systematic reviews in education, crime and justice, and social welfare."

Similar to the Cochrane Collaboration, services from the Campbell Collaboration to reviewers include specialized databases containing useful information and the clear establishment of the criteria for inclusion of studies in a meta-analysis, choice of the search strategies, and so on. Moreover, the so-called coordinating groups supervise the preparation of reviews in different fields (White, 2009).

The goals of the Campbell Collaboration include the avoidance of duplicities in systematic reviews, the minimization of biases in the results, and the constant updating process by incorporating new scientific findings. The aim of this institution is to provide support to professionals from Social, Educational, Criminological, and Behavioral Sciences (e.g., Sánchez-Meca, Boruch, Petrosino, & Rosa-Alcázar, 2002; Sánchez-Meca & Botella, 2010).

One more institution that can be regarded as a product of the *Evidence-Based Practice* approach is the Joanna Briggs Institute, which was established in the Nursing field in 1996. Visitors of the website of the Joanna Briggs Institute[3] can find the following definition: "the Institute is known for providing reliable evidence which health professionals can use to inform their clinical decision making. The Institute develops evidence in various formats for nursing, allied health and medical professionals as well as support information for consumers".

[2] www.campbellcollaboration.org

[3] www.joannabriggs.edu.au

In short, several scientific networks have been established in the last years to promote high-quality research syntheses in different fields. Their existence is very helpful not only for practitioners and policy makers, but also for meta-analysts, especially at stages such as the literature retrieval or the choice of methods for integrating results (White, 2009).

1.2 Statistical models in meta-analysis

Nowadays, different statistical models are available when carrying out a meta-analysis, and the model choice will have an influence not only on the statistical procedures for integrating the information, but also on the generalizability of the results (Hedges & Vevea, 1998). Moreover, depending on some characteristics of the meta-analytic database, some models might not be appropriate. In this section, the statistical models that can be assumed when conducting a meta-analysis will be presented, together with a summary of the main factors that should be considered for the model choice.

1.2.1 The fixed-effect model

Some meta-analytic models can be classified as *fixed-effect models*. These models assume that the parametric effect sizes are fixed but unknown constants to be estimated, and they usually assume as well that parameters are homogeneous from one study to another (Hedges & Vevea, 1998). In a fixed-effect model, also named common-effect model (cf. Borenstein et al., 2010), the variability between estimates is assumed to be wholly due to random sampling of participants for the individual studies (e.g., Schulze, 2004). Since the effect parameters from the studies included in the meta-analysis are the

only ones intended to estimate, results assuming a fixed-effect model can only be extended to studies identical to those included in the meta-analysis (Normand, 1999).

Let k denote the number of studies included in a meta-analysis and $\hat{\theta}_i$ the effect size estimate for the *i*th study. In a fixed-effect model, $\hat{\theta}_i$ can be defined as

$$\hat{\theta}_i = \theta + e_i, \tag{1.1}$$

where θ is the (common) parametric effect size, and e_i is the sampling error of $\hat{\theta}_i$, with distribution $N(0, \sigma^2_{y_i})$, with $\sigma^2_{y_i}$ being the within-study variance for the *i*th study. Although this variance needs to be estimated, nearly unbiased estimators are available for the most common outcome variables in meta-analysis, so that parametric within-study variances are typically considered as known in practice. Assuming a fixed-effect model implies using WLS techniques and, since greater weights are given to the most accurate estimates of the (common) parametric effect size, the inverse within-study variances can be employed as the weighting factor.

The fixed-effect model has been the most frequently assumed statistical model for meta-analyses published up to date in Psychology (cf. Schmidt et al., 2009). It makes sense to assume a fixed-effect model when the goal is to generalize results only to the set of studies included in the meta-analysis and we can assume that the studies are estimating a common effect in the population; in other words, a fixed-effect model is in order when we can reasonably assume that the variability exhibited by the effect estimates in the meta-analysis is due to within-study sampling error alone, not to true heterogeneity (Borenstein et al., 2010; Erez, Bloom, & Wells, 1996; Field, 2005; Hedges & Vevea, 1998). However, since it is usually unrealistic to assume that the effect estimates have a common population effect, some other alternatives are becoming more widely employed in meta-analysis to the detriment of the fixed-effect model.

1.2.2 The varying coefficient model

Laird and Mosteller (1990) proposed an alternate model that has been recently advocated by Bonett (2008, 2009, 2010), who proposed to label it as *varying coefficient model*. The main difference with the fixed-effect model is the assumption of heterogeneity between the effect parameters, that is, it is assumed that each study estimates a different parametric effect. In this model, studies are not assumed to be randomly sampled from a larger population of studies. Thus, as in the fixed-effect model, conclusions from a meta-analysis carried out with the varying coefficient model can only be extended to studies identical to those incorporated to the meta-analysis.

The varying coefficient model can be expressed with the formula

$$\hat{\theta}_i = \theta_i + e_i, \tag{1.2}$$

where θ_i is the parametric effect for the *i*th study. Heterogeneity assumption seems a more realistic option for most situations in Social and Behavioral Sciences (Aguinis et al., 2011; Schmidt, 2010). Therefore, the varying coefficient model will generally be preferred to the fixed-effect model.

1.2.3 The random-effects model

Apart from the fixed-effect and the varying coefficient models, the other leading possibility in meta-analysis is to assume a *random-effects model*. The random-effects model assumes that each study estimates a different parametric effect and, in contrast to the alternatives presented above, that the studies are randomly sampled from a broader population of studies (Borenstein et al., 2010; Overton, 1998; Sánchez-Meca & Marín-Martínez, 2010; Schmidt et al., 2009). Once the set of studies under investigation is

assumed to be a random sample, then the meta-analysis can be conceived as a double sampling process (Raudenbush, 2009; Viechtbauer, 2007a): firstly, subjects are randomly sampled for each study and, secondly, studies are randomly sampled for the meta-analysis by extracting them from a larger population of *potential studies* (Raudenbush, 2009, p. 297). Conclusions arising from a meta-analysis where a random-effects model is assumed are applicable not only to identical studies to those included in the meta-analysis, but also to other studies with similar, but not identical, characteristics that have been carried out or that can be conducted in the future.

The random-effects model can be expressed with the equation

$$\hat{\theta}_i = \mu + e_i + \varepsilon_i, \tag{1.3}$$

where μ represents the hypermean, that is, the mean from the population of parametric effects, and ε_i denotes the difference between the parameter from the *i*th study and the hypermean. It is assumed that $\varepsilon_i \approx N(0, \tau^2)$, with τ^2 being the (total) heterogeneity variance, which can be defined as the excess variation among the effect sizes over than expected from the imprecision of results within each study (Thompson & Sharp, 1999). As a result, the effect size estimates $\hat{\theta}_i$ are assumed to be normally distributed with mean μ and variance $\sigma^2_{y_i} + \tau^2$, that is, $\hat{\theta}_i \approx N(\mu, \sigma^2_{y_i} + \tau^2)$. As in the fixed-effect model, statistical techniques applied in a random-effects model will routinely include weights. If the inverse variance is the weighting scheme applied, this will now imply the addition of a second variance term, τ^2 (e.g., Viechtbauer, 2007b).

Since it incorporates two variance components, results assuming a random-effects model are usually more conservative than those obtained when assuming the remaining statistical models (Beretvas & Pastor, 2003; Brockwell & Gordon, 2001; Hedges & Vevea, 1998; Raudenbush, 1994). Also, the study weights will be more similar under a random-

effects model – large studies lose influence while small studies gain influence – than under a fixed-effect model (Borenstein et al., 2010; Schulze, 2004).

The random-effects model is more consistent than the two aforementioned alternatives with standard scientific aims of generalization, and allows for summarizing results in a more efficient way as the number of studies increases (Borenstein et al., 2010; Hunter & Schmidt, 2000; Marín-Martínez & Sánchez-Meca, 2010; Raudenbush, 2009; Sutton & Higgins, 2008). For those reasons, it has become widely applied for meta-analytic studies in Psychology and many other disciplines such as Medicine and Education.

1.2.4 Model choice

There is a general consensus to consider that the main criterion for choosing the statistical model in a meta-analysis should be the extent to which the meta-analyst aims to generalize his/her results (Borenstein et al., 2010; Hedges & Vevea, 1998; Overton, 1998; Sánchez-Meca, López-López, & López-Pina, in press; Schmidt et al., 2009). If the meta-analyst intends to generalize results to a population of studies identical to those included in the meta-analysis, then fixed-effect and varying coefficient are appropriate models. The latter seem more realistic because, in contrast to the fixed-effect model, it assumes that each study estimates a different effect parameter. In the unlikely event that all studies estimate a common population effect and generalization is only intended to the specific set of studies included in the meta-analysis, fixed-effect models constitute an optimal choice.

More often, however, generalization is intended to a larger population of studies than those included in the meta-analysis. The aim in a meta-analysis, as in any research project, is usually to generalize the results beyond the integrated units. As Schmidt and colleagues (2009) stated, "the usual goal of research (...) is generalizable knowledge (...),

which requires generalization beyond the current set of studies to other similar studies that have been or might be conducted" (p. 101). Consequently, random-effects models are conceptually more appropriate for the majority of situations when conducting a meta-analysis (Field, 2003, 2005; National Research Council, 1992).

However, applying random-effects models entails two main problems. Firstly, the studies in a meta-analysis are not randomly selected from a larger population of studies in practice and therefore, in the strictest sense, it is not appropriate to make inferences about that superpopulation. This is a criticism raised by Bonett (2008, 2009, 2010) against the use of random-effects models. Secondly, with a small number of studies, estimates of the heterogeneity variances are very inaccurate, and this might affect the statistical analyses conducted with random-effects models (e.g., Brockwell & Gordon, 2001, 2007; Hardy & Thompson, 1996).

With regards to the first problem, as stated by Laird and Mosteller (1990), "making inferences as if dealing with random samples contrary to fact is not a special issue for meta-analysis, but for all of science and technology" (p. 14). Therefore, if this criticism was extended to primary research, then no meta-analytic model would be appropriate, since the vast majority of individual studies strictly violate the random sampling assumption (cf. Edgington, 1966; Frick, 1998; Overton, 1998). However, statistical inference techniques are routinely applied in primary research, and primary researchers routinely generalize their results to a population of units. Likewise, the meta-analyst will apply random-effects models when he/she can assume, on a reasonable basis, the set of studies included in the meta-analysis to be a representative sample of a potential population of past and/or future studies. To apply random-effects models, the meta-analyst must define, also on a reasonable basis, the characteristics of the potential population of studies to which he/she aims to generalize the results.

The other problem that the meta-analyst will have to face when applying random-effects models refers to the difficulties in accurately estimating the heterogeneity variance

when the number of studies is small. Borenstein et al. (2010) proposed several solutions to this problem. One solution is to apply fixed-effect or varying coefficient models instead of random-effects models. But if the meta-analyst aims to generalize his/her results to a larger population of studies, this solution will not be satisfactory. Another solution is not to do the meta-analytic integration if the number of studies is not high enough, which means leaving unfinished an investigation which undoubtedly will have involved a great effort.

Several authors provided meta-analysts with some guide to determine which number of studies should be large enough to assume a random-effects model, and some approximate values that have been proposed in the literature are 20 (Aguinis et al., 2011; Biggerstaff & Tweedie, 1997; Brockwell & Gordon, 2001; Field, 2005) and 32 studies (Schulze, 2004). With a smaller number of studies, a reasonable goal is to generalize results only to a population of studies identical in composition and variability to those included in the meta-analysis (Raudenbush, 1994; Sánchez-Meca, López-López, & López-Pina, in press) and, therefore, assuming a varying coefficient model will be a suitable option for most situations.

An additional problem affecting random-effects models, and any other model using WLS methods, is that weights can lead to biased estimates if effect sizes and sample sizes are correlated. This criticism comes from the finding in some meta-analyses of a negative correlation between sample size and effect size and, as a consequence, some authors have proposed using OLS techniques (e.g., Shuster, 2010). The most frequent reason for a negative correlation between effect size and sample size is the existence of publication bias, in the sense that studies with small sample sizes need to estimate large effect sizes to be published. However, as Thompson and Higgins (2010) argued, this is an empirical issue, so that all meta-analyses should examine the correlation between sample size and effect size and, if a high correlation is found, then the meta-analyst should

investigate the reasons for this fact and decide whether or not to apply weighting schemes.

In summary, since the goal in meta-analysis is to generalize knowledge, random-effects models constitute an optimal alternative as long as some criteria are met (e.g., Biggerstaff & Tweedie, 1997; Schmidt, 2010). For that reason, the present dissertation is focused on random-effects models, and most of the methods described throughout this manuscript and compared by means of the simulation studies here presented are specific to these statistical models.

1.3 Moderator analyses

When the different phases in a meta-analysis were described in Section 1.1.2.4, it was mentioned that the first inferential goal in the statistical analyses is to obtain an overall effect size estimate, together with its confidence interval. The overall effect size is a very informative index in situations where studies integrated are similar enough to discard any moderating effect. However, this is rarely the case in practice, and usually the studies will differ to some extent in one or more characteristics, leading to discrepant results (Makambi, 2004; Sidik & Jonkman, 2005b). Under these conditions, the usefulness of the mean effect size becomes very limited, and its interpretation can even be misleading if one or more variables are affecting the effect size estimates (Hartung, Knapp, & Sinha, 2008; Viechtbauer, 2008). Therefore, moderator analyses are justified in the vast majority of scenarios and constitute a crucial issue in meta-analysis (Lipsey, 2009).

In addition to the overall effect size estimate, another inferential task that the meta-analyst will typically attend before conducting moderator analyses is the assessment of the amount of heterogeneity among the effect size estimates. From the last paragraph,

it might be concluded that the presence of moderators will necessarily lead to a significant result for the Q statistic. Nevertheless, that is not always the case because, as it was previously noted, the statistical power of the Q statistic is strongly dependent on the number of studies (e.g., Aguinis et al., 2011; Biggerstaff & Tweedie, 1997; Hardy & Thompson, 1998; Pereira, Patsopoulos, Salanti, & Ioannidis, 2010; Sagie & Koslowsky, 1993; Sánchez-Meca & Marín-Martínez, 1997; Thompson, 1994). Thus, it is advisable to carry out moderator analyses regardless of the statistical conclusion of the Q statistic (Baker et al., 2009; Sánchez-Meca & Marín-Martínez, 1998a).

The statistical analyses examining the influence of the study characteristics on the effect sizes are known as *moderator analyses*. While simple subgrouping of the studies can be used for that purpose (Borenstein et al., 2009), meta-analysts are increasingly employing so-called *meta-regression* models to study the influence of one or multiple moderator variables on the effect sizes (Thompson & Higgins, 2002). In a meta-regression model, the effect size estimates are used as the dependent variable, and moderators are incorporated to the model as independent variables. Not only continuous, but also categorical moderators can be included in the model, using appropriate dummy coding (Viechtbauer, 2007a).

All three statistical models presented before can be applied to examine the influence of moderator variables on the effect sizes (e.g., Bonett, 2009; Cooper et al., 2009; Hedges & Olkin, 1985), and the model choice can affect the statistical conclusions and will determine their generalizability (Aguinis et al., 2011). When a random-effects model is assumed, the set of effect sizes is treated as a random variable. Since the predictors included in the model are usually added as fixed effects, this approach then leads to a *mixed-effects meta-regression model*. Such model will be described with more detail in Chapter 3.

When results from the moderator analyses are interpreted, it is important to take into account that these analyses cannot provide causal evidence, because the meta-

analyst only observes retrospectively the characteristics of the studies, instead of manipulating levels for each of those independent variables (Baker et al., 2009; Hardy & Thompson, 1998; Thompson & Higgins, 2002; Viechtbauer, 2007a). Nonetheless, moderator analyses allow the synthesist to examine relationships that have never been explored before in the individual studies (Botella & Gambara, 2002), which may lead to interesting hypotheses to be tested in future primary research. To this respect, the researcher must be aware that a relationship found at the aggregate level (e.g., studies) might not be present at the individual level, due to the so-called ecological fallacy (Robinson, 1950).

The selection of moderators should be guided by an expert on the field under study (Baker et al., 2009; Raudenbush, 1994), and the number of moderators to be tested should be limited, in order to avoid false positive findings (e.g., Cohen, 1990; Hunter & Schmidt, 2004; Thompson & Higgins, 2002), especially when the number of studies integrated is small. In such cases, results should be interpreted very cautiously (Thompson, 1994). Another issue that requires attention at this stage is the correlation between moderators, which might also lead to an overestimation of the moderator effects (Konstantopoulos & Hedges, 2009; Viechtbauer, 2008). Regarding types of moderator variables, the potential influence of extrinsic and (especially) methodological moderators should be discarded before analyzing substantive moderator variables (Lipsey, 2009). The main purpose of this step in the analysis is to be able to obtain an explanatory, or predictive, model that contains the subset of moderator variables more statistically related to the effect sizes.

In summary, the sources of variability between the outcomes of different studies should be routinely investigated in meta-analysis, in order to increase the practical relevance of the conclusions extracted from the synthesis and the scientific understanding of the set of studies integrated (Thompson, 1994). If the aim of the meta-analyst is to generalize results beyond the sample of studies, then a random-effects model must be

assumed, leading to mixed-effects meta-regression models. Due to their relevance in meta-analysis, such models have received much attention in the last two decades, and several methodological alternatives are available at present for the estimation and statistical testing of the main parameters in a mixed-effects meta-regression model. The goal in the present dissertation was to compare several procedures to fit mixed-effects meta-regression models under different realistic scenarios, with the aim to help meta-analysts to make the best choice depending on the specific conditions of their databases. With this aim, several Monte Carlo simulation studies to be presented in subsequent chapters were carried out.

1.4 The Monte Carlo method: Applications to meta-analysis

The Monte Carlo method constitutes a very useful tool for researchers interested in comparing the properties of several statistical procedures, when analytical treatment is not feasible (Schulze, 2004). It is a method usually applied to simulation studies. In a Monte Carlo simulation study, several data sets are independently created by random number generation, using functions based on probability distributions typically implemented in any statistical package (Burton, Altman, Royston, & Holder, 2006). The strategy of simulating real data from random number generators is the main feature of the Monte Carlo method. When programming a Monte Carlo simulation study, the researcher must define in advance the mathematical distribution and parameters from which the random numbers will be obtained. In a second step, the procedures intended to compare are applied to each data set. Typically, the procedures under study are applied to a very large number of data sets, which requires iterative computations. To this respect, the rapid improvement of computers has supposed a great help for researchers, and Monte Carlo simulation studies are nowadays much less time-consuming than they used to be just at the end of the past century (F. Marín-Martínez, personal communication, December 11, 2007).

Simulation studies implementing the Monte Carlo method can provide empirical estimation that could not be achieved in any other way (Serlin, 2000). Since the true values are determined at first, these studies allow researchers to obtain accuracy measures about the parameter estimates and/or their corresponding statistical tests (Burton et al., 2006; Skrondal, 2000). Modeling the data distribution also allows the researcher to explore the performance for methods which require a set of assumptions, when some of them are not met (Harwell, 1992; Serlin, 2000). In sum, these studies can be considered as experiments where the goal is to analyze the "behavior" of the statistical methods of interest under different scenarios.

In any simulation study, the data must be generated within the framework of a prespecified model, and the set of levels of the manipulated factors is finite. As a consequence, conclusions must be restricted to the model/s and conditions accounted for by the simulation (Schulze, 2004; Skrondal, 2000). Therefore, decisions affecting the generation method for each data set should be made with the aim to include a wide range of realistic scenarios, that is, populations from which researchers are likely to extract their samples (Serlin, 2000).

Because it is a relatively young methodology, there are several issues in meta-analysis where consensus has not been reached yet, and different procedures are available to the meta-analyst at each stage of the statistical analyses. This situation makes it necessary to carry out simulation studies in order to find out which techniques can be expected to perform appropriately given the characteristics of a meta-analytic database, which is intended to help the growing community of applied researchers conducting meta-analytic reviews. Assessing the properties of different methods applied to meta-analysis implies programming simulations where each data set contains data from a whole meta-analysis.

The present dissertation is structured as a set of three empirical studies comparing different methodological alternatives when conducting moderator analyses, by fitting

mixed-effects meta-regression models. The way data were simulated and the comparative criteria for the different procedures vary from one simulation to another, as it will be seen later on.

Chapter 2

Outcome variables in meta-analysis

2.1 Effect sizes

Despite a set of studies have analyzed the same topic, they may have used different measurement instruments (e.g., different psychological tests), different statistical analyses, or both. To accomplish the purpose of a meta-analysis, the result of each single study has to be put into the same metric, so that all of the outcomes are readily comparable (Viechtbauer, 2008). This can be done by means of effect sizes, which allow meta-analysts to put results from all studies into a common scale. Therefore, effect sizes are typically the dependent variable, or the outcome variable, in a meta-analysis.

Throughout this section, an overview of the different situations where an effect size index can be computed will be presented, together with a comprehensive definition. Secondly, the requirements of effect sizes to be used as dependent variables in meta-analysis will be briefly discussed. Lastly, effect sizes will be presented as an alternative to significance tests. The next two sections will address the effect size indices considered in the empirical part of this dissertation.

2.1.1 Conceptualization and definition of effect size

It is very common to use the term "effect size" to refer to the outcome variable in a meta-analysis. However, the effect size can be operationalized in very different ways from a meta-analysis to the next. This is because the outcome variable of interest in a meta-analysis depends on such factors as the question that the meta-analysis intends to address, the design implemented in the single studies, and how the relevant variables have been measured in the studies. As a consequence, the effect size extracted from each single study can represent very different parameters from a meta-analysis to the next.

In Psychology and related areas, dependent variables are mostly continuous, and it is common to find study designs which entail a group comparison. For example, in a meta-analysis that intends to determine the effectiveness of a psychological treatment by integrating studies that compared a treatment and a control group on a continuous dependent variable at the posttest, the result of each study can be defined and quantified by means of a standardized mean difference. If the single study applied a pretest-posttest one-group design, then a different effect size index will have to be calculated, such as a standardized mean change. In addition, studies can implement a more complex design, including two groups with pretest and posttest measures, in which case the effect size index will be a standardized difference between the mean change scores of the two groups (Borenstein, 2009; Morris, 2008; Morris & DeShon, 2002).

In addition to the comparison of two or more groups, another common purpose in psychological research is the analysis of the association between two continuous variables. Some examples of fields where this kind of analysis is usually found are Heritability and Organizational Psychology studies. When the aim of the meta-analysis is to analyze the degree of association among two continuous variables, a correlation coefficient is an optimal effect size index (Borenstein, 2009). The Pearson correlation coefficient is the most popular one, although depending on how the variables have been

measured, other alternatives are also available to the researcher (e.g., Spearman correlation for ordinal variables).

Dichotomous dependent variables are less frequently employed in psychological research, but they can easily be found in sciences like Medicine. In a typical situation, a treatment and a control group are compared in terms of the proportion of occurrence of an event in each group at the posttest (e.g., deaths, recoveries from a disease, etc.). Several effect size indices can be computed for such designs, including the *phi* coefficient, the difference between two probabilities, the risk ratio, and the odds ratio (Fleiss & Berlin, 2009). The selection of the effect size index in this case will depend on the design implemented in the study, such as randomized two-group designs, cohort studies, or case-control studies (Fleiss & Berlin, 2009). Those indices can also be computed for individual studies in which one or more continuous dependent variables were dichotomized (Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003).

The aforementioned scenarios mostly refer to meta-analyses whose purpose is to assess the effectiveness of treatments, programs, or interventions. However, there are other questions that a meta-analysis can address. For example, many meta-analyses have been carried out with the purpose to assess some psychometric property of the scores from a test administration, such as the criterion validity, for which the effect size is typically a correlation coefficient between the test scores and an external criterion (Hunter & Schmidt, 2004). Other meta-analyses have assessed the reliability a given test in different applications. This property is mostly estimated by computing a reliability coefficient for each study such as, for example, a coefficient alpha to assess the internal consistency of the scale, a Pearson correlation between two applications of the test (test-retest reliability) or between two parallel forms of the test, or a concordance coefficient to assess interrater reliability. The so-called *reliability generalization* (Vacha-Haase, 1998) approach will be described with more detail in Section 6.1 of this dissertation (see also Sánchez-Meca, López-López, & López-Pina, in press).

In the Health Sciences, there is another kind of meta-analysis whose objective is to assess the precision of a diagnostic test when a measurement instrument is used to screen a population to detect cases (e.g., participants) with a given event such as, for example, to have a disorder. In this case, the effect size from each single study is defined in terms of the sensitivity and the specificity exhibited by the test (e.g., Tatsioni et al., 2005; Walter & Jadad, 1999). One of the most recommended methods when assessing the accuracy of a diagnostic instrument is the meta-analysis of receiver operating (ROC) curves (Botella, Suero, & Huang, 2012, July; Chappell, Raab, & Wardlaw, 2009).

Finally, in other cases the purpose of a meta-analysis is to estimate the proportion of cases with an event in the population. For example, a meta-analysis can be interested in estimating the prevalence of a disorder in the population, and how the prevalence rates vary among the studies. Another example is that of a meta-analysis focused on estimating the recidivism rate of delinquents once they have finished their sentences (e.g., Morales, Garrido, & Sánchez-Meca, 2010). In these cases, the effect size index is a proportion or a percentage.

As a consequence of the great variety of faces that the effect size can adopt in meta-analysis, the most comprehensive definition of effect size proposed in the literature is that recently published by Kelley and Preacher (2012): "Effect size is defined as a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (p. 140). Therefore, the question of interest might refer to "... central tendency, variability, association, difference, odds, rate, duration, discrepancy, proportionality, superiority, or degree of fit or misfit, among others" (p. 140). An excellent collection of effect sizes available for researchers has recently been elaborated by Grissom and Kim (2012).

2.1.2 Estimation and use of effect sizes in meta-analysis

As Hedges (2007) pointed out, it is important to distinguish the effect size estimate in a given study from the parameter effect in that study. The effect size estimate will vary to some extent from its parameter due to random sampling error, and this variability is accounted for by the within-study variance component in meta-analysis. Therefore, the parameter effect size for a study can be conceived as the value that would have been obtained if researchers conducting that study had been able to measure (without error) the whole population of reference. Regardless of the effect size estimator used in the meta-analysis, it has to exhibit good statistical properties, such as unbiasedness, consistency, and asymptotic efficiency (e.g., Kelley & Preacher, 2012).

The statistical methods typically applied in meta-analysis usually require a normal distribution for the effect sizes and stable sampling variances (Hedges & Olkin, 1985). To accomplish these conditions, in some cases the effect sizes have to be transformed. For example, risk ratios and odds ratios are transformed by their natural logarithm, the Pearson correlation coefficient is transformed into the Fisher's Z, a proportion is transformed into logits, and the coefficient alpha can also be transformed to normalize its distribution and to stabilize the sampling variances by means of the Hakstian and Whalen's (1976) or the Bonett's (2002) transformations.

Moreover, when the meta-analyst has selected a set of studies for the quantitative synthesis, it is not totally uncommon to find out that the computation of a single effect size measure is not feasible for all of them. For those situations, conversion formulae among most of the effect size indices enumerated along this chapter are available (Borenstein, 2009; Fleiss & Berlin, 2009; Sánchez-Meca, Marín-Martínez, & López-López, in press).

2.1.3 Effect sizes as an alternative to significance tests

Conclusions from empirical studies in Psychology and in many other fields are usually guided by the results of significance tests. Two components that will determine the statistical conclusion extracted from a significance test are the effect size and the sample size. This relationship can be expressed as (Rosenthal & DiMatteo, 2001)

Significance test = Effect size x Sample size.

Thus, the p-value associated to a significance test is dependent on the magnitude of the effect under study. At least one effect size index can be estimated for every significance test (Rosenthal, 1994; see also Grissom & Kim, 2012). Many authors have encouraged researchers to compute effect sizes from their study results along the last decades (e.g., Cohen, 1990, 1994; Kirk, 1996; Rosnow & Rosenthal, 2009; Schmidt, 2010; Wilkinson & APA Task Force on Statistical Inference, 1999). As a culmination of the so-called *effect size movement* (Robinson, Whittaker, Williams, & Beretvas, 2003), the American Psychological Association (2010) stated that, when using inferential techniques, effect sizes "are needed to convey the most complete meaning of the results" (p. 33).

In contrast to the dichotomous outcome provided by statistical significance tests (rejection vs. no rejection of the null hypothesis), effect sizes provide information about the magnitude of the relationship of interest (Grissom & Kim, 2012; Schmidt, 2010). Due to this fact, effect sizes allow scientists to draw conclusions about *practical significance* or, in the psychological field, *clinical significance* (Kirk, 1996). It should be remarked, however, that the interpretation of the practical significance of an effect size is strongly dependent on the research area, so that it must be endorsed by some expert opinion (Knapp & Sawilowsky, 2001; Robinson et al., 2003). Another strategy to interpret the magnitude of an effect size in a given research area is to compare it with the results of any meta-analysis published in the same area (Sánchez-Meca, Marín-Martínez, & López-López, 2012, July). If none of these strategies is feasible, then a cautious interpretation of effect

sizes in Psychology can be addressed by employing the standards proposed by Jacob Cohen (1988).

The remaining two sections of this chapter will focus on the effect size indices that will be employed in the simulation studies of this dissertation: the standardized mean difference applied to studies about the effectiveness of psychological treatments, and the coefficient alpha, as it is the reliability coefficient most commonly applied when assessing the reliability of the scores obtained from the administration of a measurement instrument in Psychology and related disciplines.

2.2 Integrating mean differences: The *d* family

Dependent variables are mostly continuous in psychological research, and the most commonly employed study designs entail the comparison of two or more groups in terms of the average scores on some psychological construct, typically measured by means of a test or an interview conducted by an assessor. When the metric of the dependent variable is different from one study to another, it is necessary to standardize results in order to make them comparable from study to study. Consequently, the standardized mean difference is the effect size index most frequently reported or computed *a posteriori* in psychological studies.

A standardized mean difference can be computed to compare two groups (e.g., experimental vs. control group) in terms of their average scores on a continuous dependent variable, usually at the end of an intervention program. This index is defined with the expression

$$\delta = \frac{\mu_E - \mu_C}{\sigma}, \tag{2.1}$$

where μ_E and μ_C represent the mean scores for the experimental and control groups in the population, respectively, and σ is the population pooled standard deviation.

For the *i*th study, assuming the scores of the subjects in the respective groups to be normally distributed, the standardized mean difference can be computed with the expression

$$g_i = \frac{\bar{Y}_{iE} - \bar{Y}_{iC}}{S_i}, \tag{2.2}$$

with $\bar{Y}_{iE}$ and $\bar{Y}_{iC}$ representing the mean scores for the experimental and control groups, respectively, and S_i being the pooled standard deviation that, assuming equal group variances (Ray & Shadish, 1996), is computed with

$$S_i = \sqrt{\frac{(n_{iE}-1)S_{iE}^2 + (n_{iC}-1)S_{iC}^2}{n_{iE}+n_{iC}-2}}, \tag{2.3}$$

with n_{iE} and n_{iC} being the sample sizes for the experimental and control groups, and S_{iE}^2 and S_{iC}^2 representing the variances of the group scores. Then, an unbiased estimator of δ in the *i*th study, d_i, can then be obtained with the expression (Hedges & Olkin, 1985)

$$d_i = \left(1 - \frac{3}{4(n_{iE}+n_{iC})-9}\right) g_i. \tag{2.4}$$

Moreover, an estimate of the sampling variance of d_i can be calculated with

$$\hat{\sigma}_{d_i}^2 = \frac{n_{iE}+n_{iC}}{n_{iE}n_{iC}} + \frac{d_i^2}{2(n_{iE}+n_{iC})}. \tag{2.5}$$

The sampling distribution of the d index is closely related to a non-central t-distribution (Viechtbauer, 2007b), and it is asymptotically normal. Optimal weights for this index can be computed as the inverse of the sampling variances (Borenstein et al., 2010; Marín-Martínez & Sánchez-Meca, 2010; Sánchez-Meca & Marín-Martínez, 1998b) and, for that reason, this will be the weighting scheme employed along this dissertation.

Other related indices that will not be presented here can be computed to assess the standardized mean change of a treatment group (Morris, 2000) or the differential change from pretest to posttest when comparing two groups (Morris, 2008). All of these indices can also be adjusted in order to account for the effect of a covariate (Grissom & Kim, 2012). Lastly, Larry V. Hedges (2011) derived a new index to compute d values in studies with a two-level sampling process where interventions are assigned to entire clusters, a situation that is commonly found in educational research. Regarding the interpretation of the value obtained when computing standardized mean differences, when no better criterion is available, Cohen (1988) proposed values of 0.2, 0.5, and 0.8 as reflecting effects of low, medium, and high magnitude, respectively. The value of the d index can be positive or negative, just depending on how the means in Equation (2.2) are sorted.

2.3 Integrating reliability coefficients: Coefficient alpha and its transformations

In Psychology, standardized tests are the most common measurement instruments. When a test is administered to a sample of subjects, the researcher or clinician must assess the psychometric properties of the sample scores, because these properties can affect results and statistical conclusions based on data obtained with the test (American Psychological Association, 2010; Wilkinson & APA Task Force on Statistical Inference, 1999). Since the psychometric properties will fluctuate from one application of

the test to another, a reasonable approach to obtain representative results for a given test is the integration of the results obtained across different administrations of that instrument. Hunter and Schmidt (1977, 1978, 1983) firstly proposed applying meta-analytic techniques to the integration of validity coefficients obtained across different administrations of the same test.

Apart from different types of validity, another property that must be evaluated when applying a test is the reliability, defined as the consistency or reproductibility of test scores (Anastasi & Urbina, 1997; Crocker & Algina, 1986). As it will be detailed in Section 6.1 of this dissertation, Vacha-Haase (1998) proposed applying meta-analysis to the integration of reliability coefficients obtained in different applications of a psychometric test. In this section, the main reliability measures employed as dependent variables in meta-analysis will be described.

According to the Classical Test Theory (Crocker & Algina, 1986; Gulliksen, 1987), reliability is defined as the quotient between the population variances of the true and observed scores, which can also be expressed as a squared correlation. Since the true scores are unknown in practice, some alternate procedure to estimate the score reliability is needed. Given its computational simplicity, coefficient alpha, considered as a measure of internal consistency (Crocker & Algina, 1986), is the most widely reported reliability indicator in individual studies. For the *i*th sample, a coefficient alpha estimate can be obtained with the expression

$$\hat{\alpha}_i = \frac{N_i}{N_i - 1}\left(1 - \frac{\Sigma \hat{\sigma}_q^2}{\hat{\sigma}_X^2}\right), \tag{2.6}$$

where N_i is the sample size of the *i*th sample, $\hat{\sigma}_q^2$ is the variance of the scores in the *q*th item (any of the items of the test), and $\hat{\sigma}_X^2$ is the variance of the total scores. The sampling variance of $\hat{\alpha}_i$ can be obtained with (Bonett, 2003)

$$\hat{\sigma}^2_{\hat{\alpha}_i} = \frac{2J_i(1-\hat{\alpha}_i)^2}{(J_i-1)\left\{N_i-2-\left[(J_i-2)(k-1)\right]^{1/4}\right\}}, \tag{2.7}$$

where J_i denotes the number of items of the test applied to the *i*th sample and k is the number of studies. Although the statement of a minimum reliability value can be problematic for some situations (cf. Streiner, 2003), it is generally accepted that a value of 0.7 reveals an appropriate reliability of the scores obtained in a given application of the test (Nunnally & Bernstein, 1994).

Standard meta-analytic techniques require some assumptions. One of these assumptions is that the parameters whose estimates are integrated in the meta-analysis are normally distributed, at least for large samples (Hedges, 2009). One problem arising when integrating reliability coefficients is that their sampling distribution is usually asymmetric. For that reason, it seems sensible to apply some transformation on the reliability coefficients prior to the statistical analyses. Some reliability measures (e.g., test-retest, parallel forms) are computed as a correlation, so that a suitable procedure to transform these coefficients is Fisher's Z, which was proposed as a method to normalize the distribution of Pearson correlations (Viechtbauer, 2007b). The Fisher's Z transformation is computed with the formula

$$Z_i = \frac{1}{2}\log_e\left(\frac{1+\hat{\alpha}_i}{1-\hat{\alpha}_i}\right), \tag{2.8}$$

and the sampling variance of this transformation is given by

$$\sigma^2_{z_i} = \frac{1}{N_i-3}. \tag{2.9}$$

When comparing Equations (2.7) and (2.9), it can be readily seen that variances computed for the Fisher's Z transformation will be more stable than those obtained for untransformed alpha coefficients (Sánchez-Meca, López-Pina, & López-López, 2009).

Although Fisher's Z has been the most frequently employed transformation for the meta-analytic integrations of alpha coefficients published up to date (cf. Sánchez-Meca, López-Pina, & López-López, 2008), that transformation is theoretically appropriate only when the reliability coefficients are computed as a Pearson correlation, and that is not the case for alpha coefficients. For that reason, Rodriguez and Maeda (2006) recommended the use of a transformation firstly proposed by Feldt (1969) for two samples and extended by Hakstian and Whalen (1976) for k samples. This transformation is obtained with the expression

$$T_i = \sqrt[3]{1-\hat{\alpha}_i}. \tag{2.10}$$

The sampling variance of this transformation is computed with the formula

$$\hat{\sigma}^2_{T_i} = \frac{18J_i(N_i-1)(1-\hat{\alpha}_i)^{2/3}}{(J_i-1)(9N_i-11)^2}. \tag{2.11}$$

Finally, another transformation was proposed more recently by Bonett (2002), in order to compensate for the fact that tests and confidence intervals for alpha are based on the usually unrealistic assumption (required for alpha coefficients) that the J parts of the test are parallel. This method consists of a logarithmic transformation computed with

$$L_i = Log_e\left(1-\left|\hat{\alpha}_i\right|\right), \tag{2.12}$$

with sampling variance

$$\sigma^2_{L_i} = \frac{2J_i}{(J_i-1)(N_i-2)}. \tag{2.13}$$

When some transformation is applied, the statistical results in a meta-analysis are not directly comparable to those obtained using raw coefficients as the dependent variable (Aguinis et al., 2011). To account for that issue, formulas for back-transformation

are available for the aforementioned transformations. For Fisher's Z, values can be back-transformed to the metric of the original coefficients with the expression

$$\hat{\alpha}_i = \frac{e^{2Z_i} - 1}{e^{2Z_i} + 1}. \tag{2.14}$$

For the Hakstian-Whalen transformation, the equation is

$$\hat{\alpha}_i = 1 - T_i^3. \tag{2.15}$$

Lastly, when the Bonnet's transformation was employed, back-transformation is given by

$$\hat{\alpha}_i = 1 - e^{\zeta_i}. \tag{2.16}$$

Equations (2.14) to (2.16) can be applied to mean coefficients alpha and their confidence limits, the intercept in a regression model, and the mean reliability values for each category in an ANOVA. However, to back-transform regression slopes into the original metric, a different strategy is required, given that the value obtained by a simple back transformation of the slope could be misguiding when Y_i^T is not a linear transformation of coefficient alpha. An alternative, based on the definition of the slope as the amount of change on the dependent variable as the predictor increases in one unit, is outlined below.

Let $Y_i^T = \beta_0^T + \beta_1^T X_i$ be a regression model where Y_i^T is a transformed reliability coefficient, β_0^T and β_1^T are the model coefficients expressed in the transformation metric, and X_i is a predictor. If X_i is set to values of 0 and 1, then two different predictions are obtained for the criterion, $Y_i^T[0]$ and $Y_i^T[1]$, and the slope can be regarded as the difference between both predicted values, that is:

$$Y_i^T[1] - Y_i^T[0] = \beta_0^T + \beta_1^T(1) - \left|\beta_0^T + \beta_1^T(0)\right| = \beta_1^T. \tag{2.17}$$

In Equation (2.17), both the predicted values and the slope are in the metric of the transformation. An alternative for reporting the slope in the metric of the original reliability coefficient, β_1^B, is to calculate the difference between the back transformations of the predicted values $Y_i^B[0]$ and $Y_i^B[1]$, using one of the aforementioned formulae. (Note that this procedure provides a result different to the simple back-transformation of the slope).

Chapter 3

Mixed-effects meta-regression models

3.1 The model

In a meta-analysis with k independent studies, let $\mathbf{y}$ denote a $(k \times 1)$ vector of effect sizes $\{y_i\}$, and $\mathbf{X}$ a $[k \times (p+1)]$ design matrix of full column rank with p predictor variables. The common practice in a meta-regression model is to assume effect sizes to be a random-effects variable, allowing for a broader generalizability of results (see Chapter 1 of this manuscript); on the other hand, the estimation method of the model coefficients proposed by Hedges (1982) requires to assume the predictor variables as fixed effects. This leads to a mixed-effects model, which can be expressed with the formula (Raudenbush, 1994)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \mathbf{e}, \tag{3.1}$$

where $\boldsymbol{\beta}$ is a $[(p+1)\text{x}1]$ vector containing the regression coefficients $\{\beta_0, \beta_1, ..., \beta_p\}$, $\mathbf{u}$ is a $(k\text{ x }1)$ vector of independent between-studies errors $\{u_i\}$ with distribution $N(0, \tau^2_{res})$, and $\mathbf{e}$ is a $(k\text{ x }1)$ vector of independent within-study errors $\{e_i\}$, each of them with distribution $N(0, \sigma^2_{y_i})$. While $\sigma^2_{y_i}$ is the within-study variance (or random sampling error) for the *i*th study, τ^2_{res} represents the residual heterogeneity (or between-studies) variance, that is, the remaining heterogeneity different to sampling error after adding one or more predictor variables to the model (Viechtbauer, 2008).

Note that the mixed-effects model presented in Equation (3.1) is actually an extension of the random-effects model, and that the latter could be formulated if $\mathbf{X}$ is defined as a $(k\text{ x }1)$ vector of ones. In this case, $\boldsymbol{\beta}$ is now a scalar containing the hypermean (mean of the parameter effects), and $\mathbf{u}$ is normally distributed with mean 0 and variance τ^2. For the remainder of this dissertation, τ^2 will be referred to as the total heterogeneity variance, that is, the heterogeneity variance in a model without predictors. If, moreover, the error term $\mathbf{u}$ is suppressed from Equation (3.1), then the model becomes a fixed-effect model.

Regression coefficients $\{\beta_0, \beta_1, ..., \beta_p\}$ can be estimated using the weighted least squares formula

$$\mathbf{b} = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{W}}\mathbf{y}, \tag{3.2}$$

where $\hat{\mathbf{W}}$ is a $(k\text{ x }k)$ diagonal matrix with the inverse sampling variances of the k effect sizes as elements, that is, $\{1/(\hat{\sigma}^2_{y_i} + \hat{\tau}^2_{res})\}$ under a mixed-effects model. Nearly unbiased estimators of $\sigma^2_{y_i}$ are available for all of the meta-analytic outcome variables presented in Chapter 2, and therefore $\sigma^2_{y_i}$ is usually assumed as known in meta-analysis. Assuming within-study variances to be known will work reasonably well as long as sample sizes from each study are not too small (Hedges & Pigott, 2004; Knapp, Biggerstaff, & Hartung, 2006).

Conversely, there are at least seven different estimators for τ^2_{res}, and no consensus has been reached yet in the scientific community about which one works best. These procedures will be described in the next section.

3.2 Residual heterogeneity variance estimators

Several alternatives have been proposed in the literature for the estimation of the total heterogeneity variance, τ^2, in random-effects models (Sánchez-Meca & Marín-Martínez, 2008; Sidik & Jonkman, 2005b, 2007; Viechtbauer, 2005). Most of these estimators have also been extended to the mixed-effects model, allowing for the estimation of the residual heterogeneity variance, τ^2_{res}. It is important to remark here that, for both random- and mixed-effects models, no estimator is expected to provide accurate results unless the number of studies is large enough (Aguinis et al., 2011; Bonett, 2008, 2009; Borenstein et al., 2010; Brockwell & Gordon, 2001, 2007; Hardy & Thompson, 1996). Since τ^2_{res} is included in mixed-effects weights, obtaining accurate estimates of this parameter constitutes a crucial issue in mixed-effects meta-regression models (Biggerstaff & Tweedie, 1997; Sidik & Jonkman, 2005a).

In this section, seven different estimators of τ^2_{res} for mixed-effects models are described. Four of these estimators are non-iterative, while three require iterative computations. When an iterative procedure is implemented, a starting value must be assigned to the parameter of interest, and then an adjustment value, Δ, is added to the initial estimate. This process can continue until Δ is smaller than some preset threshold (e.g., when $\Delta < 10^{-5}$), although a limit for the number of iterations can also be set for situations where convergence is never achieved. Adjustment formulae presented here for these estimators are based on the Fisher scoring algorithm, which is robust towards poor

starting values (Jennrich & Sampson, 1976) and whose computational agility usually leads to convergence quickly (Harville, 1977).

All of the estimators to be presented along this section can be succinctly expressed after defining the matrix

$$\mathbf{P} = \mathbf{W} - \mathbf{WX}(\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{W}, \tag{3.3}$$

where $\mathbf{W}$ is a diagonal weighting matrix whose elements can change from one estimator to another, as further detailed below. Note that all equations presented in this section also apply to the random-effects model, by setting $p = 0$ and with $\mathbf{X}$ being a $(k \times 1)$ vector of ones.

A value of zero for $\hat{\tau}^2_{res}$ suggests that all variability among the true effect sizes is accounted for by the predictors included in the model (Viechtbauer, 2007a). Also, due to random sampling error, most of the estimators to be presented can provide a negative estimate, which is a value outside of the parameter space for a variance component. For non-iterative estimators, the usual practice is to truncate negative values to zero. When an iterative estimator is employed, truncation is also feasible, although a simple (and preferable) strategy to avoid negative estimates is the use of step-halving (Jennrich & Sampson, 1976), which implies multiplying the adjustment value, Δ, by 1/2 (e.g., first by 1/2, then by 1/4, then by 1/8, and so on) until it becomes small enough, such that the resulting estimate stays non-negative.

Moreover, as it will be seen on the basis of the set of equations to be presented in this section, the underlying logic for all methods is to estimate the residual heterogeneity as the difference between the total variability among the true effect sizes not accounted for by the explanatory variables included in the model, which can be quantified with the heterogeneity statistic Q_E (Hedges, 1982), and the variability expected from random

sampling error alone, whose value is usually related to the degrees of freedom of the model under assessment (e.g., $df = k - p - 1$).

The Q_E statistic allows the meta-analyst to determine whether the model is well specified or if, on the contrary, there is a significant amount of unexplained heterogeneity among the effect sizes indicating the influence of additional moderators, additional random heterogeneity, or both (Viechtbauer, 2008). The Q_E statistic is obtained with the expression

$$Q_E = \mathbf{y}' \hat{\mathbf{P}} \mathbf{y}, \tag{3.4}$$

with $\hat{\mathbf{P}}$ defined in Equation (3.3). The Q_E statistic is an extension of the homogeneity test usually computed for the assessment of the heterogeneity among effect sizes in meta-analysis, Q, which was mentioned in Chapter 1 of this dissertation. Indeed, the Q test can be computed with Equation (3.4) with $\mathbf{X}$ being a $(k \times 1)$ vector of ones. Note that, when calculating Q or Q_E, the diagonal elements of $\hat{\mathbf{W}}$ are given by $\hat{w}_i = 1/\hat{\sigma}^2_{y_i}$, excluding the heterogeneity variance component (Beretvas & Pastor, 2003; Borenstein et al., 2009; Hartung et al., 2008).

Under the null hypothesis $\tau^2_{res} = 0$, the Q_E statistic follows a chi-square distribution with degrees of freedom equal to $df = k - p - 1$. The rejection of the null hypothesis would indicate a model misspecification, with a statistically significant heterogeneity unexplained by the predictors in the model (Aguinis & Pierce, 1998). Nonetheless, the Q_E test suffers from the same problems mentioned in Chapter 1 for the Q statistic regarding statistical power (e.g., Pereira et al., 2010; Sánchez-Meca & Marín-Martínez, 1997), so that a cautious interpretation of the statistical conclusion of Q_E should be recommended for most situations.

3.2.1 Hedges (HE) estimator

Hedges (1983; see also Hedges & Olkin, 1985) proposed a method of moments estimator of τ^2 for random-effects models based on ordinary least squares (OLS) estimation. The estimate is obtained by calculating the difference between an unweighted estimate of the total variance of the effect sizes and an unweighted estimate of the average within-study variance (Sánchez-Meca & Marín-Martínez, 2008). In a simulation study comparing the bias and efficiency of different τ^2 estimators in random-effects models, Viechtbauer (2005) found that the HE estimator was almost unbiased for most conditions, although it was less efficient than other procedures (HS, DL, ML, and REML estimators) that will also be presented further below.

When moderators are included in the model, the extension of the Hedges method for the estimation of the residual heterogeneity variance, τ^2_{res}, can be written as (Raudenbush, 2009)

$$\hat{\tau}^2_{HE} = \frac{\mathbf{y}'\mathbf{P}\mathbf{y} - tr(\mathbf{P}\mathbf{V})}{k - p - 1}, \tag{3.5}$$

with $tr()$ denoting the trace of the matrix in between the parentheses, $\mathbf{V}$ denoting a diagonal matrix with elements $\hat{\sigma}^2_{y_i}$ and with $\mathbf{W}$ equal to a $(k \times k)$ identity matrix $\mathbf{I}$ for the calculation of $\mathbf{P}$, in which case Equation (3.3) simplifies to $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

3.2.2 Hunter and Schmidt (HS) estimator

Hunter and Schmidt (2004) proposed an estimator of τ^2 for random-effects models which, in essence, is given by

$$\hat{\tau}^2_{HS} = \frac{\sum \hat{w}_i (y_i - \hat{\mu})^2}{\sum \hat{w}_i} - \frac{\sum \hat{w}_i \hat{\sigma}^2_{y_i}}{\sum \hat{w}_i} = \frac{\sum \hat{w}_i (y_i - \hat{\mu})^2 - k}{\sum \hat{w}_i}, \quad (3.6)$$

where $\hat{\mu} = \sum \hat{w}_i y_i / \sum \hat{w}_i$ and $\hat{w}_i = 1/\hat{\sigma}^2_{y_i}$. In this case, the HS estimator is equivalent to the difference between a weighted estimate of the total variance of the effect sizes and a weighted average of the within-study variances. In his simulation study, Viechtbauer (2005) found a negative bias for the HS estimator, which performed reasonably well in terms of efficiency for most conditions.

Although no extension has been suggested yet for the HS estimator when one or more covariates are included in the model, a logical proposal for computing this estimator in mixed-effects models is given by (Viechtbauer, López-López, Sánchez-Meca, & Marín-Martínez, 2012)

$$\hat{\tau}^2_{HS} = \frac{\mathbf{y}' \hat{\mathbf{P}} \mathbf{y} - k}{tr(\hat{\mathbf{W}})}, \quad (3.7)$$

with $\hat{\mathbf{P}}$ again defined in Equation (3.3), and the diagonal elements of $\hat{\mathbf{W}}$ given by $\hat{w}_i = 1/\hat{\sigma}^2_{y_i}$.

3.2.3 DerSimonian and Laird (DL) estimator

The estimator of τ^2 proposed by DerSimonian and Laird (1986) for random-effects models, probably the most widely employed in meta-analyses up to date, is also based on the method of moments (DerSimonian & Kacker, 2007). Although it generally constitutes a reasonable alternative for the estimation of the total heterogeneity variance in meta-analysis, this method has shown some problems in previous simulations, especially when the parameter value was very large, which led to negatively biased estimates, and when

the within-study variances were not homogeneous (Malzahn, Böhning, & Holling, 2000; Sidik & Jonkman, 2005b, 2007; Viechtbauer, 2005).

When one or more covariates are included in the model, the DL estimator is given by

$$\hat{\tau}^2_{DL} = \frac{\mathbf{y}'\hat{\mathbf{P}}\mathbf{y} - (k - p - 1)}{tr(\hat{\mathbf{P}})}, \tag{3.8}$$

with $\hat{\mathbf{P}}$ defined in Equation (3.3) and the diagonal elements of $\hat{\mathbf{W}}$ again given by $\hat{w}_i = 1/\hat{\sigma}^2_{y_i}$.

3.2.4 Sidik and Jonkman (SJ) estimator

Another alternative to estimate the residual heterogeneity variance was proposed by Sidik and Jonkman (2005b). In their simulation study comparing DL and SJ estimators, these authors reported a positive bias for the SJ method, which decreased for larger parameter values. Due to the negative bias of the DL estimator for large τ^2 values, as mentioned in Section 3.2.3, the SJ estimator was found to be a good alternative for parameter values equal or greater than 0.50 in random-effects models using log-odds ratios as the effect size index (cf. Sidik & Jonkman, 2005b). The SJ estimator is obtained by starting with an initial (rough) estimate of the heterogeneity variance, denoted by $\hat{\tau}^2_0$ and given by

$$\hat{\tau}^2_0 = \frac{\sum (y_i - \bar{y})^2}{k}, \tag{3.9}$$

where $\bar{y}$ is an unweighted mean of the effect sizes. In a mixed-effects model, the SJ estimator is computed with the expression

$$\hat{\tau}^2_{SJ} = \frac{\hat{\tau}^2_0(\mathbf{y}'\hat{\mathbf{P}}\mathbf{y})}{k-p-1}, \tag{3.10}$$

with $\hat{\mathbf{P}}$ defined in Equation (3.3) and elements $\{\hat{w}_i = 1/(\hat{\sigma}^2_{y_i} + \hat{\tau}^2_0)\}$ for the diagonal matrix $\hat{\mathbf{W}}$. In contrast to the other procedures presented along this section, the SJ estimator always provides a non-negative value, so that it never requires truncation (cf. Sidik and Jonkman, 2005b).

3.2.5 Maximum Likelihood (ML) estimator

The ML estimator is based on the joint likelihood of the regression coefficients, $\boldsymbol{\beta}$, and the residual heterogeneity variance, τ^2_{res} (Raudenbush, 1994). Since this estimator does not account for the uncertainty about the unknown regression coefficients, it is expected to provide negatively biased estimates in random-effects models, as it was found in several simulation studies (e.g., Sidik & Jonkman, 2007; Viechtbauer, 2005).

The ML estimator requires iterative computations. The process can be expressed with

$$\hat{\tau}^2_{New} = \hat{\tau}^2_{Current} + \Delta, \tag{3.11}$$

where $\hat{\tau}^2_{Current}$ is the current estimate of τ^2, its starting value being that obtained with any of the other (non-iterative) estimators. For maximum likelihood estimation, the adjustment factor in mixed-effects models using the Fisher scoring algorithm is equal to

$$\Delta_{ML} = \frac{\mathbf{y}'\hat{\mathbf{P}}\hat{\mathbf{P}}\mathbf{y} - tr(\hat{\mathbf{W}})}{tr(\hat{\mathbf{W}}\hat{\mathbf{W}})}, \tag{3.12}$$

with $\hat{\mathbf{P}}$ defined in Equation (3.3) and the diagonal elements of $\hat{\mathbf{W}}$ given by $\hat{w}_i = 1/(\hat{\sigma}^2_{y_i} + \hat{\tau}^2_{Current})$. Therefore, after each iteration, $\hat{\mathbf{W}}$ is firstly updated, then $\hat{\mathbf{P}}$, and finally Δ_{ML} can be computed to obtain $\hat{\tau}^2_{New}$.

3.2.6 Restricted Maximum Likelihood (REML) estimator

Another iterative procedure is the REML estimator, which overcomes the negative bias observed in the ML method (Thompson & Sharp, 1999), because it takes into account the uncertainty about the regression parameter estimates (Raudenbush, 1994). On the other hand, several simulation studies found a loss of efficiency in REML compared to ML under a random-effects model (e.g., Sidik & Jonkman, 2007; Viechtbauer, 2005). Moreover, Sidik and Jonkman (2007) found a negative bias in the REML estimator for large values, although the magnitude of that bias was smaller than that obtained for ML and DL estimators, also included in their study. For that reason, these authors warned against the use of DL, ML, and REML estimators unless the heterogeneity variance parameter, τ^2, is expected to be relatively small (Sidik & Jonkman, 2007, p. 1980). It should be noted, however, that most of the values set for τ^2 in the simulation conducted by Sidik and Jonkman would be considered as extremely large in Psychology (range 0 : 1.75 using log-odds ratios as the effect size index).

The REML estimator has been recommended by Raudenbush (2009) for mixed-effects models, where the adjustment for this procedure can be computed with the expression

$$\Delta_{REML} = \frac{\mathbf{y}'\hat{\mathbf{P}}\hat{\mathbf{P}}\mathbf{y} - tr(\hat{\mathbf{P}})}{tr(\hat{\mathbf{P}}\hat{\mathbf{P}})}, \tag{3.13}$$

with $\hat{\mathbf{P}}$ again defined in Equation (3.3) and elements $\left\{\hat{w}_i = 1/(\hat{\sigma}^2_{y_i} + \hat{\tau}^2_{Current})\right\}$ for $\hat{\mathbf{W}}$.

3.2.7 Empirical Bayes (EB) estimator

The last estimator considered in this section was first proposed by Morris (1983) and later adapted to the meta-analytic context (Berkey, Hoaglin, Mosteller, & Colditz, 1995). This estimator can be derived based on empirical Bayes methods (Morris, 1983) and will therefore be denoted by $\hat{\tau}^2_{EB}$. Sidik and Jonkman (2007) reported a good performance for this estimator under a random-effects model in terms of bias and mean squared error. Again, there is no closed-form solution, so that iterative methods must be used. Under a mixed-effects model, the adjustment required for the EB estimator at each iteration is computed with

$$\Delta_{EB} = \frac{k/(k-p-1)\mathbf{y}'\hat{\mathbf{P}}\mathbf{y} - k}{tr(\hat{\mathbf{W}})}, \tag{3.14}$$

with $\hat{\mathbf{P}}$ again defined in Equation (3.3) and elements $\left\{\hat{w}_i = 1/(\hat{\sigma}^2_{y_i} + \hat{\tau}^2_{Current})\right\}$ for $\hat{\mathbf{W}}$.

Furthermore, the EB estimator can be shown to be identical to another estimator, going back to the work of Paule and Mandel (1982), which was recently described in the meta-analytic context by DerSimonian and Kacker (2007). In particular, for the mixed-effects model, the Paule and Mandel's estimator is that value of $\hat{\tau}^2_{res}$ for which

$$\mathbf{y}'\hat{\mathbf{P}}\mathbf{y} = k - p - 1, \tag{3.15}$$

with the diagonal elements of $\hat{\mathbf{W}}$ given by $\hat{w}_i = 1/(\hat{\sigma}^2_{y_i} + \hat{\tau}^2_{res})$. The equivalence between these two estimators leads to some interesting properties to be described further below.

3.3 Hypothesis tests for the model coefficients

Once an estimate of τ^2_{res} has been computed, the vector of model coefficients can be obtained with Equation (3.2). The next step in a meta-regression is then to determine the precision of these estimates and to test whether the moderators actually exert a statistically significant influence on the effect sizes. Five alternatives for testing the regression coefficients are presented below.

The first one is a Wald-type test (Raudenbush, 2009), and it is the one that is most commonly applied in practice. Accordingly, this approach will be referred to as the standard method. Despite its wide use in meta-analysis, some authors argued that this method does not take into account the uncertainty of working with estimated variances, and that might produce misleading findings (Brockwell & Gordon, 2001; Van Houwelingen et al., 2002). To offset that limitation, Knapp and Hartung (2003) developed a new method by incorporating a correction factor to the traditional formula. Also, their method assumes a t-distribution for the coefficient values, instead of a normal distribution. The third method presented in this section makes use of a robust estimate of the variance-covariance matrix of the model coefficients. The fourth method here presented is a likelihood ratio test, which compares the likelihood of the model with and without the predictor of interest. Finally, a permutation test is described. While the latter is computationally more demanding than the other tests, it is, in principle, free of distributional assumptions.

3.3.1 Standard method

If $\sigma^2_{y_i}$ and τ^2_{res} were known, then the variance-covariance matrix of the model coefficients computed with Equation (3.2) would be equal to $\boldsymbol{\Sigma} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$, with the

diagonal elements of $\mathbf{W}$ given by $w_i = 1/(\sigma^2_{y_i} + \tau^2_{res})$. However, since both τ^2_{res} and $\sigma^2_{y_i}$ need to be estimated in practice, we cannot compute $\boldsymbol{\Sigma}$ directly. The standard approach is to substitute the estimates of τ^2_{res} and $\sigma^2_{y_i}$ for the unknown variance components in $\mathbf{W}$, yielding an estimate of $\boldsymbol{\Sigma}$ given by the equation

$$\hat{\boldsymbol{\Sigma}}_{STD} = (\mathbf{X}' \hat{\mathbf{W}} \mathbf{X})^{-1}, \tag{3.16}$$

where the diagonal elements of $\hat{\mathbf{W}}$ are equal to $\hat{w}_i = 1/(\hat{\sigma}^2_{y_i} + \hat{\tau}^2_{res})$. The test statistic for a particular model coefficient can then be obtained with

$$z_j = \frac{b_j}{S^{STD}_{b_j}}, \tag{3.17}$$

with b_j being the $[j+1]$ element of the $\mathbf{b}$ vector, obtained with Equation (3.2), and $S^{STD}_{b_j}$ being the square root of the $[j+1, j+1]$ element of the $\hat{\boldsymbol{\Sigma}}_{STD}$ matrix, computed with Equation (3.16). The value obtained by Equation (3.17) is then compared against the critical values of a standard normal distribution for a desired significance level (e.g., ±1.96 for $\alpha = .05$, two-sided). Although this has been almost the only method employed to test coefficients from mixed-effects meta-regression models up to date, its adequacy is strongly dependent on the accuracy of the sampling variance estimates. Consequently, if those estimates were inaccurate, then the statistical conclusion provided by the standard method might not be correct (Brockwell & Gordon, 2001; Knapp & Hartung, 2003; Sidik & Jonkman, 2005a).

3.3.2 Knapp-Hartung method

The Knapp-Hartung method (Knapp & Hartung, 2003) is based on a corrected estimate of the variance-covariance matrix of the model coefficients, given by

$$\hat{\boldsymbol{\Sigma}}_{KH} = c(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}, \tag{3.18}$$

where

$$c = \frac{\mathbf{y}'\hat{\mathbf{P}}\mathbf{y}}{k-p-1}, \tag{3.19}$$

with $\hat{\mathbf{P}}$ again defined in Equation (3.3) and the diagonal elements of $\hat{\mathbf{W}}$ given by $\hat{w}_i = 1/(\hat{\sigma}^2_{y_i} + \hat{\tau}^2_{res})$. The test statistic for a particular model coefficient is then computed with the expression

$$t_j^{KH} = \frac{b_j}{S_{b_j}^{KH}}, \tag{3.20}$$

with b_j denoting the corresponding element of $\mathbf{b}$, computed with Equation (3.2), and $S_{b_j}^{KH}$ being the square root of the respective diagonal element of $\hat{\boldsymbol{\Sigma}}_{KH}$, obtained with Equation (3.18). Under the null hypothesis $\beta_j = 0$, it is assumed that t_j^{KH} follows a t-distribution with $df = k - p - 1$ degrees of freedom, according to the authors (Knapp & Hartung, 2003). Note, however, that some other values for the degrees of freedom have been proposed (e.g., Berkey et al., 1995).

In their simulation study, using log risk ratios as the dependent variable, Knapp and Hartung (2003) found that their new method outperformed the standard one in terms of adjustment to the nominal significance level. Sidik and Jonkman (2005a) obtained similar results when comparing both methods.

For known variance components, the expected value of the correction factor c is one (Hartung et al., 2008). Also, when using the EB method for estimating τ^2_{res}, presented in Section 3.2.7, c is always equal to one for positive values of $\hat{\tau}^2_{EB}$ (Knapp & Hartung, 2003).

Knapp and Hartung (2003) originally proposed that the correction factor c should always be equal to or greater than one. A value smaller than one is likely to be obtained with Equation (3.19) in scenarios where the effect sizes are very homogeneous, so that the total variability unaccounted for by the moderators, Q_E, is even smaller than its expected value (e.g., $df = k - p - 1$) when $\tau^2_{res} = 0$. However, when working with small samples (e.g., small number of studies, small average number of participants per study, or both), such counterintuitive results can easily happen, since meta-analytic estimates are generally quite inaccurate in those situations (Hedges, 2009).

Following the recommendations provided by Knapp and Hartung (2003), the correction factor c should be truncated to one when a smaller value is obtained. With this practice, the variance estimate of b_j obtained with their method would never be smaller than the one obtained with the standard method, always leading to more conservative tests than those obtained with the standard approach. However, this practice may actually lead to over conservative results and, consequently, to a loss of power, thereby increasing the chance that relevant moderators may be missed. This will be examined with more detail in Chapter 5 of this dissertation.

3.3.3. Huber-White method

The Huber-White method is based on the work of Huber (1967) and White (1980) and was first proposed in the meta-analytic literature by Sidik and Jonkman (2005a). For this method, the variance-covariance matrix of the model coefficients is estimated with

$$\hat{\boldsymbol{\Sigma}}_{HW} = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{W}}\hat{\mathbf{E}}^2\hat{\mathbf{W}}\mathbf{X}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}, \tag{3.21}$$

where $\hat{\mathbf{E}}$ is a diagonal matrix with elements obtained from the vector $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$, and with $\{\hat{w}_i = 1/(\hat{\sigma}^2_{y_i} + \hat{\tau}^2_{res})\}$ as the elements for $\hat{\mathbf{W}}$. The test statistic for a particular model coefficient is then given by Equation (3.20), except that $S^{KH}_{b_j}$ is replaced with $S^{HW}_{b_j}$. Again, the test statistic is compared against the critical values of a t-distribution with $df = k - p - 1$ degrees of freedom.

The Huber-White method did not consistently improve the performance of the standard method in the simulation conducted by Sidik and Jonkman (2005a), using the empirical Type I error rate as the comparative criterion. However, a simple correction was recently proposed (Hedges et al., 2010). Incorporating this proposal to Equation (3.20) leads to the expression

$$t^{HW}_j = \frac{b_j}{S^{HW}_{b_j}\sqrt{k/(k-p-1)}}, \tag{3.22}$$

which yields a more conservative test, especially when k is small. However, it remains to be determined how the Huber-White method with this correction performs in comparison to the other approaches considered in this section. This will also be analyzed in Chapter 5 of this dissertation.

3.3.4 Likelihood ratio test

All of the approaches described so far are based on a test statistic that divides the model coefficient to be tested by some estimate of its standard error. An alternative approach is based on likelihood ratio testing (Bates, 2011, March; Guolo, 2012; Huizenga, Visser, & Dolan, 2011). This approach can be used in the context of ML estimation, and it is based on the change in the deviance of two models, the first including the predictor of interest and the second excluding it [in a meta-regression with a single predictor, the second model would be a random-effects model where $\mathbf{X}$ is a $(k \times 1)$ vector of ones]. The likelihood ratio test is computed with the expression

$$\chi_j^2 = -2 \ln \frac{L_0}{L_1}, \tag{3.23}$$

where L_0 is the likelihood of the null model and L_1 is the likelihood of the model including the jth moderator. The result is compared against the critical value of a chi-square distribution with one degree of freedom (e.g., 3.84 for $\alpha = 0.05$).

Previous simulation studies (Guolo, 2012; Huizenga et al., 2011) found a performance somewhat less than good for the likelihood ratio test, with empirical Type I error rates clearly over the nominal significance level.

3.3.5 Permutation test

Finally, the use of permutation tests has been suggested as another alternative in the meta-analytic context (Follmann & Proschan, 1999; Higgins & Thompson, 2004). To carry out the test for a particular model coefficient, we first obtain z_j, the test statistic based on the standard approach, given by Equation (3.17). Then, for each of the $k!$

possible permutations of the rows of the $\mathbf{X}$ matrix, the model is refitted and the value of the test statistic is recomputed. Note that each permutation requires that τ^2_{res}, $\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}$ are re-estimated. Letting z_j^m denote the value of the test statistic for the mth permutation, the (two-sided) p-value for the permutation test is then equal to two times the proportion of cases where the test statistic under the permuted data is as extreme or more extreme than under the actually observed data. In other words, the p-value is two times the proportion of z_j^m values greater than z_j when z_j is positive, or two times the rate of z_j^m values smaller than z_j if that statistic is negative.

Note that k must be at least as large as 5 before it is actually possible to obtain a p-value below $\alpha = .05$ (e.g., for $4! = 24$ permutations, the p-value can never be smaller than $2 \times 1/24 = .0833$, while for $5! = 120$, the p-value can be as small as .0167). On the other hand, as k increases, $k!$ quickly grows so large that it may not be possible in practice to obtain the full set of permuted test statistics. In that case, one can approximate the exact permutation-based p-value by going through a certain number of random permutations of the rows of the $\mathbf{X}$ matrix. Using a sufficiently large number of such random permutations ensures that the resulting p-value is stable.

The permutation approach may be especially appropriate when the data cannot be regarded as a random sample from a given population (Manly, 1997). Moreover, this method is, in principle, free of distributional assumptions. However, the use of a nonparametric approach may be less efficient than parametric methods, potentially resulting in a lower statistical power.

3.4 Model predictive power

After estimating the model parameters (τ^2_{res}, $\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}$) and testing the significance of the model coefficients, another important objective in mixed-effects meta-regression models is the estimation of the predictive power of the model. This parameter is represented by P^2 (note that P here denotes the capital Greek letter *rho*), and can be defined as the proportion of variance among the true effect sizes that can be accounted for by the predictors included in the model. Therefore, this parameter will only take values between 0 and 1. The estimates of this parameter are usually represented with the R^2 index.

The R^2 index is an effect size measure which complements the statistical conclusion provided by the significance tests of the regression coefficients, presented in the previous section of this chapter. The R^2 index is usually interpreted as a percentage of variance accounted for and, in Psychology and related fields, the guidelines stated by Jacob Cohen (1988) can be followed. According to Cohen, a 10% of variance accounted for by the predictor/s reveals an effect with practical significance of low-medium magnitude, while values around 25% already reflect an effect size of high magnitude. Nevertheless, these orientations should be contextualized and revised in the specific content area of the phenomenon under study (Knapp & Sawilowsky, 2001), with the help of the R^2 effect sizes typically found in the papers and meta-analyses on the topic, as well as the opinion of experts in the area.

When regression models are fitted using OLS techniques, the R^2 index is computed as the quotient between the sum of squares due to the regression and the total sum of squares, that is, $R^2 = SS_{Regression} / SS_{Total}$. However, this strategy is not suitable for meta-regression models because part of the total variability, due to sampling error

(within-study variances, $\sigma_{y_i}^2$), cannot be explained by the predictors in the linear model[4] (Aloe et al., 2010; Konstantopoulos & Hedges, 2009; Rodriguez & Maeda, 2006). In other words, "the only variation that linear models of effect size can explain is this between-studies variation" (Aloe et al., 2010, p. 276). Thus, a different method is required for computing an R^2 index in meta-regression models.

An alternative was proposed by Raudenbush (1994). The R^2 index proposed by Raudenbush is based on the re-estimation of the heterogeneity (or between-studies) variance after adding one or more predictors to the model. The rationale for this index is that the influence of the moderators will be reflected on the residual heterogeneity variance, τ_{res}^2, which will be smaller than the total heterogeneity variance, τ^2, as a result of including explanatory variables accounting for part of that heterogeneity. The comparison of both values provides the percentage of variance explained in the population, P^2, and this criterion can be used to assess the model predictive power (Raudenbush, 2009). In practice, $\hat{\tau}^2$ and $\hat{\tau}_{res}^2$ must be used instead of the population values, allowing for the computation of the R^2 index with the expression (Borenstein et al., 2009)

$$R^2 = 1 - \left(\frac{\hat{\tau}_{res}^2}{\hat{\tau}^2} \right). \tag{3.24}$$

If a negative value is obtained when applying Equation (3.24), it is truncated to zero, and the interpretation is that all of the variability among the true effect sizes remains unexplained after including the moderator(s).

[4] An exception to this is when meta-analyzing the raw data from a set of individual studies, in which case within-study variability can be accounted for. For more details on so-called individual participant data meta-analyses, see, for example, Cooper and Patall (2009).

The method proposed by Raudenbush (1994), therefore, constitutes a reasonable alternative to estimate the model predictive power in mixed-effects meta-regression models. Note, however, that moderator analyses can also be conducted in meta-analysis by assuming a fixed-effect model. For the so-called *fixed-effect models with moderators*, Konstantopoulos and Hedges (2009) suggested that Equation (3.24) could also be implemented. In that case, the total heterogeneity variance, τ^2, which is estimated in the model without predictors (fixed-effect model), is assumed to be wholly due to the influence of one or more unidentified moderators. Regarding the residual heterogeneity variance, τ^2_{res}, it reflects the influence of one or more additional moderators that were not included in the fixed-effects meta-regression model. The same rationale can also be applied if moderator analyses are conducted by assuming a varying coefficient model.

Both heterogeneity variance estimates employed in Equation (3.24) can be obtained using any of the methods presented in Section 3.2. As a consequence, there are at least seven different methods to compute the R^2 index using this proposal, if the same estimation method is employed for both estimates, as recommended before (Aloe et al., 2010). It is important to note that, due to sampling error, the formula proposed by Raudenbush may require or lead to truncation in several situations. First, $\hat{\tau}^2_{res}$ can be larger than $\hat{\tau}^2$ for a given meta-analytic data set, especially with small samples (small number of studies, small sample sizes, or both), leading to a negative R^2 value that is typically truncated to zero in practice (indicating that all of the heterogeneity among the effect sizes remains unaccounted for after including the moderator(s) in the model). Second, a negative value of $\hat{\tau}^2$ truncated to zero leads to division by zero in Equation (3.24), in which case R^2 is undefined. It is then common practice to set (or truncate) the value of R^2 to 0 (indicating that none of the heterogeneity among the effect sizes is accounted for by the moderators, given that there appeared to be none to begin with). Finally, with a positive value of $\hat{\tau}^2$, a negative value of $\hat{\tau}^2_{res}$ truncated to zero will lead to

an R^2 value of 1 (indicating that all of the heterogeneity among the effect sizes is accounted for the moderators included in the model).

Since an estimate of the heterogeneity variance is included in both the random- and the mixed-effects model weights, the accuracy of these estimates might also affect the result of other statistical analyses, such as the computation of an overall effect size estimate and its confidence interval in a random-effects model or the estimation and testing of the model coefficients in a mixed-effects meta-regression model. However, getting accurate estimates of τ^2 and τ^2_{res} seems even more crucial for the assessment of the predictive power in meta-regression models, because the R^2 index computed with Equation (3.24) requires estimates both of the total and the residual amount of heterogeneity (and hence, any error in these estimates may compound). The performance of the different methods for calculating R^2 will be considered in further detail in Chapter 4 of this dissertation.

Chapter 4

Study 1: Assessing predictive power in mixed-effects meta-regression models

4.1 Objectives, previous simulation studies, and hypotheses

4.1.1 Objectives of the study

The availability of different procedures to estimate the heterogeneity variance in both random- and mixed-effects models poses a problem to the meta-analyst, because the estimator choice may have an influence on the meta-analysis results. Since an estimate of the heterogeneity variance is included in both random- and mixed-effects weights, the accuracy of these estimates might affect the result of statistical analyses such as the computation of an overall effect size estimate and its confidence interval in a random-effects model, or the estimation and testing of the model coefficients in a mixed-effects meta-regression. Getting accurate estimates of the heterogeneity variance seems even more crucial for the assessment of the predictive power in meta-regression models

which, when using the procedure proposed by Raudenbush (1994; see also Section 3.4 of this dissertation), takes into account both the total and residual heterogeneity variance estimates.

In the present study, all seven heterogeneity variance estimators detailed in Section 3.2 were considered (that is, HE, HS, DL, SJ, ML, REML, and EB methods), and applied to simulated meta-analyses where the standardized mean difference, already defined in Equation (2.4), was the effect size index. This simulation compared the accuracy for the methods under different scenarios for the estimation of:

- The total heterogeneity variance in a random-effects model: τ^2 parameter.
- The residual heterogeneity variance in a mixed-effects meta-regression model with one predictor: τ^2_{res} parameter.
- The predictive power of a mixed-effects meta-regression model with one predictor, using the proposal of Raudenbush (1994; see also Section 3.4): P^2 parameter.

4.1.2 Previous simulation studies

Several simulation studies have already been conducted with the aim to compare the accuracy of various estimators of the heterogeneity variance in meta-analysis. Some of these studies employed effect size indices for dichotomous measures (e.g., Malzahn et al., 2000; Sidik & Jonkman, 2005b, 2007), while others considered indices for continuous variables (e.g., Van den Noortgate & Onghena, 2003; Viechtbauer, 2005).

In general, a positive bias has been found in the SJ estimator for small to medium parameter values (Sidik & Jonkman, 2005b, 2007), while a negative bias was reported for

the HS and ML estimators, as well as for the DL method when estimating large parameter values (Malzahn et al., 2000; Viechtbauer, 2005). The HE method was found to perform appropriately in terms of bias, although it was less efficient than other estimators (Viechtbauer, 2005). Finally, good performance was observed both for the REML and the EB estimators when considering bias and efficiency criteria jointly (Sidik & Jonkman, 2007; Van den Noortgate & Onghena, 2003; Viechtbauer, 2005).

All of these simulation studies focused on random-effects models. Therefore, it is not certain to what extent these trends would also apply to mixed-effects meta-regression models. Moreover, these studies do not indicate whether one of the various estimators for τ^2 and τ^2_{res} would be preferable when computing the R^2 index computed with Equation (3.24). For example, even though biases have been found in some of the heterogeneity estimators, since R^2 is based on the ratio of the residual and total amount of heterogeneity, it is not possible to predict whether these biases would carry over when computing R^2 or may in fact essentially cancel each other out.

4.1.3 Hypotheses of this study

Due to the results showed in previous simulations, it was expected that:

1. The SJ method would provide positively biased estimates of the heterogeneity variance in random-effects models, improving its performance for large parameter values, as reported by the authors (Sidik & Jonkman, 2005b, 2007).

2. The HS and ML methods would provide negatively biased estimates of the heterogeneity variance in random-effects models, as it was previously found (Viechtbauer, 2005).

3. The HE method would perform inefficiently when estimating the heterogeneity variance in random-effects models, as it was reported before (Viechtbauer, 2005).

4. The DL method would provide negatively biased estimates of the heterogeneity variance in random-effects models for large parameter values, as warned by several authors (e.g., Malzahn et al., 2000; Sidik & Jonkman, 2005b).

5. The REML method would perform appropriately (in terms of bias and efficiency) for the estimation of the heterogeneity variance in random-effects models, as reported by Viechtbauer (2005).

6. The EB method would perform reasonably well (in terms of bias and efficiency) for the estimation of the heterogeneity variance in random-effects models, as previously found (Sidik & Jonkman, 2007).

7. The trends for the different estimators under a random-effects model would be similar when estimating the residual heterogeneity variance in mixed-effects meta-regression models with one predictor.

8. The most precise methods in the estimation of the heterogeneity variances (DL, REML, and EB methods) would also be the most accurate options when estimating the predictive power of mixed-effects meta-regression models with one predictor.

9. The HS, ML, and SJ methods would provide biased estimates of the predictive power of meta-regression models with one predictor.

10. The number of studies would exert the greatest influence on the accuracy of the different methods, and its influence would be even more critical when estimating the model predictive power, which is computed as a ratio between two heterogeneity variance estimates.

11. A larger number of participants per study would lead to more accurate results in the estimation of the different parameters considered in this study.

4.2 An illustrative example

Else-Quest, Hyde, and Linn (2010) published a meta-analysis integrating results from the Program for International Student Assessment (PISA) in different countries in 2003. This report evaluated 15-year old students' performance in several subjects. The authors focused on mathematics and, since they were interested in gender differences, effect sizes were defined as standardized mean differences between the marks achieved by boys and girls. Positive values revealed a better performance for boys.

One of the coded characteristics for each country was the women's share of parliamentary seats, used as a moderator in this example. Twenty countries from different parts of the world were selected to illustrate the methods described earlier. Table 4.1 presents the effect size, d_i, its sampling (within-country) variance estimate, $\hat{\sigma}^2_{d_i}$, and the moderator value, $Parl_i$, for each of the 20 countries.

The set of effect sizes reported in Table 4.1 ranged from -0.17 to 0.25. These values were obtained for Icelandic and South Korean students, respectively. The women's share of parliamentary seats ranged between the 4% found for Turkish politicians and the 45% obtained for their Swedish colleagues.

Table 4.1 Data from the meta-analysis published by Else-Quest and colleagues (2010)

Country	d_i	$\hat{\sigma}^2_{d_i}$	$Parl_i$	*Country*	d_i	$\hat{\sigma}^2_{d_i}$	$Parl_i$
Australia	0.06	.0003	0.27	Mexico	0.13	.0001	0.16
Belgium	0.07	.0005	0.25	The Netherlands	0.06	.0010	0.33
Brazil	0.16	.0009	0.09	Poland	0.06	.0009	0.21
Canada	0.13	.0002	0.24	South Korea	0.25	.0008	0.06
France	0.09	.0009	0.12	Spain	0.10	.0004	0.27
Germany	0.09	.0009	0.31	Sweden	0.07	.0009	0.45
Greece	0.21	.0009	0.09	Thailand	-0.05	.0008	0.10
Iceland	-0.17	.0012	0.35	Tunisia	0.15	.0008	0.12
Italy	0.19	.0003	0.10	Turkey	0.14	.0008	0.04
Japan	0.08	.0009	0.10	USA	0.07	.0007	0.14

All seven methods compared in this study were employed to estimate the total heterogeneity variance in a random-effects model, as well as the slope, the residual heterogeneity variance, and the proportion of variance accounted for by the moderator in a mixed-effects meta-regression model with one predictor. Results are presented in Table 4.2.

Table 4.2 Estimates in random- and mixed-effects models using data from Else-Quest and colleagues (2010)

Method	$\hat{\tau}^2$	$\hat{\beta}_1$	$\hat{\tau}^2_{res}$	R^2
HE	.0077	-.3870	.0061	.2120
HS	.0052	-.3849	.0046	.1207
DL	.0058	-.3861	.0054	.0691
SJ	.0076	-.3870	.0061	.1891
ML	.0069	-.3858	.0051	.2544
REML	.0073	-.3867	.0058	.2060
EB	.0075	-.3868	.0059	.2093

As the slope estimates show, a negative relationship was found with all methods, indicating that a higher percentage of women in the parliament was associated with smaller advantages for boys in the mathematics test. Regarding the total heterogeneity variance, the lowest estimates were obtained using HS and DL methods (.0052 and .0058, respectively), while the highest estimates were provided by HE, SJ, and EB methods (.0077, .0076, and .0075, respectively). Residual heterogeneity variance estimates also showed some variability, with values ranging between .0046 (HS estimator) and .0061, obtained with the HE and SJ estimators. These differences led to notable variation among the estimates of the model predictive power depending on the estimator used. The R^2 values showed fluctuations from a 6.9% of heterogeneity accounted for by the moderator (DL estimator) to the 25.4% obtained with the ML estimator.

4.3 Simulation study

A simulation study was programmed in R using the *metafor* (Viechtbauer, 2010) package. Meta-analyses of k studies were generated, each of them based on a two-group design, comparing subjects in an experimental (E) and a control (C) group with respect to some continuous dependent variable. The scores of the n_i^E and n_i^C subjects in the respective groups were assumed to be normally distributed, using the standardized mean difference, defined in Equation (2.4), as the effect size index.

For each meta-analysis, $\boldsymbol{\theta}$ and $\mathbf{x}$ were defined as $(k \times 1)$ vectors containing parameter effects and moderator values, respectively. The predictor $\mathbf{x}$ was generated from a standard normal distribution. On the other hand, the $\boldsymbol{\theta}$ values were obtained from the expression $\boldsymbol{\theta} = \beta_0 + \beta_1 \mathbf{x} + \mathbf{u}$, where β_0 was set to 0.5 to reflect an effect of medium size according to the guidelines provided by Cohen (1988) for Social Sciences, the slope β_1 was set as described below, and $\mathbf{u}$ is an error term with distribution $N(0, \tau^2_{res})$. Note that, if the predictor is dropped from the model, then the error term $\mathbf{u}$ will have distribution $N(0, \tau^2)$.

The total heterogeneity variance, τ^2, and the model predictive power, P^2, were manipulated in the simulations. The former was set to values representative of no, low, medium, or large amounts of heterogeneity in Psychology and related fields (0, .08, .16, and .32, respectively). Regarding P^2, values of 0%, 25%, 50%, or 75% of heterogeneity accounted for the predictor were chosen, with the aim to reflect realistic conditions (Thompson & Higgins, 2002). After setting both parameter values, a value was then assigned to β_1 by means of the expression $\beta_1^2 = \tau^2 \mathrm{P}^2$. Table 4.3 reports the different values considered for these parameters, as well as the resulting values for β_1^2 and the

residual heterogeneity variance parameter, τ^2_{res}, which were computed with $\tau^2_{res} = \tau^2 - \beta_1^2$.[5]

Table 4.3 Parameter values considered in this simulation for τ^2, P^2, β_1, and τ^2_{res}

τ^2	0	.08				.16				.32			
P^2	0	0	.25	.50	.75	0	.25	.50	.75	0	.25	.50	.75
β_1^2	0	0	.02	.04	.06	0	.04	.08	.12	0	.08	.16	.24
τ^2_{res}	0	.08	.06	.04	.02	.16	.12	.08	.04	.32	.24	.16	.08

Other factors manipulated in this simulation were the number of studies in each meta-analysis (k = 5, 10, 20, 40, and 80) and the average sample size of the k studies ($\overline{N}$ = 30, 50, and 100). Note that, for the *i*th study, $N_i = n_{iE} + n_{iC}$, assuming equal group sizes. Vectors of individual sample sizes were generated with an asymmetry of +1.546, as reported by Sánchez-Meca and Marín-Martínez (1998a, p. 317) from a review of meta-analytic syntheses in Psychology. A total of $13 \times 5 \times 3 = 195$ conditions were examined. For each condition, 10,000 meta-analyses were simulated, and $\hat{\tau}^2$, $\hat{\tau}^2_{res}$, and R^2 were computed for each simulated database with the seven alternatives presented in Section 3.2 of this dissertation: HE, HS, DL, SJ, ML, REML, and EB methods.

Performance for all estimators of τ^2, τ^2_{res} and P^2 was compared using several criteria. Let $\hat{\theta}_i^j$ be an estimate obtained with any of the proposed methods in a particular condition. The bias for that condition was computed with (Marín-Martínez & Sánchez-Meca, 2010)

[5] From $\theta_i = \beta_1 X_i + u_i$, the total amount of heterogeneity in the true effect sizes, τ^2, can easily be computed with $\tau^2 = \beta_1^2 V(X_i) + \tau^2_{res} = \beta_1^2 + \tau^2_{res}$, as X_i and u_i are independent and normally distributed with mean zero and variances 1 and τ^2_{res}, respectively. This leads to the expression $\tau^2_{res} = \tau^2 \quad \beta_1^2$.

$$BIAS\left(\hat{\theta}^{j}\right)=\frac{\sum_{i}\hat{\theta}_{i}^{j}}{10{,}000}-\theta, \tag{4.1}$$

where θ is the value of the parameter of interest (see Table 4.3). The percentage of bias was then obtained with

$$\%BIAS\left(\hat{\theta}^{j}\right)=\frac{BIAS\left(\hat{\theta}^{j}\right)}{\theta}x100. \tag{4.2}$$

Moreover, the MSE was calculated with

$$MSE\left(\hat{\theta}^{j}\right)=\frac{\sum_{i}\left(\hat{\theta}_{i}^{j}-\theta\right)^{2}}{10{,}000}. \tag{4.3}$$

Finally, as described in Section 3.4 of this dissertation, the computation of the R^2 value may require truncation in various cases. When τ^2 and τ^2_{res} are both actually positive (in which case $0<\mathrm{P}^2<1$), a large rate of truncated R^2 values would reflect undesirable performance of Equation (3.24). Therefore, the proportion of R^2 values truncated to 0 or 1 was also examined for the different estimators along the simulated scenarios.

4.4 Results

4.4.1. Total heterogeneity variance

Because any negative estimates of τ^2 were truncated to zero, all estimators showed the expected positive bias under the homogeneous scenario ($\tau^2=0$). On the other hand, for the conditions with $\tau^2>0$, Table 4.4 shows the percentage of bias for the total heterogeneity variance estimates provided by each method when setting the number of studies and the average within-study sample size to values that can often be found in meta-analytic reviews (e.g., $k=20$ and $\overline{N}=50$).

Table 4.4 Percentage of bias for the total heterogeneity variance estimators with $k = 20$ and $\overline{N} = 50$

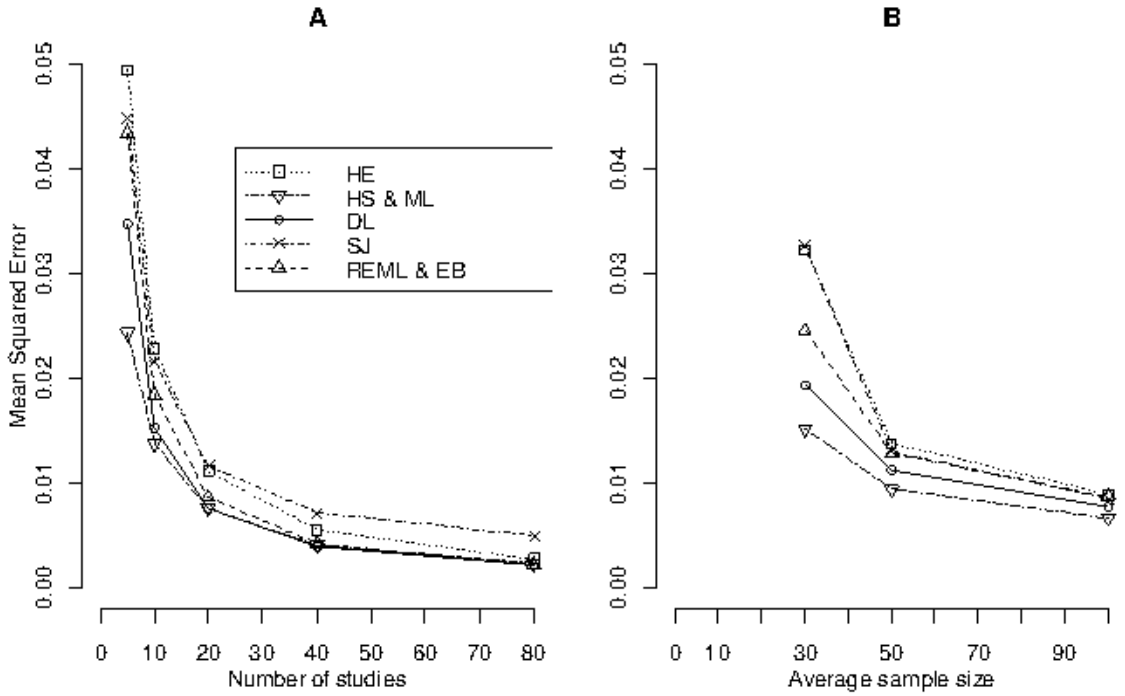
τ^2	HE	HS	DL	SJ	ML	REML	EB
0.08	2.81	-17.33	-6.23	45.19	-17.92	-6.96	-1.82
0.16	1.05	-16.03	-7.49	16.96	-14.27	-6.01	-2.36
0.32	.47	-16.83	-9.83	4.69	-11.93	-5.25	-2.36

The HS and ML estimators provided the most negatively biased estimates, with a deviation of around 16% from the parameter value. The SJ estimator showed the most (positively) biased results, although its performance improved as τ^2 increased. The DL and REML estimators performed similarly for small to medium amounts of heterogeneity, with a negative bias slightly over 5%, while the DL estimator yielded more biased results for large values of τ^2. The HE estimator showed the best results in terms of bias, with a positive deviation smaller than 3% and better results as the parameter value increased. Finally, the EB estimator performed reasonably well in terms of bias, with a negative deviation from the parameter value around 2%. Results with smaller values of k showed larger biases for all of the estimators under comparison. Conversely, the estimates obtained with 40 and 80 studies were more accurate than for $k = 20$ for the different methods. Finally, higher average sample sizes also led to more accurate results for all estimators.

Figure 4.1 presents the MSE values for the different estimators of τ^2 as a function of the number of studies and the average sample size. When comparing the estimators in terms of their relative efficiency, the SJ and HE methods provided the largest MSE values, while the HS and ML estimators showed the most efficient performance. The remaining estimators (DL, REML, and EB) performed similarly as k increased. All methods yielded more accurate estimates with a larger number of studies, as shown in Figure 4.1A, with

MSE values clearly decreasing with 20 or more studies. Moreover, an increase on the average sample size per study also led to better results, as it can be seen in Figure 4.1B.

Figure 4.1 Mean Squared Errors for the total heterogeneity variance estimators



4.4.2 Residual heterogeneity variance

Table 4.5 shows the percentage of bias for the various residual heterogeneity variance estimators, using again conditions with some heterogeneity among the true effects once a predictor is included in the regression model (τ^2_{res} = 0.02, 0.04,..., and 0.32), and setting the remaining factors to values that can be regarded as representative for meta-analyses in Psychology and related fields (e.g., $k = 20$ and $\overline{N} = 50$).

Table 4.5 Percentage of bias for the residual heterogeneity variance estimators with $k = 20$ and $\overline{N} = 50$

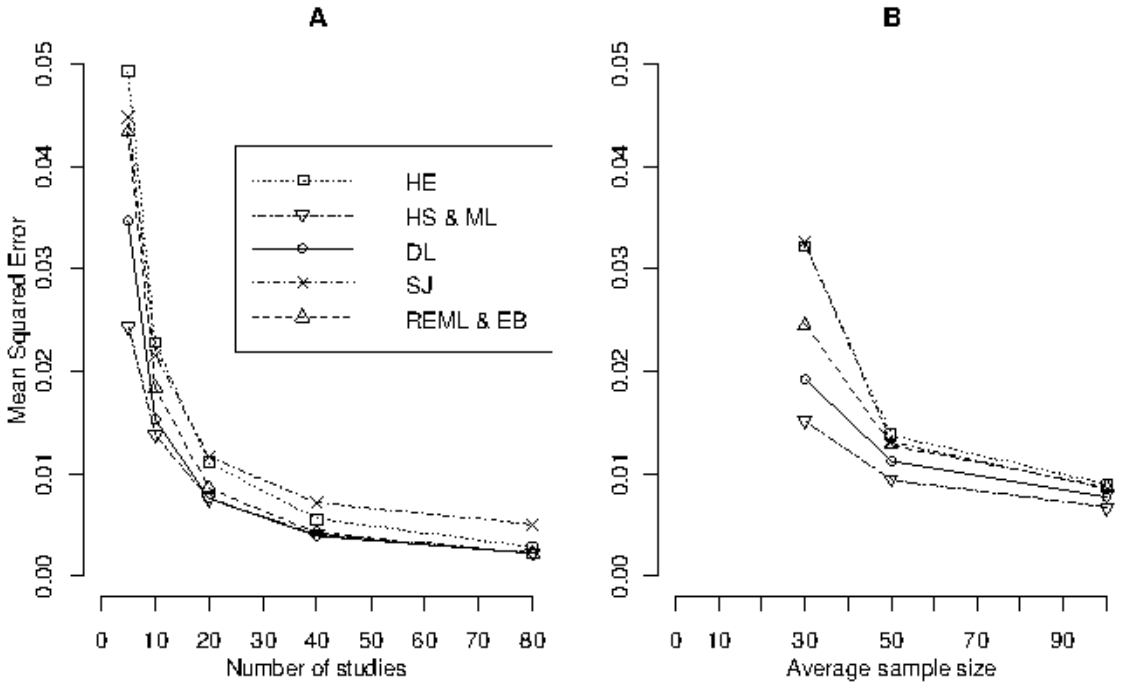
τ^2_{res}	HE	HS	DL	SJ	ML	REML	EB
0.02	42.40	-20.59	21.39	287.04	-25.75	16.35	28.85
0.04	14.12	-27.93	1.55	140.93	-30.59	-1.06	6.68
0.06	4.41	-29.75	-5.40	70.96	-31.28	-6.86	-.97
0.08	3.22	-27.45	-5.70	63.67	-28.04	-6.35	-1.33
0.12	2.26	-24.87	-6.23	32.83	-24.06	-5.58	-1.47
0.16	-.12	-24.71	-8.02	23.03	-23.08	-6.72	-3.29
0.24	.84	-22.97	-8.17	13.66	-19.54	-5.09	-2.07
0.32	.11	-23.46	-9.74	4.44	-18.62	-5.29	-2.55

Trends for the different methods when estimating the residual heterogeneity variance were very similar to the ones detailed for τ^2. Regarding bias, the SJ estimator showed again the most biased results – the positive bias was now larger than for τ^2 – unless the parameter value was large enough ($\tau^2_{res} = 0.24$ and $\tau^2_{res} = 0.32$). Moreover, HS and ML methods provided again negatively biased estimates, with a deviation from the parameter value around 25% with 20 studies, larger than the one observed for τ^2. Finally, HE, DL, REML, and EB estimators performed similarly than for τ^2.

Figure 4.2 shows the MSE results for the estimators as a function of the number of studies and the average sample size of the studies. The HS and ML methods performed very similarly, so their results are presented jointly, same as for the REML and EB estimators. As found in the results for τ^2, the number of studies showed the largest influence on the efficiency of all estimators of τ^2_{res} and the MSE values especially

decreased when going from 5 to 10 and from 10 to 20 studies. The average sample size also showed some influence on the efficiency of the estimates, with smaller MSE values obtained as $\overline{N}$ increased. The SJ and HE estimators showed the largest MSE values, while the HS and ML methods provided the most efficient estimates. All estimators except the SJ method performed similarly with $k = 80$.

Figure 4.2 Mean Squared Errors for the residual heterogeneity variance estimators



4.4.3 Model predictive power

The R^2 values obtained with all estimators were quite variable, but the estimates tended to fall closer to the parameter value as k, $\overline{N}$, τ^2, and P^2 increased. As an illustration, Table 4.6 presents the correlations between the estimates obtained with the different methods under two opposite scenarios. The lower part of this table (below the main diagonal) presents the correlations under adverse conditions ($k = 5$, $\overline{N} = 50$,

$\tau^2 = 0.16$, and $P^2 = 0.25$), while the upper part provides the correlations obtained under an optimal scenario ($k = 80$, $\overline{N} = 100$, $\tau^2 = 0.32$, and $P^2 = 0.50$).

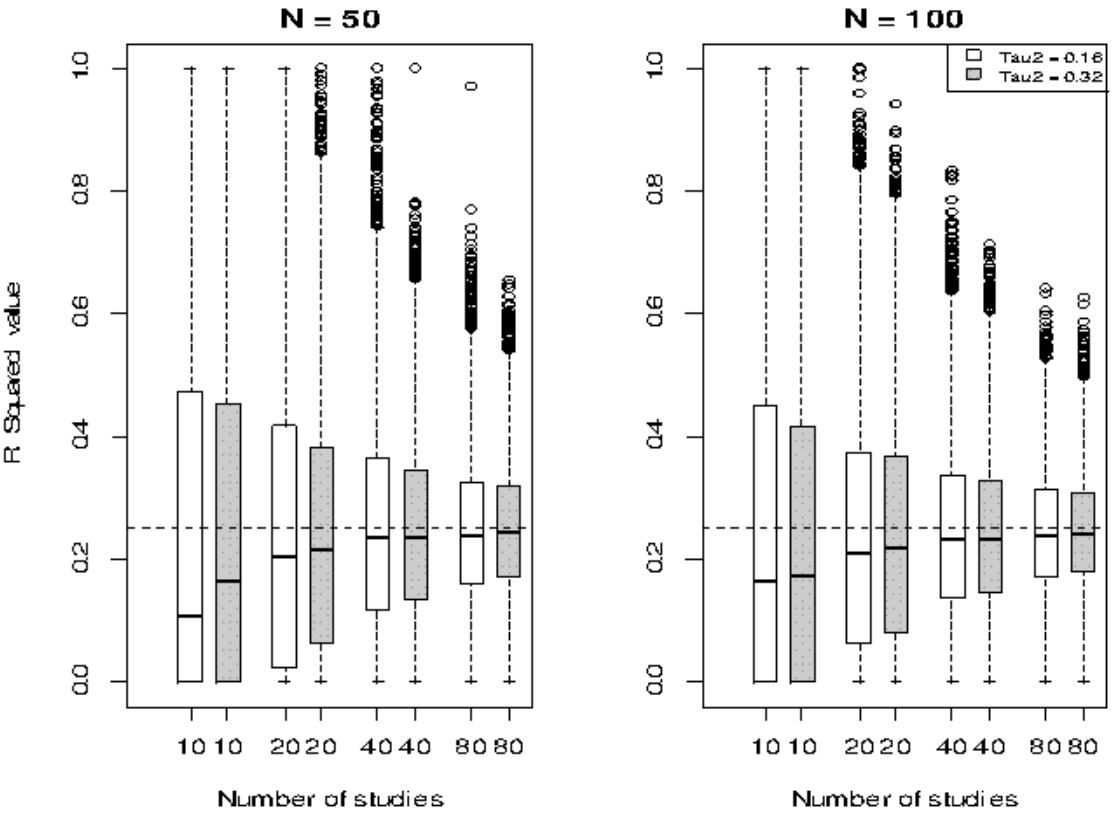
Table 4.6 Correlations between the R^2 values obtained with the different methods

	HE	HS	DL	SJ	ML	REML	EB
HE		.9727	.9731	.9934	.9958	.9960	.9991
HS	.7070		.9999	.9692	.9869	.9865	.9803
DL	.9368	.7201		.9701	.9871	.9868	.9807
SJ	.8227	.5720	.8196		.9907	.9915	.9935
ML	.7627	.8395	.7677	.5943		.9999	.9988
REML	.9322	.6796	.9516	.8221	.7591		.9989
EB	.9678	.6991	.9772	.8314	.7725	.9626	

Under adverse conditions, the highest correlations were found between the DL, REML, and EB estimators, with values over .95, while most of the remaining combinations yielded values below .90 and even below .60 (e.g., the correlation between the HS and SJ estimators). Conversely, all estimators performed very similarly under the optimal scenario, with all correlations falling above .96. Table 4.6 shows, therefore, that the differences between estimators are especially important under the most adverse conditions, while the performance for all methods tends to convergence for the optimal scenarios.

Out of the different factors manipulated in this simulation, the accuracy of the P^2 estimates was mostly influenced by the number of studies. The influence exerted by k, as well as by other factors, on the accuracy of the P^2 estimates is illustrated in Figure 4.3 using the EB estimator, which provided slightly more accurate results than the other methods under comparison, and considering scenarios where the model predictive power to be estimated was $P^2 = 0.25$.

Figure 4.3 R^2 values using the EB estimator with $P^2 = 0.25$



The boxplots presented in Figure 4.3 show a very large variability in the P^2 estimates, especially for small values of k. The picture is worrisome for a typical meta-regression, as it reveals that any R^2 value (including a truncated estimate) can be obtained unless the number of studies is large enough (40 or more studies). Results with 5 studies, which are not shown in this figure, were very unstable. Apart from the notable influence of k, an increase in the average sample size per study led to more precise estimates, while increasing the heterogeneity variance parameter led to a smaller rate of truncations.

Several descriptives were computed for the R^2 values obtained with the different estimators, considering conditions with 40 studies and setting the other factors to realistic values for a meta-regression with a single covariate (e.g., $\overline{N} = 50$, $\tau^2 = 0.16$, and

$P^2 = 0.25$). Table 4.7 presents the mean, the median, the 2.5 and 97.5 percentiles, and the rates of values truncated to zero and to one for each estimation method.

Table 4.7 R^2 values with $k = 40$, $\overline{N} = 50$, $\tau^2 = 0.16$ and $P^2 = 0.25$

Estimator	HE	HS	DL	SJ	ML	REML	EB
Mean	.2534	.2950	.2488	.1464	.3052	.2588	.2555
$P_{2.5}$	0	.0157	0	0	.0166	0	0
Median	.2311	.2752	.2281	.1332	.2843	.2390	.2351
$P_{97.5}$	.6512	.6974	.6458	.3734	.7379	.6781	.6547
$P\left(R^2 = 0\right)$	.0585	.0003	.0689	.0570	.0062	.0630	.0565
$P\left(R^2 = 1\right)$	.0021	.0029	.0017	0	.0011	.0010	.0015

Regarding the comparison of the different estimators in terms of bias, the HE, DL, REML, and EB estimators performed appropriately, with their mean estimates deviating less than .01 from the parameter value ($P^2 = 0.25$). In contrast, the HS and ML estimators showed a positive bias, while the mean estimate for the SJ estimator showed a large negative bias. Despite the negative bias for HS and ML estimators and the positive bias for the SJ estimator when estimating τ^2 (see Section 4.4.1), those deviations were larger when estimating τ^2_{res}, as it was also detailed before. Consequently, the trends for these three methods were reversed for the estimation of P^2.

In addition to the bias that was found for the HS, ML, and SJ estimators, the remaining methods showed some problems as well. When examining the percentiles presented in Table 4.7, it can be seen that there was a wide variation among the individual estimates, and that 95% of the central values ranged from 0 to 0.74. Moreover, a non-negligible proportion of the estimates (over 5%) were truncated to zero, especially for the DL and REML estimators. While the truncation rates to zero were clearly lower for the HS and ML estimators, the bias showed by these two methods advises against their use.

Finally, despite the parameter value of $P^2 = 0.25$, the HE, DL, REML, and EB methods still provided some estimates that were truncated to one. On the other hand, since the SJ estimator always yields a non-negative value for τ^2 or τ^2_{res}, R^2 can never reach 1 when using this estimator[6] and hence never required truncation at the upper end of the scale, although in turn it provided the largest bias.

Table 4.8 presents the MSE results with 40 studies and $P^2 = 0.25$ for the different estimators. Only conditions with some heterogeneity among the parameter effects $(\tau^2 > 0)$ were considered here.

Table 4.8 MSE values for the P^2 estimators with $k = 40$ and $P^2 = 0.25$

Estimator		HE	HS	DL	SJ	ML	REML	EB
$\overline{N} = 30$	$\tau^2 = 0.08$	.0959	.1271	.1036	.0390	.1498	.1195	.1078
	$\tau^2 = 0.16$	.0711	.0815	.0682	.0311	.0972	.0777	.0671
	$\tau^2 = 0.32$	.0359	.0402	.0375	.0229	.0478	.0409	.0346
$\overline{N} = 50$	$\tau^2 = 0.08$	.0637	.0770	.0636	.0288	.0871	.0695	.0642
	$\tau^2 = 0.16$	.0318	.0346	.0322	.0218	.0379	.0340	.0313
	$\tau^2 = 0.32$	.0223	.0225	.0232	.0177	.0229	.0228	.0223
$\overline{N} = 100$	$\tau^2 = 0.08$	.0287	.0299	.0284	.0205	.0308	.0289	.0285
	$\tau^2 = 0.16$	.0203	.0198	.0203	.0165	.0203	.0205	.0203
	$\tau^2 = 0.32$	.0174	.0163	.0174	.0151	.0170	.0175	.0175

[6] The SJ estimator will only provide a value of zero (for both $\hat{\tau}^2$ and $\hat{\tau}^2_{res}$) in the unlikely event that the effect sizes are homogeneous, as it can be readily seen from Equations (3.9) and (3.10).

All methods performed more efficiently as the average sample size and the total heterogeneity variance increased. When comparing the different methods, the ML and HS estimators provided the largest MSE values, while the SJ estimator was the most efficient method, especially under the most adverse conditions. Regarding the influence of the number of studies, weak performance was reported before for the method proposed by Raudenbush (1994) with a small number of studies (see Figure 4.3 and Table 4.7). With $k = 20$, trends were already similar to the ones shown in Table 4.8, although the MSE values were twice as large as for $k = 40$. With 80 studies, the MSEs were on average smaller than .04 under all of the conditions examined here, although trends for the different estimators remained the same.

4.5 Discussion

In this study, the performance of seven methods for the estimation of the total and residual heterogeneity variances, as well as the model predictive power, was assessed under a variety of realistic scenarios in applied research. The estimators here compared performed differently, especially under adverse and intermediate conditions, while all methods provided similar and accurate estimates of the parameters of interest for the most favorable conditions (e.g., large number of studies and large number of participants per study).

Regarding the results for the total heterogeneity variance, the patterns found in this simulation are comparable to the ones reported by Viechtbauer (2005). The DL, REML, and EB estimators performed reasonably well in terms of bias and efficiency, although the DL method yielded negatively biased estimates for large parameter values, as found in previous simulations (Malzahn et al., 2000; Sidik & Jonkman, 2005b, 2007; Viechtbauer, 2005). The HE estimator showed essentially unbiased results (the positive bias observed in Table 4.4 and Table 4.5 can be regarded as a consequence of truncating the negative

estimates to zero) but large MSE values, while the HS and ML methods performed very efficiently but with a negative bias. Finally, the SJ method showed a large positive bias for small parameter values, as it has been previously described (Sidik & Jonkman, 2005b), and the largest MSE values. The performance of the various estimators remained very similar after the inclusion of a moderator.

Regarding the estimation of the predictive power in meta-regression models with one predictor, no estimator performed accurately with less than 40 studies. Again, the HS, ML, and SJ estimators yielded the most biased estimates. The remaining estimators performed more precisely, although their estimates still showed wide variation even with a moderate to large number of studies, including truncated values to zero and one, as shown in Table 4.7. Given the large MSE of the SJ estimator for τ^2 and τ^2_{res}, the SJ estimator showed surprisingly efficient performance for estimating P^2, while the HS and ML methods now provided the largest MSE values.

Out of the different factors manipulated in this simulation, results from this simulation suggest that the number of studies exerts an important influence on the accuracy of the results, and that precise estimates of the heterogeneity variances and the model predictive power can only be expected with at least 20 and 40 studies, respectively. An increase in the average sample size also improved the results for all estimators. The critical influence of k on the accuracy of the heterogeneity variance estimators has already been discussed by several authors both in the context of random-effects models (e.g., Borenstein et al., 2009; Schulze, 2004) and mixed-effects models (Thompson & Higgins, 2002). The fact that results were more accurate as k and $\overline{N}$ increased is in agreement with large-sample theory, which underlies the statistical models and methods in meta-analysis (Hedges, 2009). Moreover, as shown in Figure 4.3 and Table 4.8, the estimators of the model predictive power performed more efficiently as the total heterogeneity variance increased. An explanation of this fact is that, when estimating τ^2,

a small parameter value will lead more often to negative estimates requiring truncation, and this will also lead to truncated R^2 values.

In summary, the results obtained in this simulation study suggest that about 40 studies are required to get accurate estimates of P^2 in mixed-effects meta-regression models, so that a cautious interpretation of R^2 values should be advised for meta-regression models fitted with a smaller number of studies (Thompson, 1994). Out of the different estimators here compared, the REML, DL, and EB methods showed the most accurate results across the different scenarios and criteria here considered. Although the present study focused on standardized mean differences, it is likely that these findings can be generalized to meta-analyses with other effect size measures that are (at least approximately) normally distributed. However, conclusions from this simulation are restricted to the scenarios considered here, so that further simulation studies are needed in order to account for conditions different to the ones included in the present study.

Chapter 5

A comparison of procedures to test for moderators in mixed-effects meta-regression models

5.1 Objectives, previous simulation studies, and hypotheses

5.1.1 Objectives of the study

As shown in Chapter 3 of this dissertation, different methods have been proposed in the meta-analytic literature both for estimating the amount of residual heterogeneity variance and for testing the coefficients in mixed-effects meta-regression models, and the choice of the statistical method can affect the results and statistical conclusions. In this study, several methods for mixed-effects meta-regression models were compared through

Monte Carlo simulation under some realistic scenarios in psychological research. Various methods were implemented for:

- The estimation of the residual heterogeneity variance: seven estimators, already presented in Section 3.2 of this dissertation, were employed: Hedges (HE), Hunter-Schmidt (HS), DerSimonian-Laird (DL), Sidik-Jonkman (SJ), maximum likelihood (ML), restricted maximum likelihood (REML), and empirical Bayes (EB) estimators.

- The statistical testing of the regression model coefficients: the standard method, Knapp-Hartung method, Huber-White method, likelihood ratio test, and permutation test were included. All of these methods were described in Section 3.3 of this dissertation. The Knapp-Hartung method was implemented both with and without the truncation proposed by the authors (Knapp & Hartung, 2003), leading to six different methods.

Methods from both categories were combined to generate different methodological alternatives available to meta-analysts when testing the statistical significance of one or more moderators in mixed-effects meta-regression models. There are some restrictions, however. Firstly, the likelihood ratio test was only implemented together with the ML estimator, since it is not theoretically appropriate to combine it with the remaining estimators. And secondly, the computation of the permutation test is not efficient when applying some iterative estimator, due to the fact that only one missing value along the whole set of permutations (e.g., one model for which convergence is not achieved in the estimation of the heterogeneity variance) ruins the entire process. For this reason, the permutation test was only combined with HE, HS, DL, and SJ estimators. In total, 33 combinations of procedures to test the statistical significance of the slope in a meta-regression model with a moderator, were compared for the present simulation in terms of empirical Type I error and statistical power rates, using standardized mean differences as the effect size index.

5.1.2 Previous simulation studies

Some previous simulation studies compared the performance of different methods to test for moderators in mixed-effects meta-regression models. Knapp and Hartung (2003) combined their method and the standard method with DL, (approximate) REML, and EB estimators. These authors found, using log risk ratios as the effect size index, better results for the Knapp-Hartung method compared to the standard one in terms of adjustment to the nominal significance level, irrespective of the residual heterogeneity variance estimator. Later, Sidik and Jonkman (2005a) compared the standard, Knapp-Hartung, and Huber-White tests using the DL estimator, also with log risk ratios as the effect size measure. Their results again showed a better performance for the Knapp-Hartung method in terms of adjustment to the nominal significance level. Nevertheless, no study has analyzed yet the consequences of truncating the Knapp-Hartung method, as recommended by the authors (Knapp & Hartung, 2003).

More recently, Huizenga and colleagues (2011) compared different methods in terms of empirical Type I error and statistical power rates, using standardized mean differences as the effect size index, as it was done for the present study. These authors included the likelihood ratio test, and found a slightly better control of the Type I error rate for this method compared to the standard one. They also examined a resampling method based on permutations of the residuals, and found promising results for that procedure. However, that resampling test is somewhat different to the permutation test considered in this study, and the performance for the latter has not been systematically evaluated yet.

5.1.3 Hypotheses of this study

According to the hypotheses of this study, it was expected that:

1) Differences among the methodological alternatives would not be due to the residual heterogeneity variance estimator, but rather to the method for testing the regression model coefficients, as found by Knapp and Hartung (2003).

2) The standard, Huber-White, and likelihood ratio tests for the coefficients would not control adequately the Type I error rate, as suggested by several authors (e.g., Huizenga et al., 2011; Knapp & Hartung, 2003; Sidik & Jonkman, 2005a; Thompson & Higgins, 2002).

3) The Knapp-Hartung method would provide an adequate control of the Type I error rate, as found in previous studies using different effect size measures (Knapp & Hartung, 2003; Sidik & Jonkman, 2005a).

4) The truncation proposed by Knapp and Hartung (2003) would lead to a loss of statistical power compared with the original Knapp-Hartung method.

5) The permutation test would perform appropriately in terms of empirical Type I error rates for every simulated condition.

5.2 An illustrative example

The set of methodological alternatives to test the influence of moderators in mixed-effects meta-regression models were applied to an example for illustrative purposes. Table 5.1 shows the results of 12 studies about the effect of psychological therapy on depressive symptoms for patients with obsessive-compulsive disorder, with the effect size index being the standardized mean difference, d_i, and $\hat{\sigma}^2_{d_i}$ the within-study variance for each d_i value. Data were taken from the meta-analysis conducted by Rosa-Alcázar and colleagues (2008). Most of the d_i indices were positive, indicating a higher

benefit for the treatment group compared to the control group at the posttest. Nevertheless, large variability was found from one study to another in the magnitude of the effect size estimate. In order to account for part of the heterogeneity among the effect size estimates, mixed-effects meta-regression models were fitted using the percentage of males in the clinical sample of each study as the covariate, x_i, for the analyses.

Table 5.1 Example data from the meta-analysis of Rosa-Alcázar et al. (2008)

Study	d_i	$\hat{\sigma}^2_{d_i}$	x_i
Fals-Stewart, Marks, and Schafer (1993)	.951	.0731	45.3
Fineberg, Hughes, Gale, and Roberts (2005)	.756	.1428	24.3
Freeston et al. (1997)	-.057	.1834	55.0
Greist et al. (2002)	.275	.0336	58.0
Jones and Menzies (1998)	.804	.2063	9.5
Lindsay, Crino, and Andrews (1997)	.580	.2316	33.3
Lowell, Marks, Noshirvani, and O'Sullivan (1994)	-.105	.3338	41.7
Marks, Stern, Cobb, and McDonald (1980)	-.059	.2009	30.0
Nakatani et al. (2005)	-.275	.2271	33.3
O'Connor, Todorov, Robillard, Borgeat, and Brault (1999)	1.350	.2456	30.0
Van Balkom et al. (1998)	.101	.1153	37.1
Vogel, Stiles, and Götesman (2004)	3.140	.4857	16.0

The influence of the covariate was statistically tested using all possible combinations of residual heterogeneity variance estimators and statistical tests for the regression model coefficients (see Section 5.1.1 of this dissertation). P-values obtained in the corresponding analyses are shown in Table 5.2.

Table 5.2 Results obtained with the data example from the meta-analysis of Rosa-Alcázar et al. (2008)

Method	HE	HS	DL	SJ	ML	REML	EB
Standard	.115	.056	.072	.115	.057	.076	.096
Knapp-Hartung	.120	.156	.140	.120	.155	.137	.127
Truncated Knapp-Hartung	.146	.156	.140	.146	.155	.137	.127
Huber-White	.148	.141	.144	.148	.141	.145	.146
Likelihood ratio test	-	-	-	-	.063	-	-
Permutation test	.094	.092	.134	.080	-	-	-

Although none of the analyses found a statistically significant relationship, some discrepancies among the p-values can be observed. The likelihood ratio test provided marginally significant results, as well as some applications of the standard and permutation tests. On the other hand, p-values were always greater than .10 when implementing the Knapp-Hartung and Huber-White methods, with generally higher values obtained with the latter. The example results do not suggest a clear influence of the variance estimator on the statistical conclusion, but rather of the statistical test for the moderator.

5.3 Simulation Study

In order to compare the performance of these methods, a Monte Carlo simulation study was conducted. Meta-analyses with k studies were generated, with each study based on a two-group design, comparing subjects in an experimental (E) and a control (C) group with respect to some continuous dependent variable. The scores of the n_i^E and n_i^C subjects in the respective groups were assumed to be normally distributed, using standardized mean differences as the effect size measure (further details of this index can be found in Chapter 2 of this dissertation). For the simulation study, it was assumed that a single moderator influences the size of the true effect for the *i*th study, θ_i, such that

$$\theta_i = \beta_0 + \beta_1 x_i + u_i. \qquad (5.1)$$

For each iteration of the simulation, the values of the moderator, x_i, were randomly generated from a standard normal distribution, and the random errors u_i were also generated with distribution $N(0, \tau^2_{res})$. Three different values for τ^2_{res} were considered, namely 0, 0.08 and 0.32, corresponding to the absence, a medium amount, and a large amount of residual heterogeneity in the true effects. Without loss of generality, β_0 was set equal to zero. For the slope, β_1, three conditions were examined, namely $\beta_1 = 0$, $\beta_1 = 0.2$, and $\beta_1 = 0.5$, the first yielding information on the empirical Type I error rate of the various tests, and the latter providing information about the power of the tests when the null hypothesis is in fact false.[7]

[7] Note that for each combination of the three τ^2_{res} values and the three β_1 values, the model predictive power could easily be computed with the expression $P^2 = \beta_1^2 / (\beta_1^2 + \tau^2_{res})$ (see Section 4.3 of this dissertation). Specifically, for a slope parameter of $\beta_1 = 0.2$, values of τ^2_{res} equal to 0, .08, and .32 correspond to $P^2 = 1$, $P^2 = .33$, and $P^2 = .11$, respectively, if the model predictive power is computed with the formula proposed by Raudenbush (1994). Considering now the conditions with $\beta_1 = 0.5$, the

Five different values were considered for k, namely 5, 10, 20, 40, and 80, corresponding to a small to large number of studies for the meta-analysis. After simulating k θ_i values based on Equation (5.1), the corresponding observed effect size estimates were then generated with $g_i = Z_i / \sqrt{X_i / m_i}$, where $Z_i \sim N(\theta_i, 1/n_i^E + 1/n_i^C)$, $X_i \sim \chi^2_{m_i}$, and $m_i = n_i^E + n_i^C - 2$. Then, unbiased parameter estimates, d_i, were computed by correcting g_i with Equation (2.4), already presented in Chapter 2 of this manuscript. The corresponding sampling variances, $\hat{\sigma}^2_{d_i}$, were then computed using Equation (2.5), also presented in Chapter 2.

Sample sizes of the individual studies were also manipulated, assuming $n_i = n_i^E = n_i^C$ and setting n_i either equal to (6, 8, 9, 10, 42), (16, 18, 19, 20, 52), or (41, 43, 44, 45, 77), corresponding to average sample sizes of 30, 50, and 100 subjects for the studies (individual sample sizes were chosen based on a review of published meta-analyses where a skewness value of +1.546 was found to be realistic for sample size distributions; for more details, see Sánchez-Meca & Marín-Martínez, 1998a). For the $k = 10$, $k = 20$, $k = 40$, and $k = 80$ conditions, the sample size vectors were repeated 2, 4, 8, and 16 times, respectively.

Thus, a total of 5 (k) × 3 (n_i) × 3 (β_1) × 3 (τ^2_{res}) = 135 conditions were examined. For each of these conditions, 10,000 meta-analyses were simulated. After generating the data within a particular iteration of a particular condition, the meta-regression model was fitted using the various heterogeneity estimators and then the model coefficient $\hat{\beta}_1$ was tested for statistical significance with the various procedures described earlier, using

---

corresponding values for the model predictive power, setting τ^2_{res} again to 0, .08, and .32, will now be $P^2 = 1$, $P^2 = .76$, and $P^2 = .44$. This illustrates how the increase in τ^2_{res} will generally lead to a decrease in the power of the statistical tests.

$\alpha = .05$ as the nominal significance level. For $k = 5$, an exact permutation test was carried out. For larger values of k, obtaining the exact permutation-based p-values was not feasible, so that a total of 5,000 random permutations were used for the test. The rejection rates of the various procedures were recorded for each condition. The simulation was conducted with R, using the *metafor* package to fit the meta-regression model (Viechtbauer, 2010).

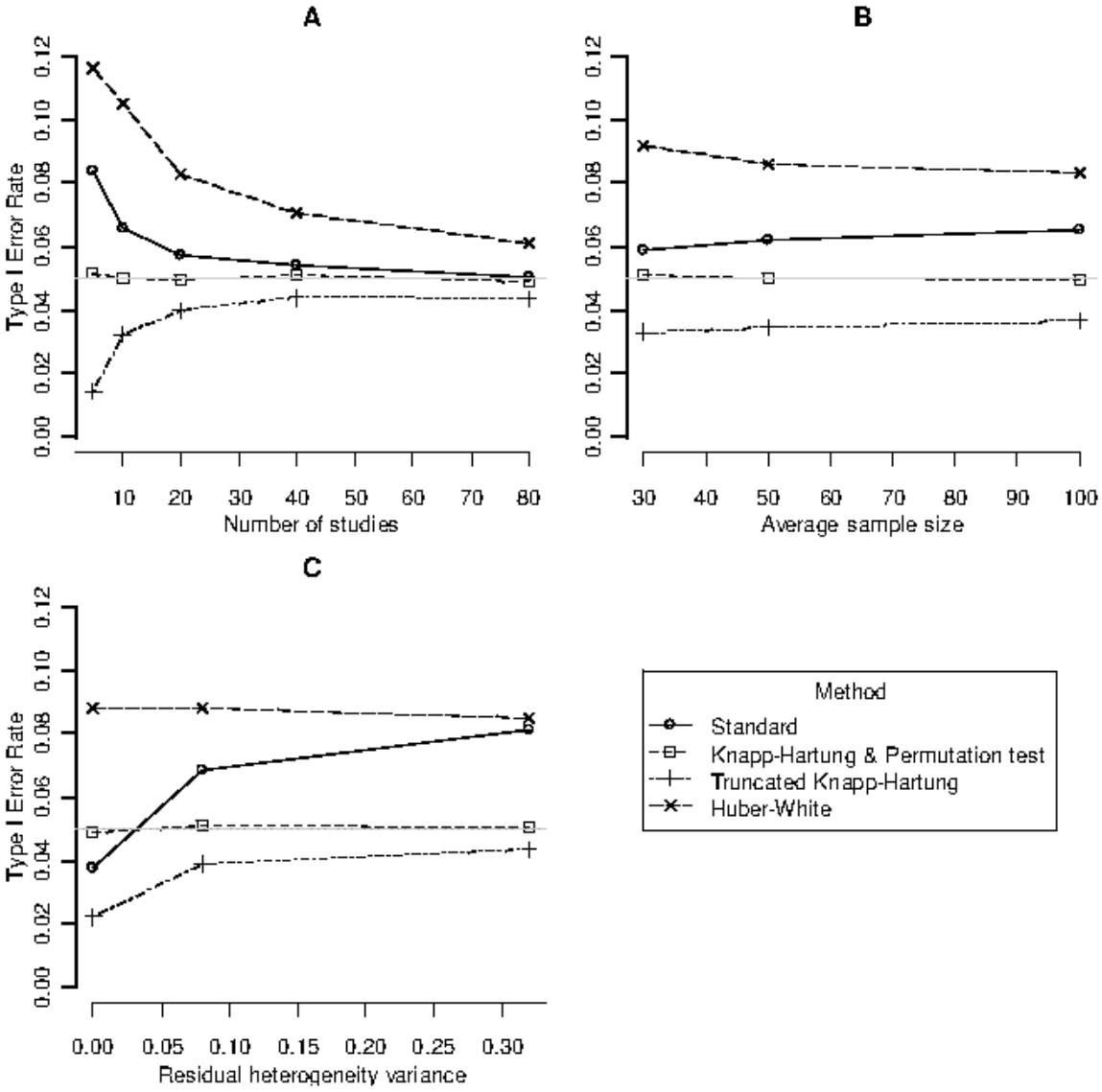
5.4 Results

In this section, the performance of the different methods is compared using 3×3 graph figures. Since no trend differences were found depending on the residual variance estimator used, only the results for the DL, ML and EB estimators are presented here. This section is divided into two parts, corresponding to the empirical Type I error rate and the statistical power of the tests, respectively.

5.4.1 Empirical Type I error rate

Setting $\beta_1 = 0$ allowed for comparing the methods in terms of their empirical Type I error rates. Note that by setting $\alpha = .05$, values around .05 indicate that the Type I error rate is adequately controlled. Figure 5.1 presents the empirical Type I error rates for the different methods using the DL estimator. Since values for the Knapp-Hartung method and the permutation test were essentially indistinguishable, results for both tests were averaged.

Figure 5.1 Empirical Type I Error Rates of the methods when using the DL estimator



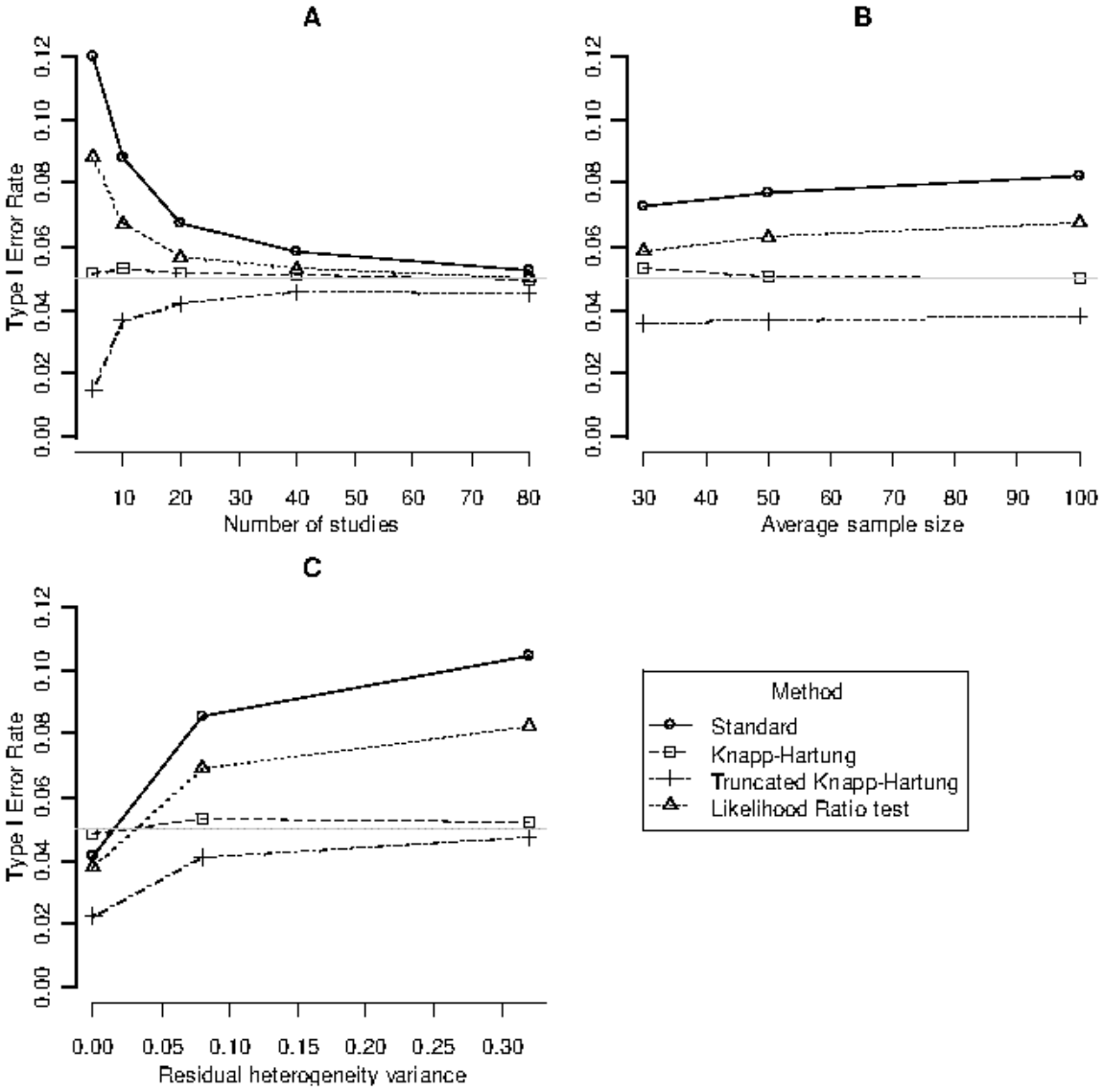
Results were very different depending on the method used to test the moderator. Both the Knapp-Hartung and permutation tests performed close to the nominal level regardless of the simulated scenario. In contrast, the truncated Knapp-Hartung method provided overly conservative results for most conditions, especially when the number of studies was small and when there was no residual heterogeneity among the true effects. On the other hand, the standard and Huber-White methods showed empirical rejection

rates clearly over the nominal significance level. The number of studies showed a similar influence on both methods, with a small number of studies corresponding to a higher proportion of incorrect rejections of the null hypothesis, especially for the Huber-White test. Finally, a larger amount of residual heterogeneity across the effect sizes also led to a greater amount of incorrect rejections for the standard method. For instance, with 40 studies, the standard method provided rejection rates around 0.04 with $\tau^2_{res} = 0$, while rates for this method were over 0.06 when $\tau^2_{res} = 0.32$.

Figure 5.2 presents the empirical Type I error rates for the different statistical tests when using the ML estimator. Rejection rates for the Huber-White method, which performed similarly than when combined with the DL estimator, were not included in this set of charts. Note that the performance of the permutation test is not analyzed here because, as stated before, this method is computationally overly demanding when combined with iterative estimators of τ^2_{res}.

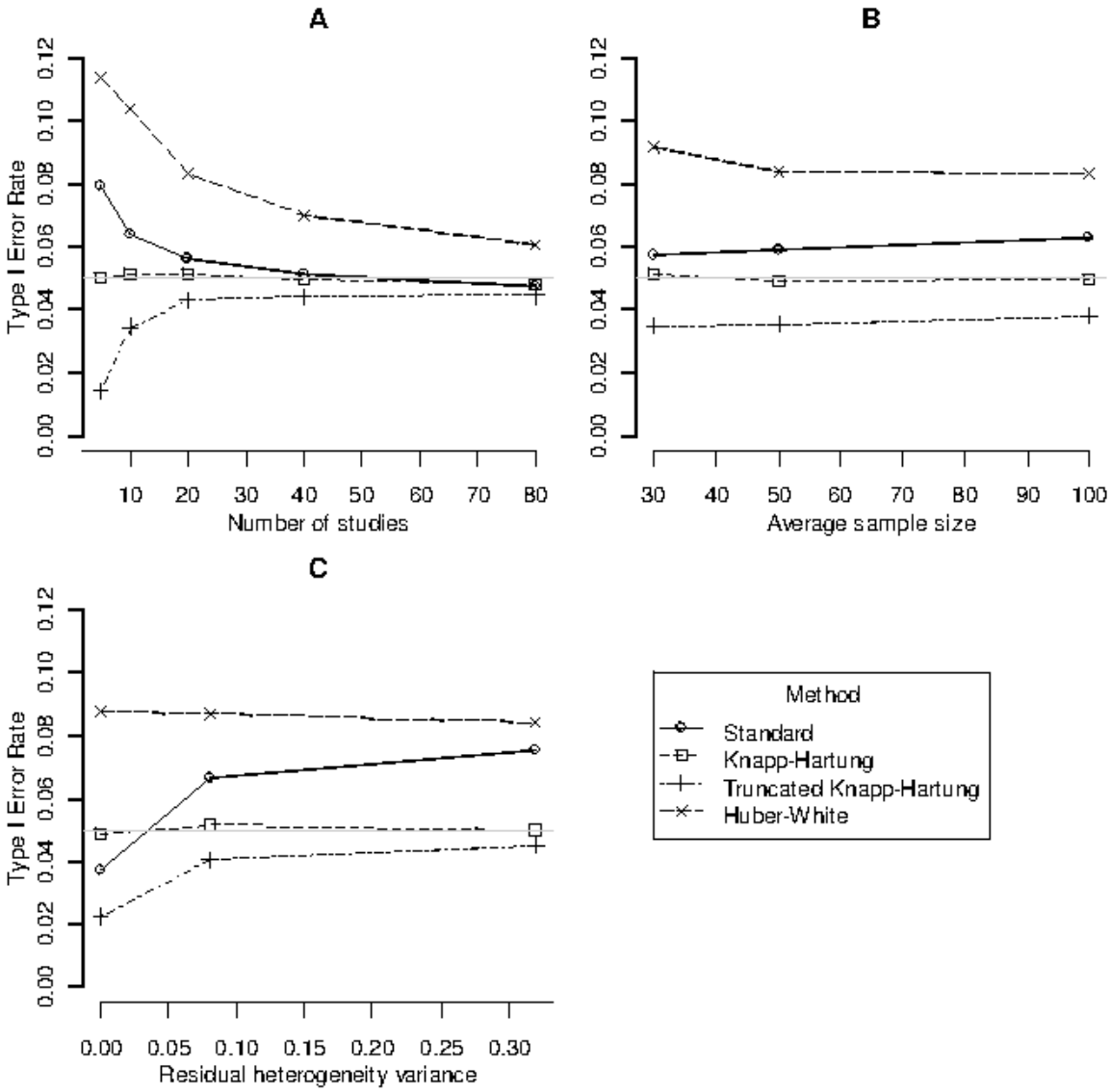
The general trends in the performance of the methods were very similar when using the DL and the ML estimator. The Knapp-Hartung method performed almost nominally irrespective of the simulated scenario. The standard method showed rejection rates clearly over 0.05, especially with a small number of studies and a large amount of residual heterogeneity among the true effects. Similar results were obtained with the likelihood ratio test, which showed rates slightly smaller than the ones observed for the standard method on average and adequate control of the empirical Type I error rate for meta-analyses with 40 or more studies. Finally, the rejection rate of the truncated Knapp-Hartung method fell below the nominal significance level, getting closer to .05 as the number of studies and the amount of residual heterogeneity increased.

Figure 5.2 Empirical Type I Error Rates of the methods when using the ML estimator



Similar trends to the ones described above were observed when using the EB estimator, whose results are shown in Figure 5.3. Again, because of the iterative computations required for the EB method, the permutation test was not combined with this estimator.

Figure 5.3 Empirical Type I Error Rates of the methods when using the EB estimator



5.4.2 Statistical power rate

Statistical power reflects the probability for a method to properly reject the null hypothesis of an absence of statistical association between the covariate and the effect

sizes when there is a true relationship (e.g., when $\beta_1 \neq 0$). Generally, power rates equal to or greater than 0.8 are considered as satisfactory in Psychology (Cohen, 1988). In order to assess the statistical power rates of the different procedures for testing the significance of regression coefficients, conditions with $\beta_1 = 0.2$ are presented here.

Figure 5.4 Statistical Power Rates of the methods when using the DL estimator

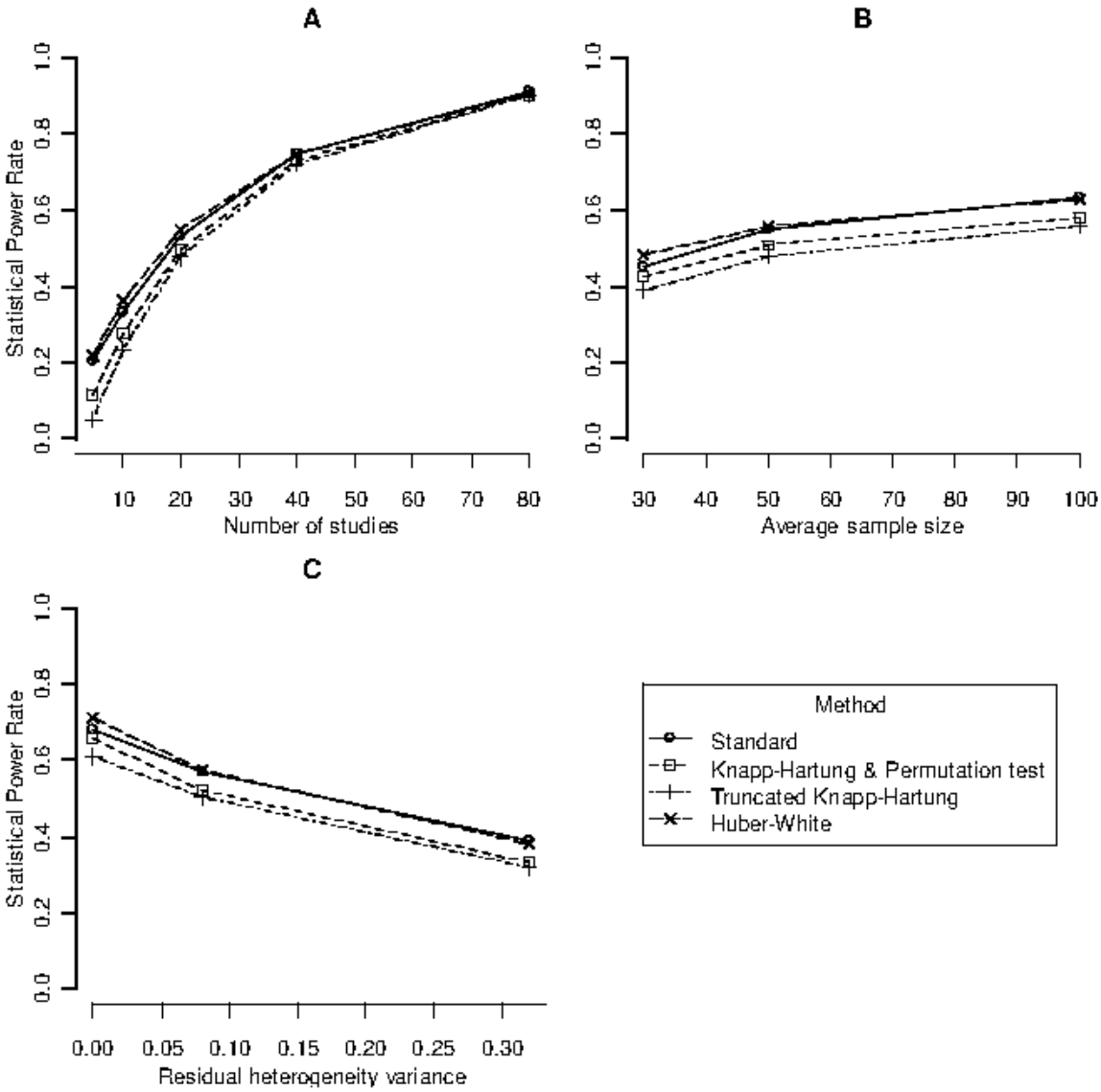


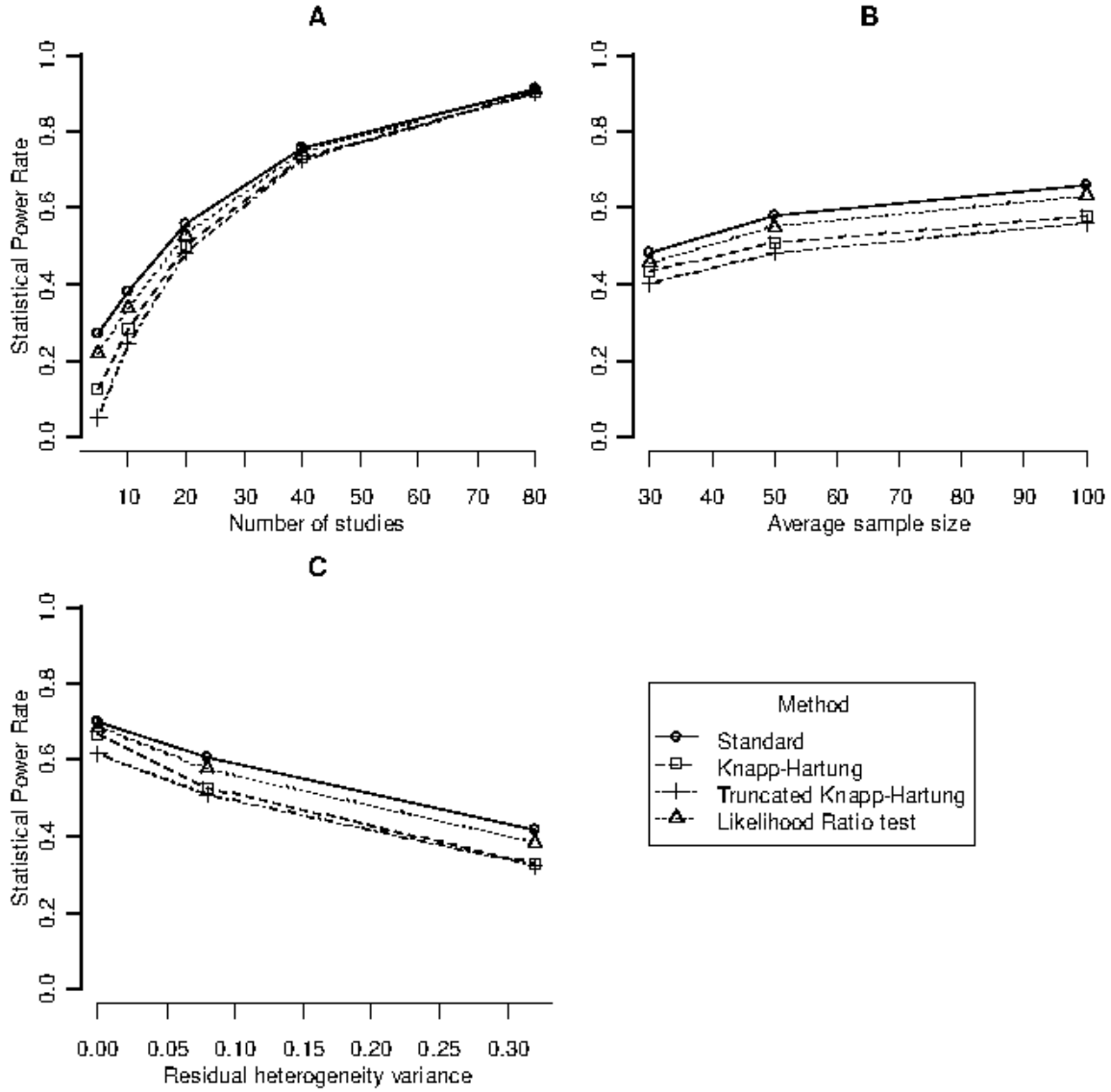
Figure 5.4 presents the power results of the methods when using the DL estimator. Again, Knapp-Hartung and permutation tests showed very similar results, so that values for both methods were averaged and are presented jointly.

When combined with the DL estimator, the standard and Huber-White methods systematically showed the highest rejection rates on average, while the truncated Knapp-Hartung method provided the lowest rates. The influence of the different conditions manipulated in the simulation was similar for all of the methods. As expected, the number of studies showed a strong positive relationship with the statistical power. Note, however, that at least 40 studies were required for the different methods to provide average power rates close to the desired value of 0.8. Furthermore, the overall power rates were slightly greater as the average sample size per study increased. Finally, the amount of residual heterogeneity showed a negative relationship with the power, with larger residual τ^2_{res} values corresponding to smaller rejection rates.

The statistical power rates for the methods when using the ML estimator are presented in Figure 5.5. Again, results for the Huber-White method were not included, since the trends for this method were similar to the ones already described in combination with the DL estimator.

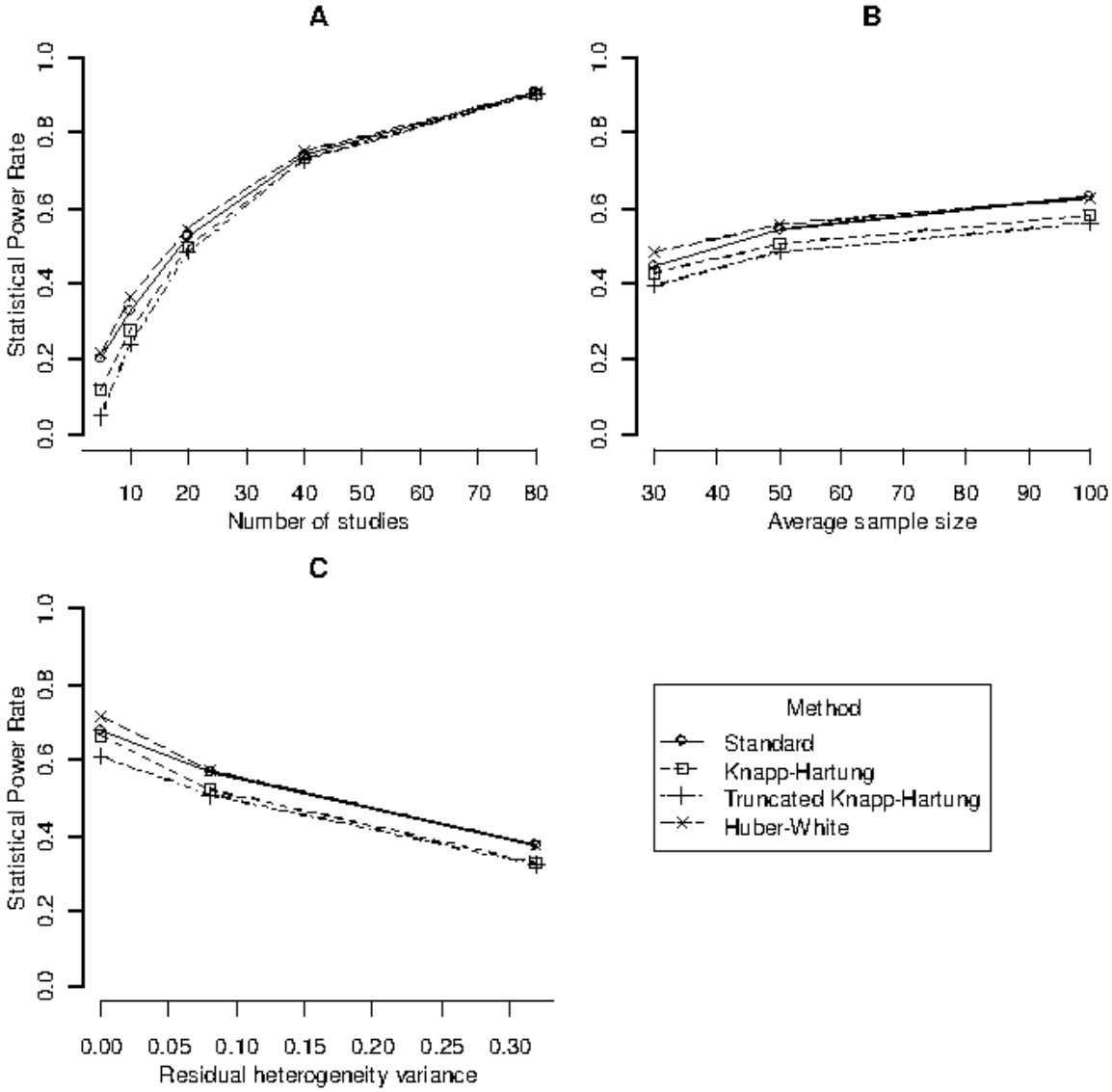
Figure 5.5 shows that the highest power rates were obtained with the standard and likelihood ratio tests, while the truncated Knapp-Hartung method yielded again the lowest rejection rates of the null hypothesis. Similar to the DL estimator, all methods combined with the ML estimator showed higher power rates as the number of studies increased, with the average sample size per study exerting a slight positive influence on the rejection rates. Also, as shown in Figure 5.5C, power for all methods decreased as the amount of residual heterogeneity among the true effects increased, with the rejection rate of the truncated Knapp-Hartung method gradually converging to that of the untruncated version of the test.

Figure 5.5 Statistical Power Rates of the methods when using the ML estimator



Lastly, Figure 5.6 shows the power results for the different testing methods when using the EB estimator. Results with this estimator did not provide any additional information, but showed the same trends described before for the four methods included.

Figure 5.6 Statistical Power Rates of the methods when using the EB estimator



Results for $\beta_1 = 0.5$ are not be presented here. With such a large slope value, all methods provided on average rejection rates over .80 with 20 or more studies. With smaller values of k, trends for the different methods were very similar to the ones described above for $\beta_1 = 0.2$.

5.5 Discussion

Several different methods are available for analyzing the association between one or more covariates and the effect size estimates under a mixed-effects model. In this chapter, a variety of different methods in the context of mixed-effects meta-regression models were compared. Specifically, seven residual heterogeneity variance estimators and six methods for testing the significance of the regression coefficients were compared in a Monte Carlo simulation study with standardized mean differences as the effect size index.

Two comparative criteria were considered for the assessment of the adequacy of each method across conditions similar to those typically found in psychological research. On the one hand, empirical Type I error rates were examined in order to assess which methods adequately control the rejection rate when a covariate is unrelated to the size of the effects. On the other hand, statistical power rates were obtained, with the aim to check which methods are more sensitive for the detection of a real moderating relationship. The results for the different procedures compared in this simulation were not found to be affected by the residual heterogeneity estimator computed. However, some differences were observed depending on the method employed for testing the regression coefficients.

Some authors have criticized that the standard method, applied in most meta-analyses when the influence of a moderator is tested under a mixed-effects model, does not take into account the uncertainty due to the variance estimation process, with the subsequent risk of reaching statistically significant results that might be inappropriate (e.g., Thompson & Higgins, 2002). When examining the empirical Type I error rates from the present simulation study, results for the standard method were in fact not satisfactory, with rates clearly over the nominal significance level in most situations, especially when some residual heterogeneity was present in the true effects and the number of studies was low. The liberal empirical Type I error rates showed by the standard method are the cause of its slightly higher statistical power exhibited in

comparison with the Knapp-Hartung and permutation tests. However, a test with deficient control of the Type I error rate should be avoided for routine use. Therefore, these results should encourage meta-analysts to consider alternative methods to the standard one, particularly when the number of studies in a research synthesis is small.

Due to the problems related to the standard method, some authors have developed various alternatives for testing the regression coefficients. The most widely employed one, up to date, is the Knapp-Hartung method, which incorporates a correction factor to the standard formula to estimate the variance-covariance matrix of the regression coefficients and whose statistical test is based on Student's t-distribution instead of the normal distribution assumed for the standard method. When this test was first proposed (Knapp & Hartung, 2003), the authors suggested truncating the correction factor to one if a smaller variance than that of the standard method was obtained. With this practice, the variance estimates of the regression coefficients would always be equal to or greater than the ones obtained with the standard method, so that the confidence intervals obtained with the Knapp-Hartung method would never be narrower than the standard ones.

The untruncated Knapp-Hartung method provided an adequate control of the Type I error rate, while truncating this method led to overly conservative results, as seen in Figures 5.1 to 5.3. Moreover, when comparing the methods in terms of their power in this simulation study, Figures 5.4 to 5.6 showed that the truncated Knapp-Hartung method provided systematically smaller rejection rates than all of the remaining methods under assessment. Therefore, results of the present study suggest better performance of the Knapp-Hartung method without the truncation of its correction factor. This is of particular concern, given that some software macros for meta-analysis (e.g., those that can be found in Stata) have implemented the Knapp-Hartung method only in combination with the truncation.

The trends described in the last paragraph for both versions of the Knapp-Hartung method, illustrated in Figures 5.1 to 5.6 for the DL, ML, and EB estimators, were also observed with the latter, despite the fact that the correction factor c is then always equal to one for positive values of $\hat{\tau}^2_{EB}$, as pointed out before. These results therefore suggest that the truncation proposed by Knapp and Hartung (2003) will make a difference especially in situations where the residual heterogeneity estimate is likely to require truncation as well (Borenstein et al., 2009).

The performance of the Huber-White and likelihood ratio tests was also assessed in the present study. As found in previous Monte Carlo simulations (Huizenga et al., 2011; Sidik & Jonkman, 2005), the results of the current simulation showed empirical Type I error rates clearly above the nominal significance level for both tests, with the Huber-White method providing higher proportions of incorrect rejections of the null hypothesis than all of the remaining methods. This trend was more evident when the number of studies was small.

Finally, the performance of a permutation test was also analyzed. This method provided results very similar to those of the (untruncated) Knapp-Hartung method. Both tests performed appropriately with respect to the empirical Type I error rates and their power rates were usually larger than those obtained for the truncated Knapp-Hartung method along the different simulated scenarios. The Knapp-Hartung method is, however, simpler to compute than the permutation test (the latter requires intensive computation), so that it seems a reasonable choice for most situations. Note, however, that the true effects were simulated as if one selects a random sample of studies from a superpopulation of studies (with normally distributed true effects). This corresponds to the usual conceptualization of the random/mixed-effects model in meta-analysis (Hedges & Vevea, 1998) and therefore also underlies the Knapp-Hartung method for testing the regression coefficients. In that sense, the Knapp-Hartung method is a suitable option as long as the set of studies can reasonably be assumed as a random sample from a broader

population of studies. On the other hand, if no random sampling of studies can be assumed, then the permutation test constitutes the most appropriate method (Manly, 1997).

The statistical power rates for all methods were clearly lower than .80 on average with less than 40 studies when the slope parameter had a small to moderate value (e.g., $\beta_1 = 0.2$ in this study). Moreover, all methods provided lower power rates as the residual heterogeneity among effect size parameters increased. An explanation for this fact is that, *ceteris paribus*, larger τ^2_{res} values will lead to a decrease in the predictive power of a model (see Section 3.4 of this dissertation).

In summary, the residual heterogeneity estimator did not show any influence on the different combinations here considered for testing the influence of a moderator under mixed-effects meta-regression models. Conversely, some discrepancies were found depending on the method applied for testing the regression coefficients. Specifically, too liberal results were obtained with the standard method, the most widely employed up to date in meta-analyses involving moderator analyses. Results of this simulation study suggest that, out of the different alternatives considered in the present study, the Knapp-Hartung method is a suitable option for most situations due to its satisfactory performance and computational simplicity. Overly conservative results were found for the Knapp-Hartung method when applying the truncation suggested by Knapp and Hartung (2003). Note that, as Figures 5.4A, 5.5A and 5.6A reveal, all of the methods compared in the present study required at least 40 studies to show power rates around 0.8 when the influence of the moderator on the effect sizes was of small to medium magnitude ($\beta_1 = 0.2$). Therefore, in order to maximize the probability to detect real moderating effects in a meta-analysis, the use of the Knapp-Hartung method without the truncation seems to be the best option.

The results of this simulation study are limited to the manipulated conditions. Although the values for the parameters and factors were chosen to represent typical conditions found in practice, additional simulation studies are needed including other scenarios not considered here and using different effect size indices (e.g., odds ratios, risk ratios, correlation coefficients).

The way moderators are tested in meta-analysis through mixed-effects meta-regression models is receiving increasing attention in the literature, and several new methods have recently been developed to conduct such analyses. Huizenga and colleagues (2011) proposed the use of a Bartlett-corrected likelihood ratio test which might improve the performance of the uncorrected likelihood ratio test regarding the control of the Type I error rate. Guolo (2012) also recently proposed a new likelihood-based test for meta-regression models. Finally, Friedrich and Knapp (2011, August) presented a new method that seems to outperform the Knapp-Hartung method in terms of coverage probability under adverse scenarios (small number of studies and very large heterogeneity of the sample sizes and true effects among the individual studies). These proposals were not considered for the present comparison of methods, although it should be very interesting to evaluate their performance in future simulation studies.

Chapter 6

Study 3: Alternatives for mixed-effects meta-regression models in the reliability generalization meta-analytic approach

6.1 The reliability generalization (RG) meta-analytic approach

Reliability is one of the most important psychometric properties to be considered when choosing a test for its administration in a specific context. However, reliability, as it is defined and estimated from the Classical Test Theory, is not a stable property for a given psychometric instrument, but rather a varying characteristic across different applications of the test (Dawis, 1987; Gronlund & Linn, 1990; Nunnally & Bernstein, 1994; Pedhazur & Schmelkin, 1991). Thus, in order to obtain a reliability estimate representative enough for

future test users, as well as to determine if one or more factors from the sample characteristics or the administration context have an influence on the reliability of test scores, the best alternative is to quantitatively integrate the reliability coefficients computed with scores from different applications of the instrument under study. And, when the aim is to carry out a quantitative synthesis, meta-analysis constitutes an optimal methodological choice (Hedges & Olkin, 1985; Sánchez-Meca & López-Pina, 2008).

Despite previous meta-analytic studies integrating reliability coefficients can be found in the literature (e.g., Churchill & Peter, 1984; Conway, Jako, & Goodman, 1995; Parker, 1983; Parker, Hanson, & Hunsley, 1988; Peter & Churchill, 1986; Salgado & Moscoso, 1996; Yarnold & Mueser, 1989), the term *reliability generalization (RG)* was firstly proposed by Vacha-Haase (1998). In an RG study, a set of reliability estimates from the same test are integrated, an overall reliability estimate is obtained and heterogeneity between the individual reliability coefficients is assessed (e.g., López-Pina, Sánchez-Meca, & López-López, 2012; Sánchez-Meca, López-López, & López-Pina, in press). Moreover, since some heterogeneity across estimates is usually found, a third objective in an RG study consists of looking for moderator variables in order to explain part of that variability.

Although Vacha-Haase's seminal paper was published just a few years ago, several dozens of RG studies have already been conducted. A great variability can be found among these studies in terms of rigor, theoretical underpinning, and methodology. This is partially due to the fact that the RG approach was not conceived as monolithic in terms of the statistical methods applied (Henson & Thompson, 2002; Vacha-Haase, 1998; Vacha-Haase & Thompson, 2011). As a consequence, there is no consensus about several methodological issues affecting the statistical analyses.

One of these issues involves reliability coefficients transformation. Some authors did not consider it necessary to transform the reliability coefficients for the statistical analyses (e.g., Bonett, 2002, 2010; Henson & Thompson, 2002; Leach, Henson, Odom, &

Cagle, 2006; Mason, Allam, & Brannick, 2007; Vacha-Haase, 1998). However, the sampling distribution for the most usual reliability coefficients (e.g., alpha coefficients and Pearson correlations) is skewed, with a larger asymmetry level as the parameter approximates to one (Rodriguez & Maeda, 2006), as it is usually the case for reliability coefficients reported in primary studies. Thus, some other authors recommended applying some transformation on the reliability coefficients in order to normalize their distribution and to stabilize their variances (e.g., Feldt & Charter, 2006; Rodriguez & Maeda, 2006). At least three different transformation formulae – already presented in Chapter 2 – have been proposed and/or applied in the RG literature.

Another issue for which different solutions have been applied so far in the RG approach is the weighting scheme of the reliability coefficients. Some authors just employed OLS analyses in their RG studies, that is, without weighting the reliability coefficients (e.g., Kieffer & Reese, 2002; Leach et al., 2006; Vacha-Haase, 1998). Nonetheless, sample sizes in RG meta-analyses are usually unequal, leading to unequal sampling variances for the reliability coefficients, so that the homoscedasticity assumption – required for OLS techniques – is rarely met (Raudenbush, 1994; Rodriguez & Maeda, 2006). When weights were included in the analyses, some researchers chose the sample size as the weighting factor (e.g., Victorson, Barocas, & Song, 2008; Yin & Fan, 2000; Zangaro & Soeken, 2005), according to the proposal of Hunter and Schmidt (2004), while some others chose the inverse variance of the reliability coefficients (e.g., Aguayo, Vargas, de la Fuente, & Lozano, 2011; Beretvas, Suizzo, Durham, & Yarnell, 2008; López-Pina, Sánchez-Meca, & Rosa-Alcázar, 2009). Inverse variances have been used as weights in most of the meta-analyses published up to date (Borenstein et al., 2010), and they are also becoming more and more frequent in the RG approach.

When the inverse variance is employed as the weighting scheme, it is necessary to assume some statistical model (Sánchez-Meca, López-López, & López-Pina, in press). In a fixed-effect model, an estimate of the within-study variance is required for the analyses,

and several formulae are available for raw reliability coefficients as well as for the different transformations proposed in the literature. This implies that, once the transformation (or no transformation) of the reliability coefficients is chosen, the estimation method for the sampling variance is essentially unique. As mentioned in Chapter 1, the fixed-effect model allows for generalizing results only to the samples whose reliability coefficients were included in the meta-analysis, and also to some external situations where the administration conditions and sample characteristics were identical to those of the studies included in the meta-analysis (Hedges & Vevea, 1998). Like in other application fields of meta-analysis, simulation studies have warned about the limitations of the fixed-effect model in the RG approach (Romano & Kromrey, 2009). The varying coefficient model mostly circumvents those limitations (Bonett, 2010) but, as in the fixed-effect model, conclusions can only be extended to samples with identical characteristics and composition to those included in the RG meta-analysis.

An alternative is to assume a random-effects model, which is considered nowadays as the most realistic option in the general meta-analytic arena (Cooper et al., 2009; National Research Council, 1992) and in the RG approach (Rodriguez & Maeda, 2006). The main reason to assume a random-effects model is that, unlike the fixed-effect and the varying coefficient models, it allows for generalizing results beyond the test administrations included in the meta-analysis (Borenstein et al., 2010; Hedges & Vevea, 1998). The random-effects model assumes that the integrated reliability coefficients are estimating a random sample of parametric reliability coefficients extracted from a bigger superpopulation. In practice, that implies estimating a second variance component, the heterogeneity variance, and different procedures to accomplish this goal are available (see Section 3.2). Since the aim in an RG meta-analysis is usually to generalize results beyond the set of studies integrated, the weighting scheme throughout this study will be that based on the inverse variance assuming a random-effects model.

Moderator analyses constitute a crucial step in the RG approach (Rodriguez & Maeda, 2006), given the fact that most of the RG studies published so far found statistically significant relationships of one or more variables to the reliability coefficients. As the psychometric theory predicts, several moderators associated to the variability of test scores have shown a statistically significant relationship with the reliability coefficients in many RG studies (e.g., standard deviation of test scores, type of population from which the sample subjects were recruited) and, for this reason, it has been argued that predictive models of the heterogeneity between reliability coefficients should always include some of them (Botella & Ponte, 2011). Other moderators which have proved a significant relationship with the reliability coefficients in previous RG studies are related to the test version (e.g., test length or original vs. adapted version).

When one or more predictors are included in the model, it becomes necessary to estimate the regression coefficients and, depending on the transformation applied to the reliability coefficients, these estimates will change to some extent. Also, a new estimate of the (now residual) heterogeneity variance, which reflects the amount of variability on the coefficients not accounted for by the moderators incorporated to the model, is required to be included into the weighting factor of mixed-effects analyses. As shown in Section 3.2 of this dissertation, different procedures are available to compute that estimate, and the estimator of choice might have an influence on the results.

Apart from this, statistical tests for the regression model coefficients are required to test the association of some moderator(s) with the reliability estimates. The method traditionally computed in RG meta-analyses for addressing that issue, which was presented in Section 3.3.1 of this dissertation, has been criticized in the last years, since its performance is strongly dependent on the accuracy of the variance estimates (Brockwell & Gordon, 2001). The correction proposed by Knapp and Hartung (2003), also mentioned in Section 3.3, has not been employed yet in any published RG meta-analysis, and it

should be interesting to assess its performance in order to determine whether its implementation in the RG context is advisable or not.

6.2 Objectives, previous simulation studies, and hypotheses

6.2.1 Objectives of the study

Since various methodological alternatives are available when fitting mixed-effects meta-regression models in the RG approach, the aim of the present study was to compare the performance of different combinations of methods under some realistic scenarios in RG studies, by means of Monte Carlo simulation. Specifically:

- Two estimators of the residual heterogeneity variance, DL and REML, were compared. Both procedures were presented in Section 3.2. The DL estimator has been almost the only procedure employed so far in RG studies assuming a random-effects model, while the REML estimator constitutes a reasonable alternative because of its appropriate performance in previous simulation studies (cf. Viechtbauer, 2005; see also Chapter 4 of this dissertation).

- Two methods for testing the significance of the model regression coefficients, standard and untruncated Knapp-Hartung procedures, were incorporated. The former has been the only method employed by RG meta-analysts up to date, while the latter represents an appealing alternative provided its good performance in previous simulation studies using different effect size indices (Knapp & Hartung, 2003; Sidik & Jonkman, 2005a; see also Chapter 5 of this dissertation). Both methods were also described in Section 3.3.

- Four outcome variables for RG studies were considered, including untransformed alpha coefficients, Fisher's Z, Hakstian-Whalen, and Bonett's transformations. All

of these outcome variables were presented in Chapter 2. Since all of them have been employed in the RG literature, another objective of the present study was to determine whether the transformation choice might have an influence on the results.

The combination of these procedures led to 16 methodological alternatives. Bias and efficiency were studied for the different estimation methods of the model regression coefficients, and the empirical Type I error and statistical power rates of their corresponding significance tests were then compared for all methodological combinations. Regarding outcome variables, coefficient alpha is the most widely reported reliability measure in primary studies and, since mixing different types of reliability coefficients is not appropriate (cf. Rodriguez & Maeda, 2006), most RG studies published so far have employed coefficient alpha as the main dependent variable. Consequently, the simulation study presented along this chapter employed alpha coefficients (transformed or untransformed) as the dependent variable.

6.2.2 Previous simulation studies

A few simulation studies have been already conducted in the RG meta-analytic approach. Mason et al. (2007) carried out a simulation study comparing the performance of different methods in mixed-effects models. However, the dependent variable in their study was the test-retest correlation instead of the alpha coefficient. Also, while these authors focused on the efficiency of the different methods included for estimating the model slope, in the present simulation bias, empirical Type I error and statistical power rates were also considered as comparative criteria.

Another simulation study was carried out by Feldt and Charter (2006), who compared different approaches for averaging internal consistency coefficients, some of

them incorporating either the Fisher's Z or the Hakstian-Whalen transformations. López-Pina and colleagues (2012) focused on the overall reliability estimate as well in their simulation study, comparing several procedures considered by Feldt and Charter (2006) under more realistic scenarios. A limitation of the Feldt and Charter's (2006) and López-Pina et al.'s (2012) studies is that they applied a fixed-effect model, instead of a random-effects one. Also, Bonett's transformation was employed in some recent simulation studies (Bonett, 2010; Romano & Kromrey, 2009). In the present study, Fisher's Z, Hakstian-Whalen, and Bonett's transformations for the reliability coefficients were included. Romano, Kromrey, and Hibbard (2010) also considered all transformations in their simulation study, although these authors assessed the performance for computing confidence intervals around the average reliability estimate. Lastly, Enders (2004) applied the Monte Carlo to the problem of handling missing data.

6.2.3 Hypotheses of this study

Regarding the hypotheses of the present study, it was expected that:

1. The residual heterogeneity variance estimator would not affect the results for the different methods, as found in previous studies (Sánchez-Meca & Marín-Martínez, 2008; see also Chapter 5 of this dissertation).

2. The alternatives including the Knapp-Hartung correction would perform better than the ones combined with the standard method in terms of empirical Type I error rate, as reported by the authors in their seminal paper (Knapp & Hartung, 2003).

3. The transformed methods would outperform the untransformed ones, especially when comparing the empirical Type I error and statistical power rates for the slope

tests, which are supposed to be more affected when the normality assumption is not met.

4. The transformed methods recommended for alpha coefficients (Hakstian-Whalen and Bonett's transformations) would perform better than Fisher's Z for the estimation and statistical testing of the meta-regression slopes.

6.3 An illustrative example

An example is presented here in order to illustrate the 16 resulting methods after combining four alternatives for transforming the reliability coefficients (including untransformed reliability coefficients), two residual heterogeneity variance estimators, and two methods for testing the regression coefficients. Data for the example were extracted from an RG study about the Hamilton Rating Scale for Depression (López-Pina et al., 2009), and are presented in Table 6.1. Considering the samples for which the 17-item version was administered, a meta-regression model was fitted using each one of the proposed methods, with the standard deviation from each sample's scores, SD_i, as the predictor, and the untransformed coefficient alpha, $\hat{\alpha}_i$, as the dependent variable.

Most of the reliability estimates were over the 0.7 boundary proposed by Nunnally and Bernstein (1994; see also Chapter 2 of this dissertation), with the exception of the estimate reported by Bent-Hamsen and colleagues (2003). Moreover, The Hamilton Rating Scale for Depression was applied to samples which ranged between 23 and 921 subjects, and the standard deviations of the total scores took values between 3 and 7.51.

Table 6.1 Data from the RG meta-analysis conducted by López-Pina et al. (2009)

Study	N_i	$\hat{\alpha}_i$	SD_i
Addington, Addington, Maticka-Tyndale, and Joyce (1992)a	50	.660	4.38
Addington, Addington, Maticka-Tyndale, and Joyce (1992)b	100	.770	7.10
Akdemir et al. (2001)	94	.750	6.89
Bent-Hamsen et al. (2003)	230	.420	3.00
Bobes et al. (2003)	165	.740	5.60
Kobak and Reynolds (2000)	921	.897	7.51
Leidy, Palmer, Murray, Robb, and Revicki (1998)	48	.860	6.32
Ramos and Cordero (1988)	135	.770	5.10
Rapp, Smith, and Britt (1990)	150	.830	6.01
Reynolds and Mazza (1998)	89	.850	5.74
Riskind, Beck, Brown, and Steer (1987)	120	.730	6.84
Rush et al. (1986)	289	.800	7.10
Rush et al. (2003)	552	.880	3.00
Stage, Middelboe, and Pisinger (2003)	49	.850	7.10
Thunedborg, Black, and Bech (1995)	23	.835	5.74

Table 6.2 presents the estimates for the model slope and the p-values for its statistical significance from each single analysis. When some transformation was applied on the reliability coefficients, the slope estimates were back-transformed. Taking the

combination of untransformed reliability coefficients with DL estimator as a reference, the regression equation was:

$$Y_i = 0.594 + 0.0326 SD_i$$

Table 6.2 Slope estimates and associated p-values from the example

	DL			REML		
	$\hat{\beta}_1$	p_{STD}	p_{KH}	$\hat{\beta}_1$	p_{STD}	p_{KH}
Raw alpha coefficients	.0326	.064	.118	.0339	.079	.107
Fisher's Z	.0442	.257	.141	.0431	.140	.147
Hakstian-Whalen	.0424	.192	.125	.0412	.116	.112
Bonett	.0441	.284	.161	.0431	.155	.166

$\hat{\beta}_1$: slope estimate. **DL, REML:** DerSimonian and Laird and Restricted Maximum Likelihood estimators for the residual heterogeneity variance. p_{STD}, p_{KH}: p-values corresponding to the standard method and the Knapp-Hartung correction for testing the regression coefficients, respectively.

Regarding the results with the 16 procedures, the slope estimates were around .033 when the untransformed reliability coefficients were employed as the dependent variable, and values over .04 were obtained when using some transformation. P-values for the slope tests showed important discrepancies depending on the method considered. Assuming a 95% confidence level, statistically significant results were not achieved in any case; however, marginally significant results were found when applying the standard method for testing regression coefficients combined with raw alpha coefficients, both for the DL and the REML estimators (p-values of .064 and .079, respectively). Conversely, the remaining methods provided p-values greater than .10.

6.4 Simulation study

A simulation was carried out to compare the 16 alternative methods for fitting mixed-effects meta-regression models presented above. The simulation was programmed in R, using *metafor* (Viechtbauer, 2010) and *MCMCpack* (Martin, Quinn, & Park, 2011) packages. This simulation was conducted under the Classical Test Theory framework (Gulliksen, 1987), because most of the tests chosen in previous RG studies were made based on that theoretical approach.

Regarding manipulated factors in this simulation, sample sizes, N_i, were generated from a log-normal distribution with a mean value of 150 participants. The asymmetry of the sample size distribution was one of the conditions manipulated, with values of +1, +2, and +3, according to empirical asymmetry values observed in previous RG databases (e.g., Botella, Suero, & Gambara, 2010; López-Pina et al., 2009; Sánchez-Meca et al., 2011). Also, the number of studies for each meta-analysis, k, was set to values of 15, 30, and 60. Lastly, for the slope parametric value, two different scenarios were considered: for the first set of conditions, a predictor variable was generated from a distribution $N(0,1)$ with no relationship to the reliability coefficients, so that the expected value for the slope was 0; for the second scenario, the error component in the test scores was generated as a function of that predictor, leading to a mean empirical slope, $\overline{\beta}_1$, of .01348 for all conditions (values between .01346 and .01350).

A key aspect in the simulation was the computation of the parametric coefficients alpha. In a first step, population test scores for each study were generated. Considering settings described in previous simulations (Bonett, 2010; Botella & Suero, 2012), a 20-item test was defined. For the calculation of each parametric coefficient alpha, a population of 10,000 subjects was defined. True scores for each of the 20 items, t_q, were generated from a multivariate normal distribution with mean 0, variance 2 for each item and covariance 0.4 for any pair of items. This provided a $(10{,}000 \times 20)$ matrix of true scores for

each study. Then, error scores for each item, e_{sq}, were generated from a normal distribution with mean 0, the variance changing from one study to the next due to the predictor value, with a range between .1 and 1.9. This resulted in another $(10{,}000 \times 20)$ matrix of error scores for each study. The observed scores for each of the 10,000 subjects in the qth item, x_{sq}, were calculated with the expression (Crocker & Algina, 1986)

$$x_{sq} = t_{sq} + e_{sq}.$$

Finally, scores for each subject in the whole test were computed as $x_s = \sum_{q=1}^{20} x_{sq}$. The parametric alpha coefficients, α_i, were computed from the database of 10,000 subjects generated for each study.

In a second step, samples of N_i subjects were taken from the respective populations, and the empirical alpha coefficients, the three proposed transformations, and their respective sampling variances were computed with the formulae presented in Section 2.3 of this dissertation. This process – generating a database of 10,000 subjects and then extracting a sample of N_i of them – was replicated k times in order to simulate the data corresponding to the k studies in an RG meta-analysis.

Once obtained the sample reliability coefficients and within-study variances for the k studies, results for each meta-analysis were obtained by fitting mixed-effects meta-regression models for the 16 statistical alternatives under comparison. For each condition, 10,000 meta-analyses were computed.

Regarding comparative criteria, the bias and the mean square error (MSE) for the slope estimates were firstly computed in the conditions where $\beta_1 \neq 0$ for each one of the 8 combinations (4 transformation methods x 2 residual heterogeneity variance estimators), providing different estimates of the model coefficients. When some

transformation was applied on the reliability coefficients, the slope values were back-transformed using the procedure described in Equation (2.17). Mathematical computations required for obtaining bias and MSE are provided below.

Let $\hat{\beta}_1^m$ be the slope estimate obtained with any of the proposed methods, and back-transformed to the reliability coefficients metric where necessary. The average of $\hat{\beta}_1^m$ for any given condition was computed with (Marín-Martínez & Sánchez-Meca, 2010)

$$AVE\left(\hat{\beta}_1^m\right) = \frac{\sum_i \hat{\beta}_{i1}^m}{10{,}000}. \tag{6.1}$$

Then, bias was obtained with

$$BIAS\left(\hat{\beta}_1^m\right) = AVE\left(\hat{\beta}_1^m\right) - \overline{\beta}_1, \tag{6.2}$$

where $\overline{\beta}_1$ is the average of the empirical parametric slopes obtained along the 10,000 meta-analyses. On the other hand, MSE was calculated with

$$MSE\left(\hat{\beta}_1^m\right) = \frac{\sum_i \left(\hat{\beta}_{i1}^m - \overline{\beta}_1\right)^2}{10{,}000}. \tag{6.3}$$

Finally, the proportion of rejections of the null hypothesis $\beta_1 = 0$, assuming a 95% confidence level, was computed for all 16 combinations. That led to compare the different methods in terms of empirical Type I error rates for conditions where $\beta_1 = 0$, and in terms of statistical power rates when $\overline{\beta}_1 \neq 0$.

6.5 Results

In order to illustrate the general trends of the simulated data, some descriptive statistics are presented in Table 6.3. Descriptives from the observed scores, x_s, were obtained after generating a database of 1,000,000 scores. Next, data from 1,000 studies were simulated with an asymmetry index of 2 for the sample size distribution, computing for each study the score variance, S_X^2, and coefficient alpha estimate, $\hat{\alpha}_i$.

Table 6.3 Descriptive statistics from the simulated data

Statistic	x_s	S_X^2	$\hat{\alpha}_i$
Minimum	-70.501	106.576	0.481
Maximum	65.646	328.786	0.818
Mean	0.003	212.211	0.719
Median	-0.007	210.336	0.720
Variance	211.966	796.081	0.001
Skewness	0.003	0.273	-0.989
Kurtosis	0.004	1.154	4.523

In the remainder of this section, the different methods described above will be assessed by means of the comparative criteria considered in the Monte Carlo simulation. Firstly, the accuracy of the slope estimates will be compared for the eight methodological alternatives (after combining four transformation methods and two residual heterogeneity variance estimators), in terms of bias and MSE. Then, the performance of the slope statistical tests will be assessed for the 16 available alternatives (as a result of

combining the 8 previous methods either with the standard method or with the Knapp-Hartung correction), in terms of empirical Type I error and statistical power rates.

6.5.1 Accuracy of the slope estimates

Tables 6.4 and 6.5 present bias and MSE results, respectively, for the eight estimators of the meta-regression model slope. In order to facilitate their interpretation, values on both tables were multiplied by 10,000, so that the reference slope value is now 134.8.

Table 6.4 Bias in the slope estimates for the different combinations of methods

k		**15**			**30**			**60**		
Asymmetry		1	2	3	1	2	3	1	2	3
Raw alpha coefficients	DL	-2.517	-0.587	-1.847	-1.962	-1.158	-0.669	-2.235	-1.999	-1.141
	REML	-2.552	-0.605	-1.802	-1.972	-1.138	-0.646	-2.238	-1.998	-1.138
Fisher's Z	DL	-2.904	-1.665	-2.412	-1.659	-0.852	-0.283	-1.401	-1.433	-0.668
	REML	-2.902	-1.611	-2.400	-1.660	-0.848	-0.275	-1.401	-1.434	-0.678
Hakstian-Whalen	DL	-3.184	-1.757	-2.767	-2.174	-1.412	-0.912	-2.129	-2.071	-1.305
	REML	-3.203	-1.719	-2.738	-2.183	-1.417	-0.895	-2.130	-2.071	-1.314
Bonett	DL	-3.850	-2.636	-3.460	-2.557	-1.773	-1.248	-2.297	-2.309	-1.555
	REML	-3.873	-2.534	-3.363	-2.565	-1.787	-1.281	-2.298	-2.318	-1.550

k : number of studies. *Asymmetry*: skewness of the sample size distribution. DL and REML: DerSimonian and Laird and Restricted Maximum Likelihood estimators for the residual heterogeneity variance.

Results in Table 6.4 show that all conditions provided negatively biased estimates of the slope parameter, although that bias was smaller than 3% for any combination of methods. Results were very similar and showed identical trends regardless of the residual variance estimator, but some differences were observed depending on the transformation method. Specifically, Bonett's transformation systematically showed the highest bias rates, with the largest percentage of bias, around -2.9%, when both the asymmetry in the sample size distribution and the number of studies were small. In contrast, raw alpha coefficients provided bias results slightly smaller than the methods involving some transformation of the reliability coefficients when the asymmetry was small, while the Fisher's Z transformation led to the smallest bias for larger values in the asymmetry of the sample size distribution and in the number of studies.

Table 6.5 MSE in the slope estimates for the different combinations of methods

k		**15**			**30**			**60**		
Asymmetry		1	2	3	1	2	3	1	2	3
Raw alpha coefficients	DL	.6545	.7341	.7651	.2900	.3050	.3209	.1355	.1413	.1438
	REML	.6545	.7341	.7647	.2899	.3049	.3207	.1355	.1413	.1439
Fisher's Z	DL	.6344	.6963	.7111	.2803	.2842	.2941	.1305	.1339	.1344
	REML	.6344	.6953	.7077	.2803	.2842	.2933	.1305	.1339	.1343
Hakstian-Whalen	DL	.6301	.6954	.7164	.2778	.2851	.2959	.1294	.1333	.1340
	REML	.6301	.6944	.7123	.2778	.2849	.2947	.1294	.1333	.1339
Bonett	DL	.6219	.6832	.7006	.2739	.2786	.2882	.1277	.1311	.1315
	REML	.6218	.6812	.6935	.2739	.2782	.2859	.1277	.1309	.1310

k : number of studies. *Asymmetry*: skewness of the sample size distribution. DL and REML: DerSimonian and Laird and Restricted Maximum Likelihood estimators for the residual heterogeneity variance.

Regarding efficiency, results in Table 6.5 show some interesting trends as well. MSEs were slightly higher for all the methods as the asymmetry values increased, but the number of studies showed a bigger influence decreasing the MSEs for larger values of k. Again, results were almost identical both for the DL and the REML estimators. Focusing on the transformation method, however, raw alpha coefficients provided the largest MSEs along all of the simulated conditions, while the smallest values were obtained when applying Bonett's transformation.

6.5.2 Performance of the hypothesis tests for the slope

Table 6.6 reports the empirical Type I error rates for the different methods under comparison, while statistical power rates are provided in Table 6.7. In both tables, only results for the DL estimator are presented, since rates obtained with the REML were very similar and fully comparable in terms of the observed trends.

Assuming a 95% confidence level, accurate results for each method should be around .05 when the slope parametric value is 0. Results presented in Table 6.6 show that the rejection rates for the standard method were clearly under the nominal significance level, with rates smaller than .01 for the Fisher's Z transformation and around .02 for the remaining transformation procedures. In contrast, the Knapp-Hartung correction performed close to the nominal level for all of the transformation methods and along all of the simulated conditions.

Table 6.6 Empirical Type I error rates for the slope tests using the DL estimator

k		15			30			60		
Asymmetry		1	2	3	1	2	3	1	2	3
Raw alpha coefficients	STD	.017	.020	.021	.014	.018	.020	.017	.018	.021
	KH	.048	.053	.054	.049	.050	.050	.050	.053	.055
Fisher's Z	STD	.006	.007	.007	.006	.006	.006	.005	.005	.006
	KH	.046	.049	.049	.050	.048	.043	.048	.048	.048
Hakstian-Whalen	STD	.017	.020	.023	.017	.019	.019	.019	.020	.021
	KH	.046	.049	.049	.049	.049	.044	.047	.049	.048
Bonett	STD	.016	.020	.022	.018	.020	.019	.020	.020	.021
	KH	.046	.048	.049	.049	.047	.043	.048	.048	.047

k : number of studies. Asymmetry: skewness of the sample size distribution. STD and KH: standard method and Knapp-Hartung correction for testing the regression coefficients.

Regarding statistical power rates, Table 6.7 shows that the lowest rates were obtained when combining the standard method and the Fisher's Z transformation. The Knapp-Hartung correction systematically led to higher power rates than those obtained for the standard method. Apart from that, all rates increased for larger values of k, while the asymmetry showed a small inverse relationship with the rates. Lastly, power rates were slightly higher when the Knapp-Hartung correction was combined with some transformation of the reliability coefficients.

Table 6.7 Statistical power rates for the slope tests using the DL estimator

k		15			30			60		
Asymmetry		1	2	3	1	2	3	1	2	3
Raw alpha coefficients	STD	.266	.271	.259	.561	.552	.556	.884	.871	.875
	KH	.369	.363	.347	.682	.666	.658	.942	.929	.925
Fisher's Z	STD	.171	.170	.166	.425	.418	.422	.810	.795	.799
	KH	.368	.364	.348	.684	.675	.667	.942	.931	.930
Hakstian-Whalen	STD	.283	.282	.271	.587	.577	.580	.902	.890	.892
	KH	.370	.363	.341	.685	.671	.661	.943	.932	.929
Bonett	STD	.284	.281	.270	.586	.578	.581	.901	.892	.894
	KH	.369	.362	.345	.684	.674	.664	.942	.931	.930

k : number of studies. *Asymmetry*: skewness of the sample size distribution. STD and KH: standard method and Knapp-Hartung correction for testing the regression coefficients.

6.6 Discussion

The present study focused on the analyses of continuous moderators by fitting mixed-effects meta-regression models using alpha coefficients as the outcome variable. In this study, different procedures for transforming reliability coefficients were compared (see Section 2.3 of this dissertation). Extensions of the DL and REML estimators for the residual heterogeneity variance (presented in Section 3.2) were also assessed, as well as the standard method for testing the regression coefficients and the adjustment proposed for Knapp and Hartung (2003) to the former, both of them presented in Section 3.3 of this dissertation. Performance for all the presented methods was compared by means of

Monte Carlo simulation, where the bias and the MSE for the slope estimates, as well as the empirical Type I error and statistical power rates for the slope tests, were the comparative criteria considered.

Out of the different methodological issues implied in the several procedures here compared, the choice of the residual heterogeneity variance estimator (DL vs. REML) produced negligible differences in the trends, the changes observed in the results for the different conditions being very small. On the other hand, the transformation method of the reliability coefficients exerted some influence on the comparative criteria considered for this study. Lastly, the method employed for testing the significance of regression coefficients (standard vs. Knapp-Hartung) showed a critical influence on the empirical Type I error and statistical power results.

Regarding transformations, in terms of bias, all methods provided negatively biased estimates of the regression coefficients, although raw alpha coefficients showed results slightly better than the ones obtained when applying some transformation, especially when the asymmetry in the sample size distribution was small. Conversely, MSEs were higher for untransformed reliability coefficients than for any of the transformed methods. However, since bias results were always smaller than 3% regarding the slope parameter, and MSE values were also small and very similar from one method to another, the conclusion should be that all four transformation methods performed reasonably well in terms of bias and efficiency. Also, from a conceptual point of view, Fisher's Z transformation should not be used with coefficients alpha, as that transformation is only appropriate when the reliability coefficients were computed as a Pearson correlation coefficient (e.g., test-retest reliability). Therefore, for coefficients alpha Hakstian and Whalen's (1976) and Bonett's (2002) transformations should be selected.

Considering now the two methods here included for testing the model coefficients, compared to the standard method, the Knapp-Hartung correction provided empirical Type

I error rates closer to the nominal significance level, performing almost nominally for all combinations and under all of the simulated scenarios. Regarding statistical power, the Knapp-Hartung correction showed higher power rates than the standard method regardless of the rest of conditions manipulated. These power rates were slightly higher when the Knapp-Hartung correction was combined with some transformation of the reliability coefficients. However, a noteworthy finding is that, when integrating 15 or 30 coefficients alpha, as it was the case for some previous RG studies, power rates were considerably lower than the .80 boundary recommended by the scientific community (Cohen, 1992). Thus, having a moderate to large number of reliability coefficients seems to be an important requirement when conducting moderator analyses in RG studies.

6.7 Usefulness and limitations of the findings presented in this chapter

The present simulation study showed that, when fitting mixed-effects meta-regression models with one covariate, the slope estimates can be negatively biased, although usually that bias is not large enough to represent a threat to the results. Also, despite MSEs for these estimates were smaller when some transformation on the reliability coefficients was applied, results were very similar when comparing different transformation methods, and MSEs decreased noticeably as the number of reliability coefficients increased. These results therefore suggest that all transformation methods here compared perform similarly in terms of bias and efficiency of the model slope estimates, so that researchers conducting RG studies should pay more attention to some other criteria before making their decisions about the statistical methods implemented.

In contrast to the previous statement, significance tests for the slope did show important differences along the methodological alternatives compared here. According to results here presented, RG researchers should take into account that testing the model

coefficients with the standard method may lead to a loss of statistical power, as Table 6.7 reflects, so that some moderators of the variability between reliability coefficients might not be identified in their RG studies unless they are integrating a large number of reliability coefficients. The Knapp-Hartung correction outperformed the standard method in terms of statistical power, with rates systematically greater than those obtained with the standard method, and showed empirical Type I error rates closer to the nominal significance level.

Regarding limitations of the methods included in the present study, the fact that only mixed-effects models were considered here might be seen as problematic, since some other options are present in published RG studies. However, the purpose of this chapter was not to assess the methodological choices implemented up to date, but rather to compare the best methodological alternatives for future studies, based on the main objectives in an RG study itself and on the current statistical alternatives to accomplish them. Since reliability is not a stable property for a given psychometric instrument (e.g., Crocker & Algina, 1986; Gronlund & Linn, 1990), the RG approach was proposed by Vacha-Haase (1998) as a way to integrate a set of reliability estimates from different applications of a test, and to guide expectations of potential test users about reliability with their sample characteristics and their administration context. That implies generalizing results to some other scenarios not necessarily identical to the ones accounted for in the RG study, and only random-effects models allow researchers for making such generalizations (cf. Beretvas & Pastor, 2003; Borenstein et al., 2010; Hedges & Vevea, 1998; Raudenbush, 2009; Sánchez-Meca, López-López, & López-Pina, in press; Schmidt et al., 2009). Assuming a random-effects model when conducting moderator analyses leads to mixed-effects models, as the ones here presented. In addition to inverse variances, sample sizes can also be considered as random-effects weights. However, results are not expected to be influenced by the choice of weights in a random-effects model, but rather by the transformation method in the coefficients (Mason et al., 2007).

Also, as in any simulation, conditions manipulated in this study cannot account for the whole universe of scenarios present in RG studies already carried out or to be done in the future. As an illustration of that, some RG studies have integrated larger numbers of sampling reliability estimates than the ones considered here (e.g., Yin & Fan, 2000), and similar results to the ones presented here for 60 studies should be expected with smaller MSEs and an additional gain of statistical power. Moreover, generating sample size values from a log-normal distribution may be a reasonable approximation to the real situation in many RG meta-analyses (Mason et al., 2007), where most of the primary studies used small to moderate inpatient samples while a few ones applied the test as a screening instrument to large samples from general population. Increasing the asymmetry of the sample size distribution produced slightly higher MSEs and smaller statistical power rates, although that factor did not show a big influence on any of the criteria compared here.

Finally, the use of coefficient alpha in this simulation study, as well as in most of the RG studies published up to date, leads to some noteworthy considerations. As Graham (2006) remarked, coefficient alpha is based on the essentially tau-equivalent measurement model. This implies that, when a coefficient alpha is computed, it is assumed that all items measure the same latent trait, although probably with a different degree of precision. Researchers estimating reliability with coefficient alpha, or retrieving alpha coefficients for carrying out an RG study, must be aware of this assumption, because its violation would directly affect the validity of the reliability estimates for a given test. The generating process of the item scores in the present simulation, which was detailed above, fulfilled the requirements of the essentially tau-equivalent measurement model.

The RG approach was recently proposed (Vacha-Haase, 1998), with the aim of applying a methodology for quantitative synthesis, meta-analysis, to the purpose of obtaining a representative reliability value along different administrations of a given test, as well as identifying which factors can explain variability across the set of reliability estimates. The latter objective implies carrying out moderator analyses, and different

alternatives for addressing that issue are available to the meta-analyst. Results of this study mainly suggest that, when a mixed-effects model is assumed for the moderator analyses in an RG study, the Knapp-Hartung correction for the statistical test of the model coefficients provides rates closer to the nominal significance level regarding Type I error, and higher power rates than the ones obtained for the standard method. Performance for that correction seems then promising in the RG approach, where it has not been applied up to date.

Chapter 7

Conclusions

Meta-analysis constitutes a great improvement to traditional, non-quantitative syntheses in terms of precision, reliability, and validity (Cooper & Hedges, 2009b). Since it was firstly proposed by Gene V. Glass (1976), this methodology has been improved and widely applied in many different fields such as Behavioral, Health, and Biological Sciences (e.g., Cooper et al., 2009; Marín-Martínez et al., 2009). In a meta-analysis, each study is usually weighted by a function of its precision (e.g., Pigott, 2001). When the results from a set of individual studies are found to be discrepant, meta-analysis allows the researcher to search for moderating influences that can explain part of that variability. And, as it was shown in Section 1.3, such moderator analyses can be conducted by fitting mixed-effects regression models.

In Chapter 3 of this dissertation, the different methods available when estimating and testing the most relevant parameters in mixed-effects meta-regression models were presented. One of these parameters is the residual heterogeneity variance, which represents the amount of unexplained heterogeneity among the individual outcomes

different to sampling error after adding one or more moderators to the model (Viechtbauer, 2008). Seven estimators of that parameter were presented in Section 3.2. Since the residual heterogeneity variance is included in the weights for meta-regression analyses when assuming a mixed-effects model, obtaining accurate estimates of this parameter constitutes an important issue. One of these analyses is the significance test of the moderator(s) included in the model, and six alternative methods to accomplish that objective were described in Section 3.3. Finally, for the estimation of the predictive power of these models, the proposal of Raudenbush (1994), which is based on the re-estimation of the heterogeneity variance after including predictors in the model, was detailed in Section 3.4. The fact that seven heterogeneity variance estimators are available leads to (at least) seven different ways for the calculation of the predictive power in meta-analytic models.

Given the amount of alternatives available to the meta-analyst when fitting mixed-effects meta-regression models, a first broad objective in this dissertation was to analyze the extent to which they can lead to different results, in order to determine which ones are preferred under a given scenario. With this aim, three simulation studies were conducted, and each one of them accounted for a wide variety of conditions that can be regarded as realistic in Psychology and related fields. A second (general) objective in the present dissertation was to check whether there are conditions under which the method choice does not make any difference on the results. On the one hand, no method was expected to perform appropriately for the most adverse scenarios. On the other hand, all methods were expected to converge (and to provide accurate results) for the optimal conditions.

The first simulation study, presented in Chapter 4, found some differences among the performance of the different heterogeneity variance estimators, which showed similar trends for both random- and mixed-effects models. On the one hand, the Hunter-Schmidt (HS), maximum likelihood (ML), and Sidik-Jonkman (SJ) methods provided negatively

biased estimates of the heterogeneity variances, while the Hedges (HE) method was unbiased but less efficient than the remaining estimators. On the other hand, the DerSimonian-Laird (DL), restricted maximum likelihood (REML), and empirical Bayes (EB) estimators showed better results, although a negative bias was found with the former for large parameter values. It seems, then, that REML and EB estimators constitute suitable options for the estimation of the heterogeneity variance parameters in meta-analytic models. The number of studies exerted a big influence on the results, and no method performed accurately with less than 20 studies. Conversely, precise estimates were obtained with 80 studies regardless of the method employed and the remaining manipulated factors.

An additional goal of the study presented in Chapter 4 was to analyze how the different methods perform for the estimation of the predictive power in mixed-effects meta-regression models, using the method proposed by Raudenbush (1994). Again, the HS, ML, SJ, and HE methods did not provide satisfactory results, and the best performance was exhibited by the DL, REML, and EB methods, with the latter showing better properties when examining the bias, truncation rates, and efficiency criteria jointly. The number of studies also exterted the greatest influence on the accuracy of all methods, and at least 40 studies were required to obtain precise estimates.

The second simulation study, described in Chapter 5, compared the performance of different methods to test for moderators in mixed-effects meta-regression models. The heterogeneity variance estimator did not show any influence in this case, but some discrepancies were observed depending on the method implemented to test the statistical significance of the regression coefficients. In previous works, it has been argued that the standard, Wald-type method for testing the coefficients in these models does not account for the fact that the variances need to be estimated in meta-analysis, leading to suboptimal results (e.g., Hardy & Thompson, 1996; Henmi & Copas, 2010). When

examining its performance in this study, the standard method did not adequately control the Type I error rate, leading to incorrect rejections of the null hypothesis.

Out of the different alternatives to the standard method examined in Chapter 5, the method proposed by Knapp and Hartung (2003) appeared as a suitable option, because of its simplicity and its appropriate empirical Type I error rates. It is worth noting, though, that this method showed a better performance without the truncation proposed by the authors, which led to a loss of statistical power. The Huber-White and likelihood ratio tests, which were also examined, did not show an appropriate control of the Type I error rate. Finally, the permutation test performed similarly to the untruncated Knapp-Hartung method. Although the latter should be preferred for most situations because of its simplicity, the permutation test constitutes a suitable option for scenarios where no random sampling of studies can be assumed (Manly, 1997). Nevertheless, about 40 studies were required for the different methods to achieve power rates around 0.80, as recommended by Jacob Cohen (1992).

The studies presented in Chapters 4 and 5 both focused on a normally distributed outcome, the standardized mean difference. Conversely, the last simulation study, presented in Chapter 6, explored some alternative outcome variables in meta-analysis within the reliability generalization framework. In this study, several methods to estimate the model coefficients and to test for moderators were compared. Regarding the outcome variables, coefficient alpha, which has an asymmetric sampling distribution, was employed together with three normalizing transformations. The results only showed slight discrepancies among the different outcome variables. Regarding the statistical methods to test for moderators, the trends were similar to the ones described in Chapter 5, with the residual heterogeneity variance estimators providing almost identical results and the Knapp-Hartung method outperforming the standard, Wald-type test both in terms of empirical Type I error and statistical power rates. Again, more than 30 studies were necessary before the methods reached satisfactory power rates.

If the findings from the three simulation studies are interpreted jointly, then the following conclusions are applicable to mixed-effects meta-regression models:

1. The heterogeneity variance estimator will not exert an important influence on the analyses when testing the significance of the model coefficients.

2. The heterogeneity variance estimator will have an influence on the results when estimating the predictive power of the model with the procedure proposed by Raudenbush (1994), and the REML, DL, and (especially) the EB estimators are expected to provide the most accurate results.

3. The method for testing the model coefficients will have an influence on the results, with the (untruncated) Knapp-Hartung method providing the most accurate results for most situations. If no random sampling of studies can be assumed, then a suitable option is to compute a permutation test.

4. About 40 studies are required to get accurate results in these models. With a smaller number of studies, the results should be interpreted cautiously.

5. Conclusions 1, 3, and 4 also hold when dealing with outcome variables that are not normally distributed (e.g., when integrating untransformed coefficients alpha).

Finally, some limitations of this dissertation, which also constitute perspectives for future research in the context of mixed-effects meta-regression models, must be remarked:

- The conclusions of this dissertation are restricted to the conditions manipulated along the three simulation studies here presented. Indeed, very interesting findings can be expected by carrying out additional simulation studies with some other conditions not included in this dissertation.

- The simulation study here presented that compared the performance of different heterogeneity variance estimators, as well as the previous studies mentioned in Chapter 4, was conducted on the basis that the normality assumption of the parameter effect sizes distribution is met. However, it does make sense to suspect that this critical assumption might not be satisfied in some practical situations, and it should be interesting for future studies to analyze how this circumstance affects the performance of the different methods.

- This dissertation includes the first systematic study of the Raudenbush's (1994) proposal for the estimation of the predictive power in mixed-effects meta-regression models, but it only accounted for normally distributed outcomes, namely the standardized mean difference. It should be interesting to analyze how the employement of outcome variables with asymmetric distribution, such as coefficient alpha, might affect the results and modify the patterns here reported.

- The procedure proposed by Raudenbush (1994) is considered as an appropriate way to compute the predictive power in meta-analytic regression models, but some other alternatives to accomplish this objective might be explored as well in the future.

- Some alternative methods for testing the model coefficients have been recently proposed (Friedrich & Knapp, 2011, August; Guolo, 2012; Huizenga et al., 2011), and it should be interesting to check whether these methods can improve the results yielded by the (untruncated) Knapp-Hartung and permutation tests.

References[8]

** Addington, D., Addington, J., Maticka-Tyndale, E., & Joyce, J. (1992). Reliability and validity of a depression rating scale for schizophrenics. *Schizophrenia Research, 6,* 201-208.

Aguayo, R., Vargas, C., de la Fuente, E. I., & Lozano, L. M. (2011). A meta-analytic reliability generalization study of the Maslach Burnout Inventory. *International Journal of Clinical and Health Psychology, 11,* 343-361.

Aguinis, H., Gottfredson, R. K., & Wright, T. A. (2011). Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior, 32,* 1033-1043.

Aguinis, H., & Pierce, C. A. (1998). Testing moderator variable hypotheses meta-analytically. *Journal of Management, 24,* 577-592.

Ahn, S., Myers, N. D., & Jin, Y. (2012). Use of the estimated intraclass correlation for correcting differences in effect size by level. *Behavior Research Methods, 44,* 490-502.

[8] References marked with one asterisk denote studies whose data were employed in the example from Chapter 5, while data from studies marked with two asterisks were included in the example shown in Chapter 6.

** Akdemir, A., Türkçapar, M. H., Örsel, S. D., Demirergi, N., Dag, I., & Özbay, M. H. (2001). Reliability and validity of the Turkish Version of the Hamilton Depression Rating Scale. *Comprehensive Psychiatry, 42,* 161-165.

Aloe, A. M., Becker, B. J., & Pigott, T. D. (2010). An alternative to R^2 for assessing linear models of effect size. *Research Synthesis Methods, 1,* 272-283.

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Anzures-Cabrera, J., & Higgins, J. P. T. (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods, 1,* 66-80.

Baker, W. L., White, C. M., Cappelleri, J. C., Kluger, J., & Coleman, C. I. (2009). Understanding heterogeneity in meta-analysis: The role of meta-regression. *International Journal of Clinical Practice, 63,* 1426-1434.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173-1182.

Bates, D. (2011, March). *Mixed models in R using the lme4 package. Part 3: Inference based on profile deviance*. Paper presented at the 8th International Amsterdam Conference on Multilevel Analysis, Amsterdam (The Netherlands).

Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499-525). San Diego, CA: Academic Press.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50,* 1088-1101.

** Bent-Hamsen, J., Lunde, M., Klysner, R., Andersen, M., Tanghøj, P., Solstad, K., & Bech, P. (2003). The validity of the Depression Rating Scale in discriminating between citalopram and placebo in depression recurrence in the maintenance therapy of elderly unipolar patients with major depression. *Pharmacopsychiatry, 36,* 313-316.

Beretvas, S. N., & Pastor, D. A. (2003). Using mixed-effects models in reliability generalization studies. *Educational and Psychological Measurement, 63,* 75-95.

Beretvas, S. N., Suizzo, M.A., Durham, J. A., & Yarnell, L. M. (2008). A reliability generalization study of scores on Rotter's and Nowicki-Strickland's Locus of Control Scales. *Educational and Psychological Measurement, 68,* 97-119.

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.

Biggerstaff, B. J., & Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine, 16,* 753-768.

** Bobes, J., Bulbena, A., Luque, A., Dal-Ré, R., Ballesteros, J., Ibarra, N., & Grupo de Validación en Español de Escalas Psicométricas (GVEEP) (2003). Evaluación psicométrica comparativa de las versiones en español de 6, 17 y 21 ítems de la Escala de valoración de Hamilton para la evaluación de la depresión [Comparative Psychometric evaluation of the 6-, 17-, and 21-item Spanish versions of the Hamilton Rating Scale for Depression]. *Medicina Clínica, 120,* 693-700.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27,* 335-340.

Bonett, D. G. (2003). Sample size requirements for comparing two alpha coefficients. *Applied Psychological Measurement, 27,* 72-74.

Bonett, D. G. (2008). Meta-analytic interval estimation for bivariate correlations. *Psychological Methods, 13,* 173-181.

Bonett, D. G. (2009). Meta-analytic interval estimation for standardized and unstandardized mean differences. *Psychological Methods, 14,* 225-238.

Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods, 15,* 368-385.

Borenstein, M. J. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 221-235). New York: Russell Sage Foundation.

Borenstein, M. J., Hedges, L. V., Higgins, J. C., & Rothstein, H. (2005). *Comprehensive meta-analysis: A computer program for research synthesis* (Vers. 2.2). Englewood, NJ: Biostat, Inc.

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.

Borman, G. D., & Grigg, J. A. (2009). Visual and narrative interpretation. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 497-519). New York: Russell Sage Foundation.

Botella, J., & Gambara, H. (2002). *¿Qué es el meta-análisis? [Meta-analysis: what is it?]* Madrid: 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., & Ponte, G. (2011). Effects of the heterogeneity of the variances on reliability generalization: An example with the Beck Depression Inventory. *Psicothema, 23,* 516-522.

Botella, J., & Suero, M. (2012). Managing heterogeneity of variances in studies of internal consistency generalization. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 8,* 71-80.

Botella, J., Suero, M., & Gambara, H. (2010). Psychometric inferences from a meta-analysis of reliability and internal consistency coefficients. *Psychological Methods, 15,* 386-397.

Botella, J., Suero, M., & Huang, H. (2012, July). *Meta-analysis of ROC curves when the reference has imperfect accuracy*. Paper presented at the V European Congress of Methodology, Santiago de Compostela (Spain).

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.

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.

Chalmers, I., Hedges, L. V., & Cooper, H. (2002). A brief history of research synthesis. *Evaluation & the Health Professions, 25,* 12-37.

Chappell, F. M., Raab, G. M., & Wardlaw, J. M. (2009). When are summary ROC curves appropriate for diagnostic meta-analyses? *Statistics in Medicine, 28,* 2653-2668.

Churchill, G. A., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research, 21,* 360-375.

Clarke, M. (2009). Reporting format. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 521-534). New York: Russell Sage Foundation.

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: Eribaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*, 98-101.

Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*, 243-253.

Conway, J. M., Jako, R. A., & Goodman, D. F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565-579.

Cooper, H. M. (1998). *Integrating research: A guide for literature reviews* (2nd ed.). Thousand Oaks, CA: Sage.

Cooper, H. (2007). *Evaluating and interpreting research syntheses in adult learning and literacy*. Boston, MA: National College Transition Network, New England Literacy Resource Center/World Education, Inc.

Cooper, H., & Hedges, L. V. (2009a). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 3-16). New York: Russell Sage Foundation.

Cooper, H., & Hedges, L. V. (2009b). Potentials and limitations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 561-572). New York: Russell Sage Foundation.

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.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14,* 165-176.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, & Winston.

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology, 34*, 481-489.

DerSimonian, R., & Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials, 28*, 105-114.

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

Duval, S. J., & Tweedie, R. L. (2000a). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455-463.

Duval, S. J., & Tweedie, R. L (2000b). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95,* 89-98.

Edgington, E. S. (1966). Statistical inference and nonrandom samples. *Psychological Bulletin, 66,* 485-487.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin, 136,* 103-127.

Enders, C. K. (2004). The impact of missing data on sample reliability estimates: Implications for reliability reporting practices. *Educational and Psychological Measurement, 64,* 419-436.

Erez, A., Bloom, M. C., & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology, 49*, 275-306.

*Fals-Stewart, W., Marks, A. P., & Schafer, J. (1993). A comparison of behavioral group therapy and individual behavior therapy in treating obsessive-compulsive disorder. *Journal of Nervous and Mental Disease, 181,* 189-193.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34,* 363-373.

Feldt, L. S., & Charter, R. A. (2006). Averaging internal consistency reliability coefficients. *Educational and Psychological Measurement, 66*, 215-227.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17,* 120-128.

Field, A. P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics, 2,* 105-124.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods, 10,* 444-467.

*Fineberg, N., Hughes, A., Gale, T., & Roberts, A. (2005). Group cognitive behaviour therapy in obsessive-compulsive disorder (OCD): A controlled study. *International Journal of Psychiatry in Clinical Practice, 9,* 257-263.

Fleiss, J. L, & Berlin, J. A. (2009). Effect sizes for dichotomous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 237-253). New York: Russell Sage Foundation.

Follmann, D., & Proschan, M. (1999). Valid inference in random effects meta-analysis. *Biometrics, 55*, 732–737.

* Freeston, M. H., Ladouceur, R., Gagnon, F., Thibodeau, N., Rhéaume, J., Letarte, H., & Bujold, A. (1997). Cognitive-behavioral treatment of obsessive thoughts: A controlled study. *Journal of Consulting and Clinical Psychology, 65,* 405-413.

Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods, Instruments, & Computers, 30*, 527-535.

Friedrich, T., & Knapp, G. (2011, August). *Generalised confidence intervals for meta regression.* Paper presented at the 58th World Statistics Congress (ISI/2011), Dublin (Ireland).

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

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

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in Social Research*. Beverly Hills, CA: Sage.

Glesser, L. J., & Olkin, I. (2009). Stocastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 357-376). New York: Russell Sage Foundation.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics, 1*, 223-231.

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement, 66,* 930-944.

* Greist, J. H., Marks, I. M., Baer, L., Kobak, K. A., Wenzel, K. W., Hirsch, M. J., Mantle, J. M., & Clary, C. M. (2002). Behavior therapy for obsessive-compulsive disorder guided by a computer or by a clinician compared with relaxation as a control. *Journal of Clinical Psychiatry, 63,* 138-145.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.) New York: Taylor & Francis Group, LCC.

Gronlund, N. E., & Linn, R. L. (1990). *Measurement and evaluation in teaching* (6th ed.). New York: Macmillan.

Gulliksen, H. (1987). *Theory of mental tests*. New York: Wiley.

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

Hakstian, A. R., & Whalen, T. E. (1976). A *k*-sample significance test for independent alpha coefficients. *Psychometrika, 41,* 219–231.

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.

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine, 17,* 841-856.

Hartung, J., Knapp, G., & Sinha, B. K. (2008). *Statistical meta-analysis with applications*. New Jersey: John Wiley & Sons.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72,* 320-338.

Harwell, M. R. (1992). Summarizing Monte Carlo results in methodological research. *Journal of Educational Statistics, 17,* 297-313.

Hedges, L. V. (1982). Fitting continuous models to effect size data. *Journal of Educational Statistics, 7,* 245-270.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93,* 388-395.

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics, 17,* 279-296.

Hedges, L. V. (2007). Meta-analysis. In C. R. Rao & S. Sinharay (Eds.), *The handbook of statistics,* (pp. 919-953). Amsterdam: Elsevier.

Hedges, L. V. (2009). Statistical considerations. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 37-47). New York: Russell Sage Foundation.

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics, 36,* 346-380.

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

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203-217.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9*, 426-445.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39-65.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*, 299-332.

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.

Henson, R. K., & Thompson, B. (2002). Characterizing measurement error in scores across studies: Some recommendations for conducting "reliability generalization" studies. *Measurement and Evaluation in Counseling and Development, 35*, 113-126.

Herbison, P., Hay-Smith, J., & Gillespie, W. J. (2006). Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology, 59*, 1249-1256.

Higgins, J. P. T., & Green, S. (Eds.) (2008). *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: Wiley.

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

Higgins, J. P. T., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine, 23*, 1663-1682.

Howell, R. T., & Shields, A. L. (2008). The file drawer problem in reliability generalization. A strategy to compute a fail-safe N with reliability coefficients. *Educational and Psychological Measurement, 68*, 120-128.

Hox, J. J., & de Leeuw, E. D. (2003). Multilevel models for meta-analysis. In S. P. Reise & N. Duan (Eds.), *Methodological advances, issues, and applications* (pp. 90-111). Mahwah, NJ: Lawrence Eribaum Associates.

Huber, P. (1967). The behavior of maximum-likelihood estimates under nonstandard conditions. In L. M. LeCam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 221-233). Berkeley: University of California Press.

Huedo-Medina, T., 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.

Hunt, M. (1997). *How science takes stock*. New York: Russell Sage Foundation.

Hunter, J. E., & Schmidt, F. L. (1977). A critical analysis of the statistical and ethical implications of various definitions of test fairness. *Psychological Bulletin, 83,* 1053-1071.

Hunter, J. E., & Schmidt, F. L. (1978). Differential and single group validity for employment tests by race: A critical analysis of three recent studies. *Journal of Applied Psychology, 63*, 1-11.

Hunter, J. E., & Schmidt, F. L. (1983). Quantifying the effects of psychological interventions on employee job performance and work force productivity. *American Psychologist, 38*, 473-478.

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment, 8,* 275-292.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting errors and bias in research findings* (2nd ed.). Newbury Park, CA: Sage.

Jackson, D., Riley, R., & White, I.R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine, 30,* 2481-2498.

Jennrich, R. I., & Sampson, P. F. (1976). Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics, 18,* 11-17.

* Jones, M. K., & Menzies, R. G. (1998). Danger ideation reduction therapy (DIRT) for obsessive-compulsive washers: A controlled trial. *Behaviour Research and Therapy, 36,* 959-970.

Jüni, P., Altman, D. G., & Egger, M. (2001). Assessing the quality of randomised controlled trials. In M. Egger et al. (Eds.), *Systematic reviews in health care* (pp. 87-108). London: BMJ Pub. Group.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological methods, 17,* 137-152.

Kieffer, K. M., & Reese, R. J. (2002). A reliability generalization study of the Geriatric Depression Scale. *Educational and Psychological Measurement, 62,* 969-994.

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

Knapp, G., Biggerstaff, B. J., & Hartung, J. (2006). Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal, 48,* 271-285.

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

Knapp, T. R., & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education, 70,* 65-79.

** Kobak, K. A., & Reynolds, W. M. (2000). The Hamilton Depression Inventory. In M. E. Maruish (Ed.), *Handbook of psychological assessment in primary care settings* (pp. 423-461). Mahwah, NJ: Lawrence Eribaum Associates.

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods, 2,* 61-76.

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.

Kraemer, H. C., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23-31.

Laird, N. M., & Mosteller, F. (1990). Some statistical methods for combining experimental results. *International Journal of Technology Assessment in Health Care, 6,* 5-30.

Lau, J., Ioannidis, J. P. A., & Schmid, C. H. (1998). Summing up evidence: One answer is not always enough. *Lancet, 351,* 123-127.

Leach, L. F., Henson, R. K., Odom, L. R., & Cagle, L. S. (2006). A reliability generalization study of the Self-Description Questionnaire. *Educational and Psychological Measurement, 66,* 285-304.

** Leidy, N. K., Palmer, C., Murray, M., Robb, J., & Revicki, D. A. (1998). Health-related quality of life assessment in euthymic and depressed patients with bipolar disorder. *Journal of Affective Disorders, 48,* 207-214.

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.

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

* Lindsay, M., Crino, R., & Andrews, G. (1997). Controlled trial of exposure and response prevention in obsessive-compulsive disorder. *British Journal of Psychiatry, 171,* 135-139.

Lipsey, M. W. (2009). Identifying interesting variables and analysis opportunities. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 147-158). New York: Russell Sage Foundation.

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

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. New York, NY: Oxford University Press.

López-López, J. A. (2008). *El enfoque meta-analítico de generalización de la fiabilidad: Una aplicación al Inventario de Obsesiones y Compulsiones de Maudsley [The reliability generalization meta-analytic approach: An application to the Maudsley Obsessive-Compulsive Inventory]*. Unpublished manuscript, University of Murcia.

López-Pina, J. A., Sánchez-Meca, J., & López-López, J. A. (2012). Métodos para promediar coeficientes alfa en los estudios de generalización de la fiabilidad [Methods for averaging alpha coefficients in reliability generalization studies]. *Psicothema, 24,* 161-166.

López-Pina, J.A., Sánchez-Meca, J., & Núñez-Núñez, R. M. (2011, July). *Validez y fiabilidad de una escala para la evaluación de la calidad de estudios primarios en meta-análisis [Validity and reliability of a scale for the evaluation of quality of primary studies in meta-analysis]*. Paper presented at the XII Congreso de Metodología de las Ciencias Sociales y de la Salud, San Sebastián (Spain).

López-Pina, J. A., Sánchez-Meca, J., & Rosa-Alcázar, A. I. (2009). The Hamilton Rating Scale for Depression: A meta-analytic reliability generalization study. *International Journal of Clinical and Health Psychology, 9*, 143-159.

* Lowell, K., Marks, I. M., Noshirvani, H., & O'Sullivan, G. (1994). Should treatment distinguish anxiogenic from anxiolytic obsessive-compulsive rumiations? *Psychotherapy and Psychosomatics, 61,* 150-155.

Malzahn, U., Böhning, D., & Holling, H. (2000). Nonparametric estimation of heterogeneity variance for the standardised difference used in meta-analysis. *Biometrika, 87,* 619-632.

Makambi, K. H. (2004). The effect of the heterogeneity variance estimator on some tests of treatment efficacy. *Journal of Biopharmaceutical Statistics, 14,* 439-449.

Manly, B. F. J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology*. London: Chapman & Hall.

Marín-Martínez, F. (1996). *Enfoques meta-analíticos: Un estudio comparativo mediante simulación Monte Carlo [Meta-analytic approaches: A comparative study through Monte Carlo simulation]*. Unpublished Dissertation, University of Murcia.

Marín-Martínez, F., & Sánchez-Meca, J. (1999). Averaging dependent effect sizes in meta-analysis: A cautionary note about procedures. *The Spanish Journal of Psychology, 2,* 32-38.

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.

Marín-Martínez, F., Sánchez-Meca, J., & López-López, J. A. (2009). El meta-análisis en el ámbito de las Ciencias de la Salud: Una metodología imprescindible para la eficiente acumulación del conocimiento [Meta-analysis in the Health Sciences: An indispensable methodology for the efficient accumulation of knowledge]. *Fisioterapia, 31,* 107-114.

* Marks, I. M., Stern, D. M., Cobb, J., & McDonald, R. (1980). Clomipramine and exposure for obsessive-compulsive rituals: I. *British Journal of Psychiatry, 136,* 1-25.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software. 42(9):* 1-21.

Mason, C., Allam, R., & Brannick, M. T. (2007). How to meta-analyze coefficient-of-stability estimates. *Educational and Psychological Measurement, 67,* 765-783.

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 537-560). New York: Russell Sage Foundation.

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 Med 6(6):* e1000097.

Morales, L. A., Garrido, V., & Sánchez-Meca, J. (2010). Treatment effectiveness in secure corrections of serious (violent or chronic) juvenile offenders. Stockholm, Sweden: Edita Norstedts Västerås.

Moreno, S. G., Sutton, A. J., Thompson, J. R., Ades, A. E., Abrams, K. R., & Cooper, N. J. (2012). A generalized weighting regression-derived meta-analysis estimator robust to small study-effects and heterogeneity. *Statistics in Medicine, 31,* 1407-1417.

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.

Muncer, S. J., Craigie, M., & Holmes, J. (2003). Meta-analysis and power: Some suggestions for the use of power in research synthesis. *Understanding Statistics, 2,* 1-12.

* Nakatani, E., Nakagawa, A., Nakao, T., Yoshizato, C., Nabeyama, M., Kudo, A., Isomura, K., Kato, N., Yoshioka, K., & Kawamoto, M. (2005). A randomized controlled trial of Japanese patients with obsessive-compulsive disorder: Effectiveness of behavior therapy and fluvoxamine. *Psychotherapy and Psychosomatics, 74,* 269-276.

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

Normand, S. L. T. (1999). Tutorial in biostatistics. Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine, 18,* 321-359.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 177-203). New York: Russell Sage Foundation.

Overton, R. C. (1998). A comparison of fixed-effects and mixed (random-effects) models for meta-analysis tests of moderator variable effects. *Psychological Methods, 3,* 354–379.

* O'Connor, K., Todorov, C., Robillard, S., Borgeat, F., & Brault, M. (1999). Cognitive-behaviour therapy and medication in the treatment of obsessive-compulsive disorder: A controlled study. *Canadian Journal of Psychiatry, 44,* 64-71.

Parker, K. (1983). A meta-analysis of the reliability and validity of the Rorschach. *Journal of Personality Assessment, 47,* 227-231.

Parker, K. C. H., Hanson, R. K., & Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin, 103,* 367-373.

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

Pearson, K. (1904). Report on certain enteric fever inoculation statistics. British *Medical Journal, 3,* 1243–1246.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: an integrated analysis.* Hillsdale, NJ: Eribaum.

Pereira, T. V., Patsopoulos, N. A., Salanti, G., & Ioannidis, J. P. A. (2010). Critical interpretation of Cochran's Q test depends on power and prior assumptions about heterogeneity. *Research Synthesis Methods, 1,* 149-161.

Peter, J. P., & Churchill, G. A. (1986). Relationships among research design choices and psychometric properties of rating scales: A meta-analysis. *Journal of Marketing Research, 23,* 1-10.

Pigott, T. D. (2001). Missing predictors in models of effect size. *Evaluation and the Health Professions, 24,* 277-307.

** Ramos, J. A., & Cordero, A. (1988). A new validation of the Hamilton Rating Scale for depression. *Journal of Psychiatry Research, 22,* 21-28.

** Rapp, S. R., Smith, S. S., & Britt, M. (1990). Identifying comorbid depression in elderly medical patients: Use of extracted Hamilton Depression Rating Scale. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 243-247.

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.

Ray, J. W., & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? *Journal of Consulting and Clinical Psychology, 64,* 1316-1325.

Reed, J. G., & Baxter, P. M. (2009). Using reference databases. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 73-101). New York: Russell Sage Foundation.

Review Manager (2011). *RevMan* [Computer program]. Version 5.1. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration.

** Reynolds, W. M., Mazza, J. J. (1998). Reliability and validity of the Reynolds Adolescent Depression Scale with young adolescents. *Journal of School Psychology, 36,* 295-312.

** Riskind, J. H., Beck, A. T., Brown, G., & Steer, R. A. (1987). Taking the measure of anxiety and depression. *The Journal of Nervous and Mental Disease, 175,* 474-479.

Robinson, D. H., Whittaker, T. A., Williams, N. J., & Beretvas, S. N. (2003). It's not effect sizes so much as comments about their magnitude that mislead readers. *Journal of Experimental Education, 72,* 51-64.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 10,* 75-98.

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306-322.

Romano, J. L, & Kromrey, J. D. (2009). What are the consequences if the assumption of independent observations is violated in reliability generalization meta-analysis studies? *Educational and Psychological Measurement, 69,* 404-428.

Romano, J. L, Kromrey, J. D., & Hibbard, S. T. (2010). A Monte Carlo study of eight confidence interval methods for coefficient alpha. *Educational and Psychological Measurement, 70,* 376-393.

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

Rosenberg, M. S., Adams, D. C., & Gurevitch, J. (1999). *MetaWin: Statistical software for meta-analysis with resampling tests* (Vers. 2.0). Sunderland, MA: Sinauer Associates.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 231-244). New York: Russell Sage Foundation.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin, 118*, 183-192.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59-82.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99,* 400-406.

Rosnow, R. L, & Rosenthal, R. (2009). Effect sizes: Why, when, and how to use them. *Zeitschrift für Psychologie / Journal of Psychology, 217*, 6-14.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds) (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, UK: Wiley.

** Rush, A. J., Giles, D. E., Schlesser, M. A., Fulton, C. L., Weissenburger, J., & Burns, C. (1986). The Inventory for Depressive Symptomatology (IDS): Preliminary findings. *Psychiatry Research, 18,* 65-87.

** Rush, A. J., Triverdi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., Markowitz, J. C., Ninan, P. T., Kornstein, S., Manber, R., Thase, M. E., Kocsis, J. H., & Keller, M. B. (2003). The 16-item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self Report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry, 54,* 573-583.

Sagie, A., & Koslowsky, M. (1993). Detecting moderators with meta-analysis: An evaluation and comparison of techniques. *Personnel Psychology, 46,* 629-640.

Salgado, J. F., & Moscoso, S. (1996). Meta-analysis of interrater reliability of job performance ratings in validity studies of personnel selection. *Perceptual and Motor Skills, 83,* 1195-1201.

Sánchez-Meca, J. (1986). La revisión cuantitativa: Una alternativa a las revisiones tradicionales [The quantitative review: An alternative to traditional reviews]. *Anales de Psicología, 3,* 79-107.

Sánchez-Meca, J., & Ato-García, M. (1989). Meta-análisis: Una alternativa metodológica a las revisiones tradicionales de la investigación [Meta-analysis: A methodological alternative to traditional research reviews]. In J. Arnau & H. Carpintero (Eds.), *Tratado de Psicología General* (pp. 617-669). Madrid: Alhambra.

Sánchez-Meca, J., Boruch, R. F., Petrosino, A., & Rosa-Alcázar, A. I. (2002). La Colaboración Campbell y la práctica basada en la evidencia [The Campbell Collaboration and the evidence-based practice]. *Papeles del Psicólogo, 83,* 44-48.

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. (in press). Some recommended statistical analytic practices when reliability generalization (RG) studies are conducted. *British Journal of Mathematical and Statistical Psychology*.

Sánchez-Meca, J., & López-Pina, J. A. (2008). El enfoque meta-analítico de generalización de la fiabilidad [The reliability generalization meta-analytic approach]. *Acción Psicológica, 1-2*, 37-64.

Sánchez-Meca, J., López-Pina, J. A., & López-López, J. A. (2008). Una revisión de los estudios meta-analíticos de generalización de la fiabilidad [A review of the reliability generalization meta-analytic studies]. *Escritos de Psicología, 1-2*, 107-118.

Sánchez-Meca, J., López-Pina, J. A., & López-López, J. A. (2009). Generalización de la fiabilidad: Un enfoque meta-analítico aplicado a la fiabilidad [Reliability generalization: A reliability-applied meta-analytic approach]. *Fisioterapia, 31*, 262-270.

Sánchez-Meca, J., López-Pina, J. A., López-López, J. A., Marín-Martínez, F., Rosa-Alcázar, A. I., & Gómez-Conesa, A. (2011). The Maudsley Obsessive-Compulsive Inventory: A reliability generalization meta-analysis. *International Journal of Clinical and Health Psychology, 11*, 473-493.

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. (1998a). Testing continuous moderators in meta-analysis: A comparison of procedures. *British Journal of Mathematical and Statistical Psychology, 51*, 311-326.

Sánchez-Meca, J., & Marín-Martínez, F. (1998b). Weighting by inverse variance or by sample size in meta-analysis: A simulation study. *Educational and Psychological Measurement, 58*, 211-220.

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., volume 7, pp. 274-282). Oxford: 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., Marín-Martínez, F., & López-López, J. A. (2011). Meta-análisis e intervención psicosocial basada en la evidencia [Meta-analysis and evidence-based psychosocial intervention]. *Psychosocial Intervention, 20,* 95-107.

Sánchez-Meca, J., Marín-Martínez, F., & López-López, J. A. (2012, July). *Improving inferential methods: Contributions from meta-analysis*. Paper presented at the V European Congress of Methodology, Santiago de Compostela (Spain).

Sánchez-Meca, J., Marín-Martínez, F., & López-López, J. A. (in press). Metodología del meta-análisis [Methodology of meta-analysis]. In F. J. Sarabia (Ed.), *Metodología para la investigación en marketing y dirección de empresa [Methodology for research on marketing and business management]* (2nd ed.). Madrid, Spain: Pirámide.

Schmidt, F. L. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science, 5,* 233-242.

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.

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods, 5,* 230-240.

Shadish, W. R., Chacón-Moscoso, S., & Sánchez-Meca, J. (2005). Evidence-based decision making: Enhancing systematic reviews of program evaluation results in Europe. *Evaluation, 11,* 95-109.

Shadish, W. S., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 257-277). New York: Russell Sage Foundation.

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., Porter, A. C., Tugwell, P., Moher, D., & Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BioMed Central, 7(10).*

Shuster, J. J. (2010). Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine, 29,* 1259-1265.

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

Sidik, K., & Jonkman, J. N. (2005b). Simple heterogeneity variance estimation for meta-analysis. *Applied Statistics, 54,* 367-384.

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

Skrondal, A. (2000). Design and analysis of Monte Carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research, 35,* 137-167.

Slavin, R., & Smith, D. (2009). The relationship between sample sizes and effect sizes in systematic reviews in education. *Educational Evaluation and Policy Analysis, 31,* 500-506.

Sutton, A. J., Duval, S. J., Tweedie, R. L., Abrams, K. R., & Jones, D. R. (2000). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal, 320,* 1574-1577.

** Stage, K. B., Middelboe, T., & Pisinger, C. (2003). Measurement of depression in patients with chronic obstructive pulmonary disease (COPD). *Nordic Journal of Psychiatry, 57,* 297-301.

Stevens, J. R., & Taylor, A. M. (2009). Hierarchical dependence in meta-analysis. *Journal of Educational and Behavioral Statistics, 34,* 46-73.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: Harvard University Press.

Streiner, D. L (2003) Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80,* 99-103.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 435-452). New York: Russell Sage Foundation.

Sutton, A. J., & Higgins, J. P. T. (2008). Recent developments in meta-analysis. *Statistics in Medicine, 27,* 625-650.

Tatsioni, A., Zarin, D. A., Aronson, N., Samson, D. J., Flamm, C. R., Schmid, C., & Lau, J. (2005). Challenges in systematic reviews of diagnostic technologies. *Annals of Internal Medicine, 142,* 1048-1055.

Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal, 309,* 1351-1355.

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

Thompson, S. G., & Higgins, J. P. T. (2010). Comments on 'Empirical vs natural weighting in random effects meta-analysis'. *Statistics in Medicine, 29,* 1270-1271.

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine, 18,* 2693-2708.

** Thunedborg, K., Black, C. H., & Bech, P. (1995). Beyond the Hamilton Depression scores in long-term treatment of manic-melancholic patients: Prediction of recurrence of depression by quality of life measurements. *Psychotherapy and Psychosomatics, 64,* 131-140.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58,* 6-20.

Vacha-Haase, T., & Thompson, B. (2011). Score reliability: A retrospective look back at 12 years of reliability generalization studies. *Measurement and Evaluation in Counseling and Development, 44,* 159-168.

Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 129-146). New York: Russell Sage Foundation.

Valentine, J. C., Cooper, H., Patall, E. A., Tyson, D., & Robinson, J. C. (2010). A method for evaluating research syntheses: The quality, conclusions, and consensus of 12 syntheses of the effects of after-school programs. *Research Synthesis Methods, 1,* 20-38.

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35,* 215-247.

* Van Balkom, J. L. M., de Haan, E., van Oppen, P., Spinhoven, P., Hoogduin, K. A. L., & van Dyck, R. (1998). Cognitive and behavioral therapies alone versus in combination with fluvoxamine in the treatment of obsessive-compulsive disorder. *Journal of Nervous and Mental Disease, 186,* 492-499.

Van Den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (in press). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*.

Van den Noortgate, W., & Onghena, P. (2003). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement, 63,* 765-790.

Van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine, 21,* 589-624.

Victorson, D., Barocas, J., & Song, J. (2008). Reliability across studies from the functional assessment of cancer therapy-general (FACT-G) and its subscales: A reliability generalization. *Quality of Life Research, 17,* 1137-1146.

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. (2007a). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie / Journal of Psychology, 215,* 104-121.

Viechtbauer, W. (2007b). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine, 26,* 37-52.

Viechtbauer, W. (2007c). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology, 60,* 29-60.

Viechtbauer, W. (2008). Analysis of moderator effects in meta-analysis. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 471-487). Thousand Oaks, CA: Sage.

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. *A comparison of procedures to test for moderators in meta-regression models.* Unpublished manuscript.

* Vogel, P. A., Stiles, T. C., & Götesman, K. G. (2004). Adding cognitive therapy elements to exposure therapy for obsessive-compulsive disorder: A controlled study. *Behavioural and Cognitive Psychotherapy, 32,* 275-290.

Walter, S. D., & Jadad, A. R. (1999). Meta-analysis of screening data: A survey of the literature. *Statistics in Medicine, 18,* 3409-3424.

White, H. (1980). A heteroscedastity-consistent covariance matrix and a direct test for heteroscedasticity. *Econometrica, 48,* 817-838.

White, H. D. (2009). Scientific communication and literature retrieval. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 51-71). New York: Russell Sage Foundation.

Wilkinson, L., & APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Wilson, D. B. (2009). Systematic coding. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 159-176). New York: Russell Sage Foundation.

Yarnold, P. R., & Mueser, K. T. (1989). Meta-analyses of the reliability of Type A behaviour measures. *British Journal of Medical Psychology, 62*, 43-50.

Yin, P., & Fan, X. (2000). Assessing the reliability of Beck Depression Inventory scores: Reliability generalization across studies. *Educational and Psychological Measurement, 60*, 201-223.

Zangaro, G. A., & Soeken, K. L. (2005). Meta-analysis of the reliability and validity of Part B of the Index of Work Satisfaction across studies. *Journal of Nursing Measurement, 13*, 7-22.