universität wien

Diplomarbeit

Titel der Arbeit

Reanalysis of classic meta-analyses (1977-82) using state-of-the-art methods:

A systematic review and meta-analytical results

Verfasserin

Barbara Zimmermann

Angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer. nat.)

Wien, im November 2011

Studienkennzahl:	298
Studienrichtung:	Psychologie
Betreuer:	Assistenzprof. Privatdoz. MMag. DDDr. Martin Voracek

Danksagung

Vor allem möchte ich Herrn Assistenzprof. DDDr. Martin Voracek für seine gleichermaßen fordernd und fördernde Art der Betreuung dieser Diplomarbeit danken, welche die vorliegende Arbeit in der bestehenden Form ermöglichte.

Danken möchte ich auch Herrn Mag. Nader und Herrn Mag. Pietschnig für ihre Anregungen zu Powerberechnung und Ergebnisdarstellung.

Dank gilt auch meinen Studienkolleginnen und Freundinnen Johanna Kafka, Ramona Knapp, Kathrin König, Claudia Krutis und Heidi Lees für die bereichernden Diskussionen und ihre kritischen Anmerkungen zur vorliegenden Arbeit.

Besonderer Dank gilt nicht zuletzt meiner Familie und Stefan Silberfeld für ihre Unterstützung und Motivation während aller Phasen des Studiums.

Table of contents

Introduction

1 Introduction

Is it possible to reanalyze classic meta-analyses using state-of-the-art methods? Can modern methods be applied to the reported original data to answer the original research questions? Will there be new or different conclusions drawn from the data when using today's methods?

This work aspires to give answers to the proposed questions through a systematic review of relevant literature and meta-analysis. However, first of all an examination of the term *classic meta-analysis* is needed, because it has not yet been defined.

1.1 Definition of the term classic meta-analysis

First associated with the term meta-analysis is Gene V. Glass: It was him who gave the statistical procedure its name (Glass, 1976) and conducted and published the first so called meta-analysis with Mary L. Smith in 1977 (Hunt, 1997). In his renowned article, "Primary, Secondary, and Meta-Analysis of Research" (Glass, 1976), he gave a definition: "Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976, p. 3). Today, more than 30 years later, meta-analysis is still defined in the same manner: "Meta-analysis refers to the statistical synthesis of results from a series of studies" (Borenstein, Hedges, Higgins, & Rothstein, 2009, p. xxi). Accordingly, meta-analysis can be seen as a method to analyse and statistically integrate results of several studies.

In this work, the word *classic* refers to the period of time during which the first meta-analyses were published, starting in 1977 with Smith and Glass's meta-analysis on the effectiveness of psychotherapy. It was not only this team who developed methods to statistically integrate research findings around this time. In the field of social psychology, Rosenthal and Rubin integrated the results of 345 studies on interpersonal expectancy effects in 1978 and in the field of personnel psychology, Hunter, Schmidt, and Hunter synthesised studies on the validity of employment tests in 1979 (Cooper, 2010). The end of the period of classic meta-analyses is set at the year 1982 for the reasons depicted and

explained below. First, the development of publications concerning meta-analysis was considered. Therefore an exemplary literature search was executed in the electronic literature database PsycINFO. The search term *meta-analysis* was used in the title field and search was restricted to the time span 1977 to 1987 (retrieved July 4, 2011). Outcome of the depicted literature search is shown in Figure 1. Important for the present work is that the number of obtained results per year grows rapidly from 1982 onwards. The rapid growth in number of articles concerning meta-analysis from 1982 on is the first reason for setting the end of the period of classic meta-analyses at that point.

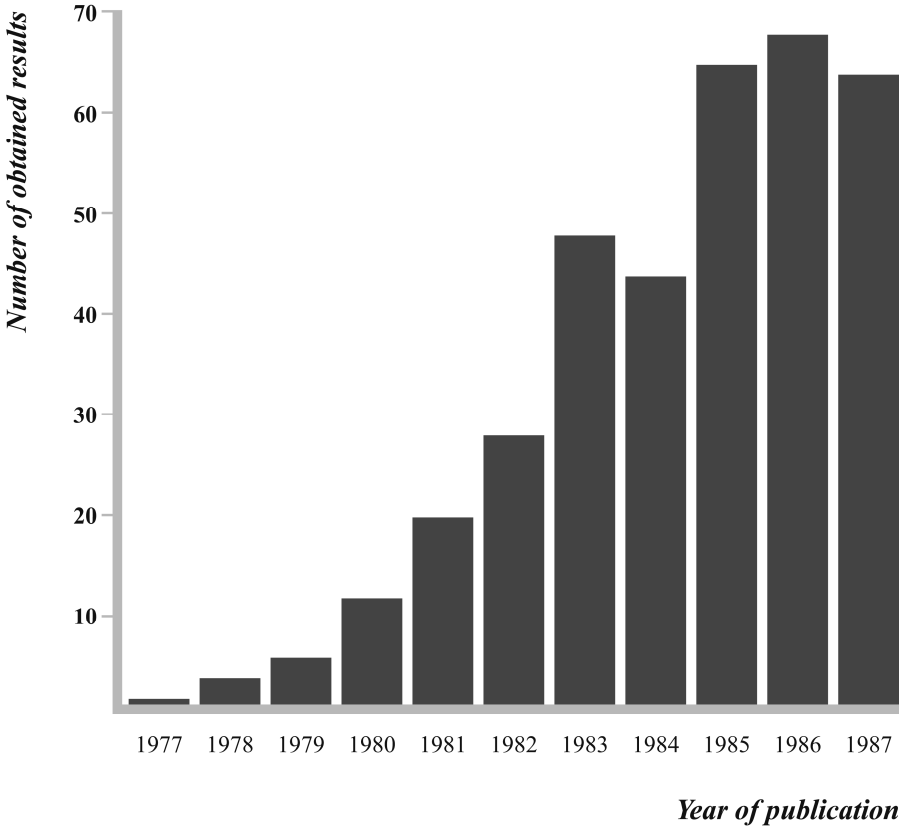


Figure 1. Results of an exemplary literature search in the electronic literature database PsycINFO with the search term meta-analysis for the years 1977 to 1987.

The second reason is an important shift in statistical theory in 1983, when Larry Hedges (1983) presented a random-effects model for the analysis of effect sizes. Statistical theory prior to Hedges (1983) assumed fixed population effect sizes. Put in other words, the fixed-effect model assumes that all studies have the same true effect size and therefore share one common effect size as summary effect, whereas in a random-effects model true effects vary between studies. The summary effect hence is an estimate of the mean of a distribution of true effects (Borenstein et al., 2009). The advancement in statistical theory

with the publication of Hedges' article in 1983 serves to justify setting an end to the period of classic meta-analyses at the year 1982. Furthermore, a characteristic of published articles of classic meta-analyses can be seen in the absence of forest plot and funnel plot. In a funnel plot the relationship between study size and effect size is shown graphically, it allows to check for publication bias visually (Borenstein et al., 2009), and acquired its name because it looks like a funnel (Light & Pillemer, 1984). A forest plot displays the results of a meta-analysis graphically. Each included original study is presented in a separate row showing its effect size and the corresponding confidence interval. The last row shows the summary effect (Petticrew & Roberts, 2006). Hedges and Olkin (1989) already plotted the effect sizes for each included study on separate horizontal lines along with the according confidence interval and can therefore be seen as a precursor of today's forest plot.

To summarise the argumentation above and give a compact definition: Classic meta-analyses are in this work defined as meta-analyses published between 1977 and 1982 using fixed-effect models and having no accompanying funnel plots or forest plots.

1.2 Previous reanalyses and replications of meta-analyses

After defining classic meta-analyses, which are supposed to be reanalyzed, thought should be given to reanalysis itself. In his article on data analysis at three levels, Glass (1978) described secondary analysis as "the re-analysis of data for the purpose of answering the original research question with better statistical techniques, or answering new questions with old data" (Glass, 1978, p. 3). Even though this definition refers to the reanalysis of original research studies and not meta-analyses, it perfectly describes one of the central intentions of the present work. The original data of classic meta-analyses reported in published articles is supposed to be reanalyzed using state-of-the-art methods to answer the original research questions.

Not exactly the same but related to reanalysis is the act of replication. A replication is the repetition of a research study to reassess its results (Bortz & Döring, 2006). Shapiro and Shapiro (1982), for example, conducted a replication of Smith and Glass's (1977) meta-analysis. Shapiro and Shapiro (1982) wanted to reassess the results and therefore conducted their own meta-analysis with altered conditions. An example for reanalysis is

Rosenthal and Rubin's (1982) work. Rosenthal and Rubin (1982) reanalyzed a meta-analysis on cognitive gender differences conducted by Hyde (1981), applying advanced statistical techniques to the original data and extending previously drawn conclusions. Accordingly, an important difference between reanalysis and replication is the basis of data used for analysis. For reanalysis existing old data is used as basis. In a replication study data will be collected anew and this new data is used as basis. The exact description of the two terms is important because replication studies as well as reanalyses will be taken into account to see whether reanalysis of a meta-analysis is at all possible and to get an overview of what has been done so far concerning reanalysis of classic meta-analyses. The reason to consider both is the similarity between reanalysis and replication.

In the following, two examples are given to demonstrate that reanalysis of meta-analysis is possible. Reanalysis in this context means that the same meta-analytical techniques to integrate results were applied to the same set of studies by independent scholars. Thus if correspondence existed between the different analyses, reanalysis is considered possible and reasonable.

The first example is a project in which four reputable scholars with good methodological skills independently conducted a meta-analysis based on a nearly common set of preselected studies (Schneider, 1990). The project depicted had been initiated by the National Institute of Education (NIE) to resolve the problem of inconsistent results concerning the effects of desegregation on academic achievement of African-American students. Seven scholars met and agreed on the use of common criteria for selecting studies to include in the meta-analysis, which resulted in a set of 19 studies to be reanalyzed (Schneider, 1990). Despite the intention of the sponsors, individual scholars in the course of work determined different subsets to be methodologically adequate (Press, 1990). One of the seven scholars was a methodological expert who commented on the results of the other scholars and two of the remaining six did a review of factors other than desegregation and critical discussion of the work of the other scholars respectively instead of a meta-analysis (Schneider, 1990). When the results of the four analyses were compared, high degree of correspondence could be declared. The small discrepancies between the results were due to differences in the studies included, the way of effect size calculation and different control groups used (Ingram, 1990). Despite the fact that the meta-analyses differed in methodological details, the obtained integrated results were notably similar (Cordray, 1990). The conclusion drawn from this project was that

reanalysis is possible and reasonable, as was shown by the general correspondence between the results of meta-analyses conducted independently by different scholars.

Cooper and Rosenthal (1980) had a different intention but applied a similar approach. Cooper and Rosenthal (1980) intended to compare conclusions drawn from an identical set of studies after using a traditional or statistical review procedure. For this purpose, graduate students and faculty members of a university were randomly assigned to one of the approaches. Statistical reviewers had to retrieve *p*-levels from the original studies and then calculate the overall probability using the unweighted Stouffer method for independent studies, for which they got instructions how to apply it. A total of 95% of the statistical reviewers retrieved the correct *p*-levels and 84% correctly combined these *p*-levels (Cooper & Rosenthal, 1980). Of interest for this work is the statistical condition only. Even if modern methods for statistically integrating research results (e.g., Borenstein et al., 2009) differ from the one applied in the described study, it is an important finding that a high correspondence between the results had been obtained. Again this serves to confirm the assumption that reanalysis of meta-analysis is possible and reasonable. Allen and Preiss (1993) even go so far as to say that there is a need to replicate any meta-analysis for confirmation and consolidation of findings. Furthermore, Allen and Preiss (1993) stated that it is important to replicate meta-analyses to prevent misleading results.

Further support for the assumption that reanalysis of meta-analysis is possible comes from Schmidt, Oh, and Hayes (2009), who reanalyzed meta-analyses using the same meta-analytical technique to integrate results as the original meta-analyses (fixed-effect model) and received results nearly identical to the results of the original meta-analyses. Furthermore, Schmidt et al. (2009) reanalyzed these meta-analyses using methods different to the ones applied by the original meta-analyses, namely random-effects models.

Schmidt et al. (2009) reanalyzed data from five large meta-analytic studies. All of them reported multiple meta-analyses and therefore encompassed 68 separate meta-analyses. The five studies were published in *Psychological Bulletin* between 1988 and 2006 and all employed a fixed-effect model. Schmidt et al. (2009) reanalyzed all of the 68 meta-analyses first applying the fixed-effect meta-analysis procedure and affirmed that the obtained results were nearly identical to the results of the original studies. Then the data sets of the 68 meta-analyses were reanalyzed applying two different random-effects meta-analysis procedures. The goal was to compare the results when a fixed-effect model was applied to data sets with the results when random-effects models were applied to the same data sets (Schmidt et al., 2009).

Schmidt et al. (2009) reanalyzed meta-analyses published between 1988 and 2006 using state-of-the-art methods (in this case random-effects models). The purpose of the present work is similar, except that classic meta-analyses are supposed to be reanalyzed using state-of-the-art methods.

Classic meta-analyses had already been reanalyzed, but in turn they had been reanalyzed using methods to integrate results that were modern in the period of 1977 to 1982. Landman and Dawes (1982) and Shapiro and Shapiro (1982), for example, were concentrated on replicating the first ever meta-analysis (Smith & Glass, 1977). The aim was to reassess the results using methods that were state-of-the-art in the time of classic meta-analyses. The two studies (Landman & Dawes, 1982; Shapiro & Shapiro, 1982) are described in more detail below, together with other examples of classic meta-analyses being reanalyzed/replicated using statistical methods of that time.

The meta-analysis conducted by Smith and Glass (1977) was the first so called meta-analysis and concerned the effectiveness of psychotherapy (Hunt, 1997). Included in the meta-analysis (Smith & Glass, 1977) were published literature as well as dissertations and fugitive literature. Studies included had to compare at least one therapy group to an untreated or different therapy group. The effect size was calculated by subtracting the mean of the control group from the mean of the treatment group and dividing the result by the standard deviation of the control group. In all, 833 effect sizes were calculated from 375 studies. There were more effect sizes than studies because some studies reported more than one effect size. The overall result was that psychotherapy is effective. An average advantage of 0.68 standard deviation units of the treated over the control group was obtained across studies. When behavioral and nonbehavioral therapies were compared no difference in effectiveness was found (Smith & Glass, 1977). Landman and Dawes (1982) replicated Smith and Glass's (1977) meta-analysis to find out whether the same results were obtained when only appropriately controlled studies, studies with random assignment of subjects to treatment or control group, were included. For this purpose they randomly selected 65 studies from the list of included studies of Smith and Glass's (1977) meta-analysis (which included 93 additional ones to the 375 original ones) excluding unpublished dissertations and books. Only studies judged to be appropriately controlled and therefore only studies with high quality, 42 out of the prior 65, were included for meta-analysis. Furthermore, Landman and Dawes (1982) were interested whether placebo effects contributed to the outcome, something Smith and Glass (1977) had not examined. For that purpose, all of the studies with placebo controls of the set of 42 studies were

statistically integrated as a group. Moreover, Landman and Dawes (1982) were interested whether statistical nonindependence of results had an influence on overall results. Two procedures were utilized to examine this issue. One of them was to compare results obtained when each effect size was the unit of analysis, as Smith and Glass (1977) did, to results obtained when each study was the unit of analysis. Calculation of effect sizes was the same as in Smith and Glass's (1977) meta-analysis, as were the applied meta-analytic procedures. Results obtained by Landman and Dawes (1982) supported Smith and Glass's (1977) results. The average effect size (0.78 *SD* units) for the 42 studies was similar and even larger than that of the original meta-analysis. A placebo effect was observed, which however was less than that of the treatment. Finally nonindependence of results had no influence on the overall result (Landman & Dawes, 1982). Shapiro and Shapiro (1982) also replicated Smith and Glass's (1977) meta-analysis. The purpose of the replication was to assess effectiveness of psychotherapy through meta-analysis with conditions altered based on criticism of Smith and Glass's (1977) work. Only studies comparing two or more treatment groups to a control group were included. Control groups had to be untreated or minimally treated. More behavioral studies were included compared to Smith and Glass (1977). Furthermore, dissertations were excluded and categories for characterizing outcome measurement were refined. A total of 143 studies were included and these studies had been published between 1974 and 1979. Effect sizes were computed as specified in Smith, Glass, and Miller (1980), in all 1828 effect sizes were calculated (Shapiro & Shapiro, 1982). Smith et al. (1980) is an expansion of Smith and Glass's (1977) work (Smith et al., 1980). Shapiro and Shapiro (1982) concluded from their meta-analysis, that psychotherapy was effective, which is consistent with the findings of Smith and Glass (1977). The obtained overall effect size (0.93 *SD* units) was even larger than that of the original meta-analysis. Moreover, behavioral and cognitive methods were found to be superior and dynamic and humanistic methods were found to be inferior (Shapiro & Shapiro, 1982).

Both studies, Landman and Dawes (1982) as well as Shapiro and Shapiro (1982), are replication studies and therefore data sets, on which the respective calculations were based, differ from the data set of the original meta-analysis. The same applies to the study of Eagly and Carli (1981), which is a meta-analysis concerning the same research question as the original meta-analysis, but was based on a different sample of included studies. Eagly and Carli (1981) had criticized the sample of studies of Cooper's (1979) meta-analysis of sex differences in influenceability and intended to conduct a meta-analysis on

this topic based on a broader sample of findings. An example for reanalysis of a classic meta-analysis using methods that had been state-of-the-art in that period of time (1977-1982), is the work of Rosenthal and Rubin (1982). Rosenthal and Rubin (1982) reanalyzed a meta-analysis on cognitive gender differences conducted by Hyde (1981), applying advanced statistical techniques to the original data and extending previously drawn conclusions.

1.3 Present research project and problems to be expected

As described above, on the one hand classic meta-analyses had been reanalyzed using statistical methods to integrate results that were modern in the period of 1977 to 1982. On the other hand, older but not classic meta-analyses had been reanalyzed using state-of-the-art methods. Reanalysis of classic meta-analyses using state-of-the-art methods with the intention of answering the original research questions has not yet been undertaken and is therefore attempted in this work. It is important to emphasise that in the present work, in contrast to replication studies, original data reported for classic meta-analyses are supposed to be reanalyzed. State-of-the-art methods in the field of meta-analysis are presented in Borenstein et al. (2009), Cooper (2010), Lipsey and Wilson (2001) and Petticrew and Roberts (2006), among others. Software used to execute calculations is Comprehensive Meta-Analysis, version 2.2.030 (Borenstein, Hedges, Higgins, & Rothstein, 2005).

Problems that could arise when carrying out a reanalysis of meta-analyses are for instance described by Schmidt et al. (2009). As described above, Schmidt et al. (2009) conducted a reanalysis of five meta-analyses using state-of-the-art methods. Besides being published in *Psychological Bulletin* between 1978 and 2006, a criterion for inclusion among others was that meta-analyses presented data tables containing the most important data on included original studies. Required information were effect sizes, sample size and other information necessary for coding of studies. To the surprise of Schmidt et al. (2009), only few meta-analyses met this criterion. It was stated that the sample of meta-analyses selected for reanalysis was typical of meta-analyses that were published in *Psychological Bulletin* the last 20 years concerning methods, except that they presented all data required for reanalysis. Schmidt et al. (2009) reported that 169 of meta-analyses published in *Psychological Bulletin* between 1978 and 2006 could be classified as using fixed-effect

models, random-effects models or both. Only five studies could be found that met the criteria of using a fixed-effect model and reporting data needed for reanalysis (Schmidt et al., 2009). That means that only around 3% (5 out of 169) of the meta-analyses screened for inclusion were identified as appropriate.

Fricke and Treinies (1985) examined a sample of 67 meta-analyses published between 1977 and 1984 and concluded that only 15% (10 meta-analyses) reported effect size and sample size for each included original study and therefore only 15% could be reproduced without consulting the reports of primary research. Press (1990) also commented on the absence of relevant data in reports of meta-analyses. It was noted that many meta-analyses reported only effect sizes, although other information such as standard errors ought to be reported. Moreover, it was recommended that meta-analyses presented data tables containing sample sizes (for treatment and control groups), effect size, values of important background variables and standard errors (for each treatment group) for each included study (Press, 1990).

Jennions and Möller (2002) examined 44 meta-analyses in the field of biology to assess the relationship of magnitude of effect size and year of publication. A total of 81 meta-analyses (published between 1991 and 2001) initially seemed to be suitable for inclusion, but 37 meta-analyses had to be excluded on the basis of different criteria. Not providing effect sizes for original studies was one of the exclusion criteria (Jennions & Möller, 2002). Around 54% (44 out of 81) of the meta-analyses were found to be eligible for inclusion. Problems similar to the ones described by Schmidt et al. (2009) concerning the quality of reporting of meta-analyses seem to occur also in fields other than psychology, like in this case biology (Jennions & Möller, 2002). Furthermore, it became obvious that meta-analyses published between 1991 and 2001 could still show deficiencies concerning the reporting of effect sizes (Jennions & Möller, 2002).

Even on the level of primary research studies similar reporting problems can be found, as described by Cooper (1979). In a meta-analysis of sex difference in conformity, 12 of the 38 studies that were supposed to be included in the meta-analysis provided no statistics (Cooper, 1979). Cooper (2010) depicted the frustration when purchased original studies do not contain the necessary information and Lipsey and Wilson (2001) emphasised that it was “distressingly common”(p. 35) that information provided is not sufficient to calculate effect sizes. Hence the same problems concerning the quality of reporting can be found at the level of meta-analyses as well as at the level of original studies. As Tom Cook in an Annual Meeting symposium on secondary analysis expressed

it: “You can get the data if you have chutzpah or if you‘re sociometrically well-connected” (Glass, 1978, p. 3).

It is to be expected that the same problems as described above concerning reporting quality will be encountered in the course of this work. One reason to believe this is that even meta-analyses published after the end of the period of classic meta-analyses show deficiencies in reporting (Schmidt et al., 2009; Jennions & Möller, 2002). Another reason is the occurrence of the same problems in another area than psychology (Jennions & Möller, 2002). Finally, also the basis of meta-analysis, primary research studies display the very same problem (e.g., Lipsey &Wilson, 2001).

Consequently, in the course of this work it is first investigated, whether the reporting quality of classic meta-analyses would allow for reanalysis using state-of-the-art methods. If so, the data could then be reanalyzed with the intention of answering the original research questions using modern methods. Thereafter the question whether new or different conclusions could be drawn from the original data using state-of-the-art methods could be approached, comparing the original results to the results obtained through reanalysis.

Methods

2 Methods

2.1 Literature search

Meta-analyses published in peer-reviewed journals between 1977 and 1982 reported in English language were located through different literature search strategies. There were no restrictions placed upon the search regarding the topic of the meta-analyses. First, appropriate keywords (meta-analy*, quantitative synthesis, statistical review, research synthesis, integrating findings, quantitative review, combining results, integrative review, research integration) were entered in the title fields of the electronic literature databases PsycINFO and Web of Science. Second, a bibliography (Fricke, 1982) was screened for further classic meta-analyses. Third, modern books on meta-analysis and systematic reviews (Borenstein et al., 2009; Cooper, 2010; Lipsey & Wilson, 2001; Petticrew & Roberts, 2006) and also older books concerned with meta-analysis (Cook et al., 1992; Cooper, 1984; Fricke & Treinies, 1985; Glass, McGaw, & Smith; 1981; Hunter, Schmidt, & Jackson, 1982; Light & Pillemer, 1984; Rosenthal, 1984; Smith, Glass, & Miller, 1980; Treinies & Fricke, 1983; Wachter & Straf, 1990) were screened for classic meta-analyses as well as an article (Chalmers, Hedges, & Cooper, 2002) and a book (Hunt, 1997) on the history of meta-analysis. Finally, retrieved classic meta-analyses themselves were read to search for possible further classic meta-analyses mentioned therein.

2.2 Inclusion criteria

Besides being published in a peer-reviewed journal between 1977 and 1982 and being reported in English language, a study had to meet four criteria to be included for reanalysis. First, the study had to be a meta-analysis. A definition is given by Borenstein et al. (2009): “Meta-analysis refers to the statistical synthesis of results from a series of studies” (p. xxi). The most common approach and the one Borenstein et al. (2009) and others (e.g., Lipsey & Wilson, 2001) primarily focused on in their books, and therefore also the one focused on in this work, is meta-analysis of effect sizes. In this approach every study of a set of studies contributes an estimate of some statistic, then the dispersion in these effects is assessed and a summary effect could be calculated as well. Other

approaches are meta-analyses that combine *p* values or psychometric meta-analyses (Borenstein et al., 2009). Meta-analyses that combine *p* values are explicitly excluded not only because it is a different approach of meta-analysis, but also because for *p* values effect magnitude is confounded with sample size (Lipsey & Wilson, 2001). Therefore, more precisely the first criterion was that the study had to be a meta-analysis of effect sizes. The other criteria concerned the information reported for the meta-analysis. As already mentioned above, Press (1990) recommended that meta-analyses at least presented data tables containing sample sizes (for treatment and control groups), effect size, values of important background variables and additionally standard errors (for each treatment group) for each included study. Contemporary guidelines for reporting of meta-analyses, PRISMA Standards (Liberati et al., 2009) and MARS (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) give a comprehensive account of the information that should be reported for a meta-analysis. To be able to reanalyze a meta-analysis, one of the most important pieces of information needed is the effect size for every included primary research study. The effect size is "a value which reflects the magnitude of the treatment effect or (more generally) the strength of a relationship between two variables, is the unit of currency in a meta-analysis" (Borenstein et al., 2009, p. 3). Accordingly, the second criterion was that an effect size was reported, or was computable from presented information, for each included original study. To enable the application of state-of-the-art methods to data of classic meta-analyses, the data need to meet the requirements of those methods. Hence, it is important that the effect sizes intended to use for calculations could be handled by modern meta-analysis software. Definitely satisfying this requirement are the effect sizes common today. These are effect sizes based on means (unstandardized mean difference (*D*), standardized mean difference (*d*, *g*), response ratio (*R*)), effect sizes based on binary data (risk ratio (*RR*), odds ratio (*OR*), risk difference (*RD*)) and effect sizes based on correlational data (correlation (*r*)) (Borenstein et al., 2009). Therefore, the third criterion was that reported effect sizes were common today or at least possible to enter into spreadsheets of the software intended to execute calculations, which in this case was Comprehensive Meta-Analysis, version 2.2.030 (Borenstein et al., 2005). The fourth criterion concerned other important information needed to reanalyze a meta-analysis using state-of-the-art methods. To calculate the summary effect according to state-of-the-art methods of meta-analysis (Borenstein et al., 2009), each included study is weighted. A more precise study is assigned more weight than a study with poor precision because a more precise study carries more

information. This is because the larger the sample the more variance decreases and hence the more precise is the estimate of effect size. Studies are weighted according to their precision and therefore the assigned weight is the inverse of the variance of the respective study. Obviously the variance for each study is necessary to compute weights. For computing the variance for each study the respective sample size of the treatment and control group is needed (Borenstein et al., 2009). Therefore other information needed to reanalyze a meta-analysis was variance (or standard error) for each original study, or information to calculate variance. Press (1990) mentioned that also values of important background variables should be reported.

Additionally reported data on variables examined in the original meta-analysis could be useful to reproduce further analyses concerning these variables. The fourth criterion was that variance (or standard error) for each included study, or information to calculate variance, was reported. Any additional reported data on variables examined in the original meta-analysis was a plus. To summarise the above:

Four criteria had to be met by a study to be included for reanalysis which were:

1. The study had to be a meta-analysis of effect sizes.
2. An effect size was reported, or was computable from presented information, for each included original study.
3. Reported effect sizes were common today or at least possible to enter into spreadsheets of the software intended to execute calculations.
4. Variance (or standard error) for each included study, or information to calculate variance, was reported.

After closer examination of the remaining studies that met these four criteria, a fifth criterion covering further requirements for reanalysis not considered so far had to be adopted.

2.3 A general overview of literature search results

Literature search described above yielded a total of 102 articles which were all published in peer-reviewed journals between 1977 and 1982 and reported in English language. As mentioned above, Borenstein et al. (2009) defined meta-analysis as follows: “Meta-analysis refers to the statistical synthesis of results from a series of studies” (p. xxi). When applying the definition of meta-analysis given by Borenstein et al. (2009), 94 of the

articles found could be classified as a meta-analysis. Of the 94 meta-analyses, only three reported data for each included original study required for reanalysis using state-of-the-art methods. Examination of all 102 studies using the four inclusion criteria is described in the following section.

Results

3 Results

An extensive literature search based on the strategies and rules stated above resulted in a total of 102 articles. This present corpus of literature comprises articles published in peer-reviewed journals between 1977 and 1982 and reported in English language. Each article was inspected to examine whether it met the four inclusion criteria. The articles were examined in a stepwise procedure regarding the four criteria. First, all of the 102 studies were assessed to see whether they met the first criterion. The ones not meeting it were excluded from reanalysis. The remaining studies were assessed whether they met the second criterion. Again the ones not meeting it were excluded from reanalysis and the other studies were assessed whether they met the third criterion. Then the ones meeting the third criterion were assessed whether they met the fourth criterion. At the end of this procedure only those studies meeting all four criteria remained, which were then screened again for any further problems for reanalysis. A list of all articles and the corresponding reasons for exclusion or inclusion can be found in Appendix A.

3.1 Articles excluded from reanalysis

The first criterion was that the study had to be a meta-analysis of effect sizes. In the first step, all articles not presenting a meta-analysis were excluded. To repeat the previous, meta-analysis refers to the statistical integration of results from a set of studies (Borenstein et al., 2009). The most common approach and the one Borenstein et al. (2009) and others (e.g., Lipsey & Wilson, 2001) primarily focused on in their books is meta-analysis of effect sizes. Other approaches are meta-analyses that combine *p* values or psychometric meta-analyses (Borenstein et al., 2009). Applying the definition of meta-analysis given by Borenstein et al. (2009), 94 studies could be classified as meta-analysis and eight articles had to be excluded because of not presenting a meta-analysis. Kennedy (1978) evaluated an education program where different educational models were employed. It was a primary analysis rather than a meta-analysis because there was no statistical integration of results from a set of studies dealing with the same topic. Instead results from different educational models were statistically integrated to gain insight into how they worked. In the same way Hereford (1979) evaluated a program concerning the Keller Plan. Results from different

courses were statistically integrated instead of results from a series of studies. Kazrin, Durac, and Agteros (1979) evaluated the efficacy of psychotherapy in a humorous way but did not conduct a meta-analysis. Ladas (1980) described a microanalysis of different studies on the topic of note taking from lectures but no statistical integration of results was performed. Berk and Chalmers (1981) discussed the results of a series of studies on cost and efficacy if inpatient care was substituted by ambulatory care. Only a few papers had reported enough data and it was concluded that further data had to be collected for decision making (Berk & Chalmers, 1981). Smith and Land (1981) also described and discussed results of different studies, in this case on the effect of low-inference teacher clarity variables on student achievement and student perception of teacher effectiveness, but did not perform a statistical integration of these results. Cotton and Cook (1982) discussed the results of a meta-analysis conducted by Johnson, Maruyama, Johnson, Nelson, and Skon (1981) and drew different conclusions from the results. Hattie and Hansford (1982) shortly depicted the results of a meta-analysis, which was described in more detail elsewhere, and compared them to the results of a literature review. The article was excluded because the meta-analysis used for the purpose of comparing the two methods was described elsewhere. In a second step, meta-analyses using a different approach than that of effect sizes were excluded. Of the 94 meta-analyses, 16 were excluded because a different approach had been used. Six studies had to be excluded because they were psychometric meta-analyses, which is also called the Hunter-Schmidt approach to meta-analysis (Borenstein et al., 2009). Another five studies were meta-analyses that combined *p* values and/or *Z* scores, and thus were excluded too. Each of the other five studies was excluded based on different reasons respectively. An approach different to the one required was employed by Anonymous (1980). Again this served as a reason for exclusion. For each of six trials on the benefits of aspirin after myocardial infarction, Anonymous (1980) compared the number of deaths among aspirin-takers and the expected number of deaths if aspirin had no effect. The difference between the two numbers for each trial was calculated and then the values were summed up to see whether there were fewer deaths than expected in aggregate. Also, Blanchard, Andrasik, Ahles, Teders, and O'Keefe (1980) employed an approach different to the one required and therefore the meta-analysis conducted had to be excluded. Blanchard et al. (1980) had intended to compare the effects of psychological treatments for headache to each other and to the effects of placebo using meta-analysis. To answer the research question a measure called percent improvement score had been employed, where the end of treatment value was subtracted from the baseline value, the

result then was divided by the baseline value and after that the total result was multiplied by 100. For each of the included original studies this percent improvement score had been calculated for different dependent variables for each treatment condition respectively. Furthermore, the average degrees of improvement had been calculated for each treatment condition and then the conditions had been compared to each other (Blanchard et al., 1980). Inglis and Lawson (1982) examined whether there was an influence of sex on the effects of unilateral brain damage on intelligence test results. Only few of the original studies had reported separate scores for males and females. Hence, the proportions of males and females reported in the individual studies were considered in a linear regression analysis to answer the research question. Therefore because no effect sizes were calculated, this meta-analysis was excluded too. Tornatzky and Klein (1982) carried out a sign test for a set of studies, wherein only the direction of the findings, in this case correlations, was taken into account. Neither magnitude nor statistical significance of the individual findings was of interest in this approach. The reason for exclusion in this case again was the application of an alternate approach. Finally, there was the meta-analysis conducted by Cooper (1979), which was a special case. In this meta-analysis *Z* scores were combined to answer the research questions, but interestingly also *d* indexes were reported for each included original study. It was pointed out that effect sizes are important to describe the strength of a relationship. The several hypotheses were tested combining *Z* scores, but in one case, when the size of the effect was in question, *d* indexes were averaged. Even though effect size data (*d* indexes) were reported for every original study included in the meta-analysis, the meta-analysis had to be excluded because its focus was on combining *Z* scores as a method to answer the research questions. In summary, after examining whether the 102 retrieved articles met the first criterion it could be concluded, eight studies had to be excluded because of not reporting a meta-analysis. Of the remaining 94 studies another 16 meta-analyses were excluded because of not applying the required approach of meta-analysis, namely meta-analysis of effect sizes. Therefore, 78 meta-analyses were left to be assessed whether they met the second criterion.

The second criterion was that an effect size was reported, or was computable from presented information, for each included original study. A total of 61 meta-analyses had to be excluded because they did not meet the second criterion. For example, Smith and Glass (1977), Rosenthal and Rubin (1978), and also Kulik and Kulik (1982), among many others, reported aggregate results only. There was no information provided on the single original studies that had been included in the meta-analysis. As a consequence, effect sizes for the

included original studies were not available. Andrews, Guitar, and Howie (1980) did provide information on each included original study and even information concerning effect sizes, but not the required effect sizes themselves. Instead, Andrews et al. (1980) reported the number of effect sizes that had been calculated for each original study. Cohen (1980) in turn reported major characteristics of the included studies, like design features, nature of feedback and outcome measures, but again no effect sizes. Ide, Parkerson, Haertel, and Walberg (1981) described the results of each included study in detail, but only sometimes provided data. Consequently, the effect sizes that had been the basis of the meta-analysis were not available.

A different kind of problem, but also concerning reporting of effect size data, became obvious in the meta-analysis of Iverson and Walberg (1982). In this case, the mode of reporting effect sizes obstructed inclusion for reanalysis. In the meta-analysis of Iverson and Walberg (1982) correlations were the units of analysis. Therefore these units would need to be available to reanalyze the meta-analysis, but instead of reporting all the correlations that had been used for computations, the authors reported type and range of correlation for each included original study. The effect sizes, namely correlations, used by Iverson and Walberg (1982) for their meta-analysis could not be determined based on the given ranges of correlations. For this reason it was concluded that the meta-analysis did not meet the second criterion and thus was excluded. These studies are only some of the total of 60 meta-analyses that were excluded because not providing the required information on effect sizes for each included original study. The list of all 102 articles that were assessed whether they met the criteria required for reanalysis and the respective reasons for exclusion or inclusion can be found in Appendix A. Of the prior 78 meta-analyses that were assessed whether they met the second criterion, 17 meta-analyses actually fulfilling the second criterion remained. Those 17 meta-analyses thus reported effect sizes (or information to calculate them) for all original studies included in the meta-analysis. In the next step these were assessed whether they met the third criterion.

The third criterion was that reported effect sizes were common today or at least possible to enter into spreadsheets of the software intended to execute calculations, namely Comprehensive Meta-Analysis, version 2.2.030 (Borenstein et al., 2005). Of the remaining 17 meta-analyses five had to be excluded because they did not meet the third criterion. Meta-analyses conducted by Levy, Iverson, and Walberg (1980), Kavale (1981), Iverson and Levy (1982), Mumford, Schlesinger, and Glass (1982) and Wampler (1982) were excluded because of employing an effect size, where the mean of the control group is

subtracted from the mean of the experimental group and the result divided by the standard deviation of the control group. It is applied when the standard deviations of the (two) groups are heterogeneous (Glass et al., 1981). This kind of effect size cannot be processed by the meta-analysis software intended to execute calculations. The exclusion of the five mentioned meta-analyses left 12 meta-analyses to be assessed whether they met the fourth criterion.

The fourth criterion was that variance (or standard error) for each included study, or information to calculate variance, was reported. Of the 12 remaining meta-analyses another six meta-analyses had to be excluded because of missing information. As already stated above, some crucial information besides the effect size is needed to conduct a meta-analysis according to state-of-the-art methods. To calculate the summary effect, each included study is weighted. The assigned weight is the inverse of the variance of the respective study. Therefore, the variance for each study is necessary to compute weights and for computing the variance for each study the respective sample size of the treatment and control group is needed if a standardized mean difference was the effect size (Borenstein et al., 2009). Hall (1978) examined gender differences concerning the ability to decode nonverbal cues. In a table, Hall (1978) reported, if available, the size of the tested sample as a whole, direction of the effect and effect size, in this case Cohen's *d*, for each included original study. As can be seen, Hall (1978) did not report variance or standard error for each included study. Furthermore, it was not possible to calculate variance for each included study because Hall (1978) had not reported separate sample sizes for the treatment and the control group but only the size of the tested sample as a whole. Hence, information to calculate variance was missing and as a consequence the meta-analysis conducted by Hall (1978) had to be excluded. The same problem was encountered in the meta-analysis conducted by Arkin, Cooper, and Kolditz (1980). Arkin et al. (1980) compared two conditions and presented a *d* index for each included original study. As was the case in the meta-analysis of Hall (1978), the size of the tested sample as a whole per included study was reported instead of the required separate sample size for each condition. Hence, this meta-analysis had to be excluded, too. Also, Smith (1980) compared two conditions and reported an effect size (standardized mean difference) for each included study. The difference to the meta-analyses described above is that Smith did not even report any sample size. The missing information on separate sample sizes of the examined groups served as a reason for exclusion of the meta-analysis. Based on the same line of argument another two meta-analyses had to be excluded. Burger (1981) again

compared two conditions, reporting the associated effect size for each included study, but additionally reported the sample size for one condition only. Hyde (1981) examined gender differences as did Hall (1978), and in the same manner reported the size of the tested sample as a whole per included study only, besides the effect size. Finally, a sixth meta-analysis had to be excluded because information to calculate variance was missing. Cooper, Burger, and Good (1981) examined gender differences concerning locus of control beliefs of school children using meta-analysis. Cooper et al. (1981) gave an account of the raw data that had been reported in the original research studies included in the meta-analysis. Besides information on authors, year and grade of school for each included original study, Cooper et al. (1981) reported means for different subscales for females and males separately as well as separate sample sizes for females and males. Thus, requirements to calculate an unstandardized mean difference (*D*) were met; but data to calculate its variance were missing. To calculate the variance of *D*, the sample standard deviations of the two groups were necessary in addition to the sample sizes in the two groups (Borenstein et al., 2009). As can be seen from the description of the data reported by Cooper et al. (1981), data on the sample standard deviations of the two groups had not been presented. Hence, the variance of *D* could not be calculated and as a consequence the meta-analysis had to be excluded. After exclusion of six meta-analyses because they were not meeting the fourth criterion, six meta-analyses remained.

After all 102 articles had been examined in a stepwise procedure regarding the four criteria, it appeared that six meta-analyses met all of the four. It was assumed that these six classic meta-analyses could be reanalyzed using state-of-the-art methods. However, closer examination necessitated the adoption of a fifth criterion covering any other requirement for reanalysis not considered so far and therefore concerning any other problems obstructing reanalysis. Another three meta-analyses had to be excluded because of specific problems described in detail below.

Glass and Smith (1979) examined the relationship between class size and achievement. Included for meta-analysis were a total of 77 studies yielding 725 effect sizes. Effect sizes were calculated as follows: the estimated mean achievement of the larger class was subtracted from the estimated mean achievement of the smaller class; the result was then divided by the estimated within-class standard deviation. It was assumed that the standard deviation was homogenous across the two classes. Therefore, the effect size employed by Glass and Smith (1979) met the third criterion. In contrast to the effect size employed by Glass and Smith (1979), the effect size employed by Levy, Iverson, and

Walberg (1980) and others described above, assumed standard deviations to be heterogeneous across different groups. This distinction is crucial to decide whether the effect size could be processed by the software supposed to execute calculations. In their article, Glass and Smith (1979) mentioned that a longer report existed containing the entire data set on which the meta-analysis was based. The report (Glass & Smith, 1978) was accessible online and could thus be accessed easily. In its appendix, the data set used by Glass and Smith (1979) was found. Besides other information, the effect size and also the separate sample sizes of the groups compared were reported for each included original study. All of the four criteria for inclusion seemed to be met and still no obstacle was obvious. Only when attempting to enter the reported data into a spreadsheet the problem became apparent. Parts of the document were illegible. After an unsuccessful literature search for a more legible version of the report, the first author of the report, Gene V. Glass (personal communication, April 14th, 2011), was contacted through E-mail for assistance. However, unfortunately a different version could not be made available. It is important that reanalysis is based on exactly the same basis of data as the original meta-analysis to enable comparisons of the results. Because it was not feasible to reconstruct the data basis for reasons of illegible data the meta-analysis conducted by Glass and Smith (1979) had to be excluded. A problem concerning reporting, especially the reporting of effect sizes, obstructed inclusion of the meta-analysis conducted by Kremer and Walberg (1981). Kremer and Walberg (1981) examined the relation between science learning and achievement and three constructs, namely student motivation, home or family environment and peer environment using meta-analytical techniques. Correlation was the effect size in this case. The first criterion, employing meta-analysis of effect sizes, and third criterion, using a common effect size were thus obviously met. In several tables, Kremer and Walberg (1981) reported study features, characteristics of tested subjects along with sample size and findings described verbally as well as median correlation and box score for each included original study. It appeared that the second criterion was also met because an effect size had been reported for each included original study. However, Kremer and Walberg (1981) explained that “mean correlations were computed for each construct area using the raw correlations reported in individual studies” (p. 14). In the presented tables Kremer and Walberg (1981) had not reported the so-called raw correlations, which were the basis of calculations but, among other information, only one (or no) correlation for each included original study, in most cases the median of reported correlations. All correlations in all studies and also the median correlations though had been presented in

leaf diagrams. In these diagrams correlations were arranged according to the three constructs. Again it appeared that the second criterion was met. However there was a problem in that the correlations reported in the diagrams could not be attributed to the corresponding original studies. On the one hand, correlations that had been the basis of calculations were reported in diagrams, and on the other hand, all other information on the original studies, including the required sample size, was reported in tables. All of the information necessary to reanalyze this meta-analysis using the same basis of data was reported in the article, but unfortunately non-matchable. Therefore, this meta-analysis had to be excluded. The last meta-analysis that had to be excluded was the one conducted by Williams, Haertel, Haertel, and Walberg (1982). Williams et al. (1982) conducted a meta-analysis to examine the effect of leisure-time television on school achievement. Effect size employed in this case was correlational. Hence, the first criterion was met, it was a meta-analysis of effect sizes. Williams et al. (1982) reported mean and standard deviation of correlations, and also the number of correlations that had been coded, for each of the 23 included original studies. The second criterion, that an effect size was reported for each included original study, was thus met. Also, the third criterion was met, because the effect size was correlational and therefore a common effect size. To calculate the variance of a correlation, the sample size is needed besides the correlation itself (Borenstein et al., 2009). Williams et al. (1982) in their appendix reported the sample size for almost all of the included studies, besides other information. The fourth criterion was thus met, too. Furthermore, for each included study many different characteristics had been coded. On the one hand, study characteristics such as location or year of the study and on the other hand sample characteristics such as age and sex had been coded. This was special as the meta-analysis was conducted in a stepwise procedure. In a first step, study characteristics were examined. For this analysis step, the reported mean correlations per included original study were used. In a second step, sample characteristics were examined. In this case, calculations were based on 274 single correlations related to sample characteristics, which had not been reported in the article. Because the necessary data to reanalyze the second step were missing, it was considered to only reanalyze the first step of the meta-analysis, for which data seemed to be available. But while attempting to reanalyze the first step, new problems appeared. Williams et al. (1982) had reported the results concerning the locations of the included studies arranged according to regions in the USA and otherwise states in general. The problem was to reconstruct the classification of regions in the USA employed by Williams et al. (1982). Williams et al. (1982) had subdivided the USA into five regions.

It was attempted to find out how regions in the USA were commonly subdivided and which locations fit to which region. For that reason the website of the U.S. Census Bureau was consulted. The U.S. Census Bureau however divides the USA into four census regions (U.S. Department of Commerce Economics and Statistics Administration, U.S. Census Bureau, n.d.). Even though other rather unofficial websites had been screened for information, where actually regions other than that used by Williams et al. (1982) were found, the classification employed by Williams et al. (1982) remained inscrutable. Moreover, the number of correlations classified was not 23, as was the number of studies included in the meta-analysis, but 20, even though a category 'not specified' existed. This phenomenon, that the number of correlations did not correspond to the number of included original studies, appeared concerning other characteristics, but in this case was accompanied by an explanation of Williams et al. (1982). For example one of the methodological characteristics was whether random sampling had been applied or not. A total of 14 correlations had been classified as 'Yes' and another 10 correlations as 'No', resulting in a total amount of 24 correlations. The reason for that could be found in a note which explained that one of the included studies had been treated as two separate studies. Causing problems was the fact that only 23 correlations had been reported instead of 24 correlations. On the study-level, the data set included 24 data points, as described by Williams et al. (1982). Hence, reanalysis based on the same data set as the meta-analysis conducted by Williams et al. (1982) seemed to be impossible, because necessary information was missing. The problems concerning the regions of the USA described above and further the problem of missing data obstructed reanalysis. Finally, a third problem that appeared should be described only briefly, because the problems described above already served as a reason for exclusion. The included original studies had been classified according to their research design as either surveys or quasi-experiments. Following the information given in a table, 20 studies had been classified as surveys and four as quasi-experiments. Three studies could easily be classified as quasi-experiments but there were two further that made decision difficult, because they were so similar. It is important to note that the study that had been treated as two separate studies could be classified as a survey. Even if this dilemma probably was to be resolved through discussion with senior researchers, it could not be ensured that the resulting decision was consistent with the decision made by Williams et al. (1982). A reanalysis based on the same prerequisites could therefore not be ascertained. Altogether, the described problems made the exclusion of the meta-analysis inevitable.

After the stepwise examination of 102 articles only three remained to be included for reanalysis using state-of-the-art methods and thus fulfilling all of the required criteria. In summary, 99 articles had to be excluded. First, eight articles were excluded because no meta-analysis had been conducted and another 16 meta-analyses were excluded because of not employing the required approach of meta-analysis, namely meta-analysis of effect sizes. Then a total of 61 meta-analyses were excluded because effect sizes had not been reported, or had not been computable from presented information, for each included original study. After that another five meta-analyses had to be excluded because an effect size not common today had been employed. Then again, six meta-analyses were excluded because of missing information concerning variance. Finally, another three meta-analyses had to be excluded because of specific problems obstructing reanalysis. A total of three meta-analyses met all of the inclusion criteria and showed no reasons that would necessitate exclusion. Thus, the meta-analyses conducted by Sparling (1980), Desilva, Hennekens, Lown, and Casscells (1981) and Stampfer, Goldhaber, Yusuf, Peto, and Hennekens (1982) were included for reanalysis using state-of-the-art methods. Figure 2 (on the following page) shows a flow chart of the described stepwise procedure. A list of all articles and the corresponding reasons for exclusion or inclusion can be found in Appendix A.

In the following section, a reanalysis of the three meta-analyses included is presented. Each of the meta-analyses is dealt with separately. First, the original meta-analysis and its results are described. Then, the original set of data is reanalyzed using state-of-the-art methods. Comparison of results and conclusions drawn from the comparison are provided in the Discussion section.

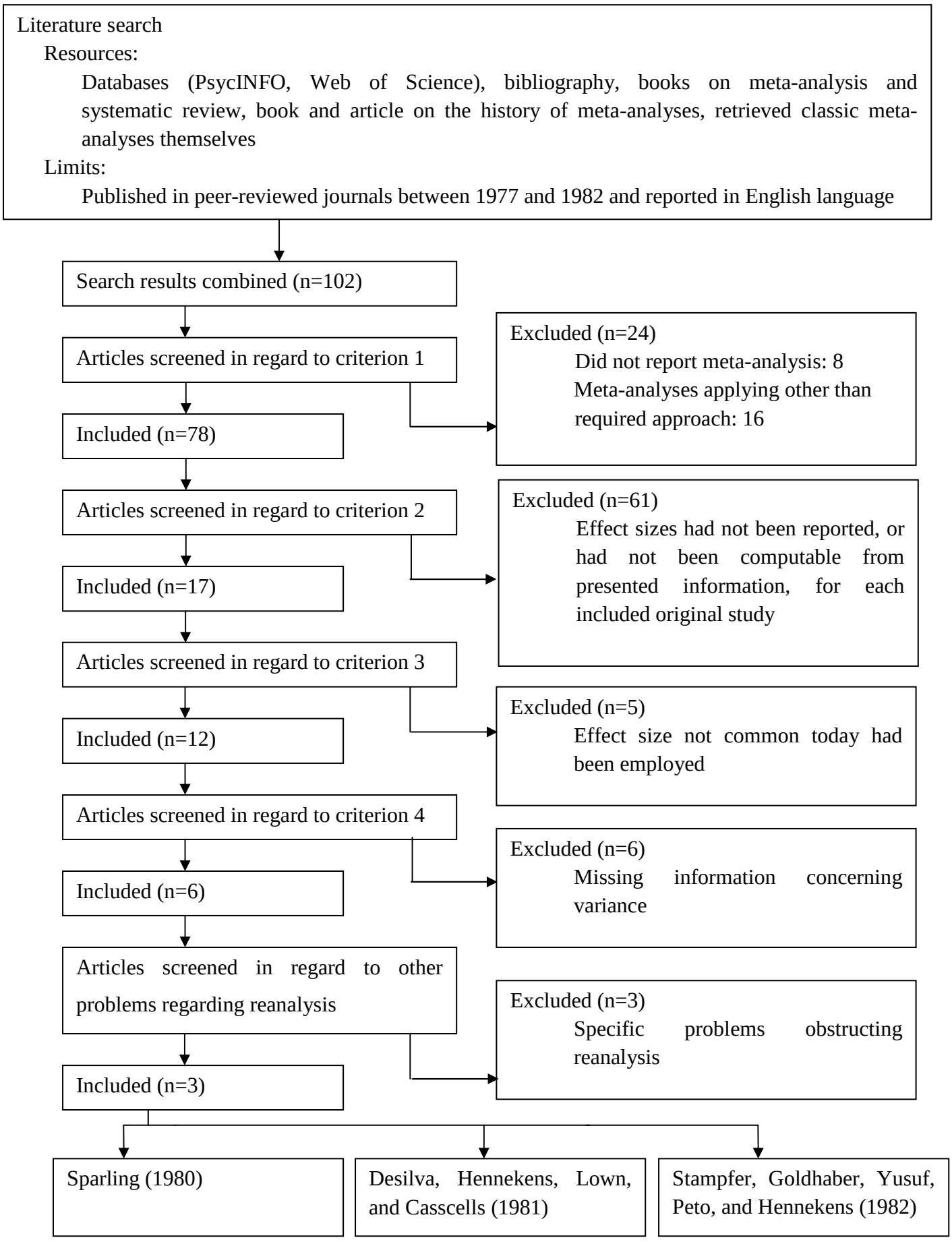


Figure 2. Flow chart of the stepwise examination of articles regarding five criteria required to be met to enable reanalysis using state-of-the-art methods.

3.2 Reanalysis of included meta-analyses

In this section the actual reanalysis of the three classic meta-analyses using state-of-the-art methods is presented. First an overview of methods applied in all three cases is given. Then, each meta-analysis is dealt with separately. At first the original meta-analysis is described, concerning the research questions addressed, study characteristics of the primary studies included for meta-analysis, and the methods and results obtained. Next, the original data set is reanalyzed using state-of-the-art methods addressing the same research questions and additionally publication bias and power analysis. For this the data presented in the article of the original meta-analysis were entered into spreadsheets. To avoid transcription errors, the values were counterchecked by an independent observer. Results obtained, when state-of-the-art methods were applied, were contrasted to the results of the original meta-analysis in a table (see Tables C1-C3, Appendix C). Comparison of the results of the original meta-analysis and the respective reanalysis as well as conclusions drawn from this comparison can be found in the Discussion section of the present work.

3.2.1 Methods of analysis

The three classic meta-analyses were reanalyzed using state-of-the-art methods with the intention of answering the original research questions. In the present work, state-of-the-art methods are meta-analytic procedures as described in the books already mentioned above (Borenstein et al., 2009; Cooper, 2010; Lipsey & Wilson, 2001; Petticrew & Roberts, 2006). Applied meta-analytic procedures, primarily followed the instructions of Borenstein et al. (2009), which are described below.

The objective of a meta-analysis is not only to compute a summary effect but also to examine the pattern of effects. If it was to be expected that all of the studies share a common effect size and therefore have the same true effect size, then true heterogeneity is zero. Observed dispersion of effect sizes in this case is only due to within-study error. If it was assumed that the true effect sizes vary between the studies, the observed dispersion includes true variance (real heterogeneity in effect sizes) and within-study (random) error. Hence, besides computing the summary effect it is important to examine the dispersion of effect sizes (Borenstein et al., 2009). Measures of heterogeneity computed in the present work are Cochran's *Q statistic* (df = number of studies minus one) and its *p* value (Borenstein et al., 2009) and the I^2 index. The *Q statistic* and its *p* value are concerned with

the null hypothesis that the true dispersion is exactly zero. The I^2 index reflects the proportion of the observed variance that is true (Borenstein et al., 2009). The I^2 index was interpreted according to Higgins, Thompson, Deeks, and Altman (2003), wherein values of I^2 of 25%, 50%, and 70% are described as low, moderate, and high, respectively.

Associated with heterogeneity is the fundamental decision of either applying a fixed-effect or random-effects model. Borenstein, Hedges, Higgins, and Rothstein (2010) recommended that the fixed-effect model was to be used if the studies are all "functionally identical" (p. 105) and if it was aimed to calculate the common effect size, which would not be generalized to other populations than the one of the analysis. When it is not assumed that all the studies share the same true effect size, and it is aimed to generalize the findings of the meta-analysis, the application of a random-effects model is appropriate. Borenstein et al. (2009) warned that the decision should not be based on the result of the test for heterogeneity. Borenstein et al. (2010) mentioned that it is sometimes practiced that a statistically non-significant test for heterogeneity is used as evidence that the studies share a common effect size. Because the test of heterogeneity often has poor power its result could be misleading (Borenstein et al., 2010). Overton (1998) however brought up the idea of using both models in the same meta-analysis to reveal the extent to which results are dependent on the assumptions of the respective model. For the present work the idea of Overton (1998) was adopted and the data were reanalyzed using both models.

Existent true variance could be explained by covariates. For this, methods such as subgroup analysis or meta-regression could be used (Borenstein et al., 2009). In the course of reanalysing the three meta-analyses, subgroup analysis was applied if appropriate as well as sensitivity analysis. Sensitivity analysis examines the robustness of the findings (Borenstein et al., 2009) and in the present work was applied to examine the influence of single studies on the overall result.

Furthermore, publication bias, when the sample of studies included for meta-analysis is not representative for all relevant studies (Borenstein et al., 2009), was addressed using different approaches. First one was the visual inspection of funnel plots. In a funnel plot the relationship between study size and effect size is displayed. Since sampling error is random, the studies are distributed symmetrically about the mean effect size when there is no publication bias. Therefore asymmetry is a sign of publication bias. As the visual inspection is a rather subjective approach, other methods to test the relationship between sample size and effect size were executed. These were Begg and Mazumdar's rank correlation method and Egger's regression test (Borenstein et al., 2009).

For the mentioned methods to make sense, a reasonable number of studies and amount of dispersion in the sample sizes is needed (Borenstein et al., 2009), which could be a problem in the present work because meta-analyses to be reanalyzed included only a few studies. Furthermore, the Duval-Tweedie trim-and-fill method was utilised to test for publication bias. The Duval-Tweedie trim-and-fill method aims to establish a symmetrical distribution of studies, observed studies complemented by theoretically missing ones, about an unbiased estimate of the effect size utilising an iterative procedure. Funnel plots generated by computer programs that incorporate trim-and-fill include observed as well as imputed studies, so that a change of effect size could be detected, when imputed studies are included (Borenstein et al., 2009).

Finally retrospective power analysis was carried out. Muncer, Craigie, and Holmes (2003; see also Muncer, Taylor, & Craigie, 2002) suggested adapting a power criterion to meta-analysis to ensure that only studies with sufficient power to detect the assumed effect size are included. This ensures that a summary effect is produced that is based on studies that have sufficient power to support it. The power criterion suggested by Muncer, Taylor, and Smith (1999, as cited in Muncer et al., 2002) is .8 for health related meta-analyses and .5 for meta-analyses in the social sciences. The procedure to incorporate power analysis into meta-analysis suggested by Muncer et al. (2003) was intended to be used a priori but could also be employed for evaluation of existing meta-analyses. The latter was undertaken in the present work, following the instructions of Muncer et al. (2003) and using the statistical open-source software package R, version 2.13.1 (R. Development Core Team, 2011) to execute calculations concerning power. First, the power of each study to test for the original summary effect was calculated, as well as the average power of the studies in the meta-analysis. Thereafter, studies meeting the power criterion were meta-analytically integrated anew and a new summary effect was calculated. The power of each study to test for the new summary effect was again calculated and studies meeting the power criterion were combined into a new meta-analysis. This iterative procedure ended when there was no further change in the effect size and power calculated and thus the overall power of a meta-analysis was acceptable. The resulting new summary effect was reported together with the average power of the studies included in the new meta-analysis. The average power of the studies to detect the initial effect size could then be compared to the result of the described iterative procedure.

Software used to execute the meta-analytical calculations was Comprehensive Meta-Analysis, version 2.2.030 (Borenstein et al., 2005). The PRISMA Standards described by Liberati et al. (2009) were used as guidelines for the depiction of the conducted reanalyses and the according results.

Reporting of results of the three reanalyses was structured in four parts respectively: synthesis of results, additional analyses (subgroup analysis, sensitivity analysis), analyses concerning publication bias, and finally computations regarding retrospective power analysis. This structure was followed not only in the text, but also in the respective tables contrasting the results of the reanalysis and the corresponding original meta-analysis (Tables C1-C3, Appendix C).

3.2.2 Reanalysis of Sparling (1980)

Summary of the original meta-analysis

The meta-analysis conducted by Sparling (1980) examined the magnitude of sex difference in maximal oxygen uptake (VO_2 max). The goal was to calculate an overall effect for three different expressions of VO_2 max: expressed in absolute terms (liters/minute), relative to body weight (ml/minute*kg BW) and relative to fat-free weight (ml/minute*kg FFW). The reason for computing different expressions is that part of the sex effect appears to be related to differences between men and women in body weight and body fatness. Primary research studies had to include participants of both sexes in late adolescence or older and had to present measures of body composition as well as measures of VO_2 max. A total of 13 studies were included for meta-analysis. Measures of VO_2 max were collected testing participants on either a bicycle or treadmill ergometer. In two studies participants were tested on both, leading to 15 effect size measures being the basis for computations of the summary effect in case of VO_2 max expressed in absolute terms and relative to body weight. In case of VO_2 max expressed relative to fat-free weight the computation of the summary effect was based on 10 effect size measures. The summary measure in this meta-analysis was a point-biserial correlation (r_{pb}). The summary effect of the sex difference in VO_2 max expressed in absolute terms (liters/minute) was $r_{pb} = .81$ ($r_{pb}^2 = .66$), whereas expressed relative to body weight (ml/minute*kg BW) it was $r_{pb} = .70$ ($r_{pb}^2 = .49$). The summary effect of the sex difference in VO_2 max expressed relative to fat-free weight (ml/minute*kg FFW) was $r_{pb} = .59$ ($r_{pb}^2 = .35$). The variance in VO_2 max explained by sex was reduced when VO_2 max was expressed relative to body weight, even more when expressed relative to fat-free weight. Thus magnitude of sex difference in VO_2 max could be substantially reduced when variability in aerobic capacity due to disparities

in body weight and body fatness was taken into account. Another research question addressed was, whether the sex difference between trained men and women differed from the sex difference between untrained men and women, again examined concerning the three expressions of VO_2 max. For this analysis a percentage difference between men and women was calculated (M-F/F*100) for the different expressions for each of the 15 units of analysis. Of the 13 studies five involved participants who were by definition trained (males > 55 ml/minute*kg BW, females > 45, Sparling, 1980, p. 548). An average value was then calculated for the trained and untrained groups per expression of VO_2 max. For liters/minute the value for the trained men and women was 46% and 60% for the untrained ones. For ml/minute*kg BW the values were 25% versus 30% and for in ml/minute*kg FFW the values were 12% versus 13%. The sex difference in the trained groups was less than the one in the untrained groups in the first two comparisons and similar in the last comparison (Sparling, 1980).

The original data set presented in the article of Sparling (1980) is reproduced in Tables B1 and B2, which can be found in Appendix B. Reanalysis of the original meta-analysis yielded the following results. It needs to be noted that for each of the three expressions of VO_2 max separate meta-analytical calculations were executed, but were based on the same group of studies.

Reanalysis

Synthesis of results

The results of the 13 included primary studies were statistically integrated employing the single study as unit of analysis, thus computations of the summary effect were based on 13 effect sizes measures. An exception were computations concerning the expression of VO_2 max relative to fat-free weight, where the basis was nine effect size measures.

When a fixed-effect model was employed, the summary effect of sex differences in VO_2 max expressed in absolute terms (liters/minute) was r_{pb} = .809 (*95% CI* = 0.784-0.832, $p < .001$). Employing a random-effects model the result was r_{pb} = .810 (*95% CI* = 0.771-0.843, $p < .001$). Both models yielded a significant and substantial sex difference in VO_2 max expressed in absolute terms. Between-study effect heterogeneity was significant

(Q = 22.773, df = 12, p = .03) and moderate (I^2 = 47.305). In Figure D1 and D2 (Appendix D) the according forest plots of the fixed-effect (Figure D1) and random-effects analysis (Figure D2) can be found.

The summary effect of sex differences in VO_2 max expressed relative to body weight (ml/minute*kg BW) when a fixed-effect model was employed was r_{pb} = .729 (*95% CI* = 0.695-0.761, p < .001), whereas when a random-effects model was employed the result was r_{pb} = .720 (*95% CI* = 0.641-0.784, p < .001). Again both models yielded a significant and substantial sex difference in VO_2 max, in this case expressed relative to body weight. But the magnitude of sex difference was smaller than when maximal oxygen uptake was expressed in absolute terms. Thus the magnitude of sex difference was smaller when body weight was taken into account. Measures of heterogeneity showed a significant (Q = 46.541, df = 12, p < .001) and high (I^2 = 74.216) effect heterogeneity across studies. These results indicate a large amount of true variance. In Figure D3 and D4 (Appendix D) the according forest plots of the fixed-effect (Figure D3) and random-effects analysis (Figure D4) can be found.

When a fixed-effect model was employed, the summary effect of sex differences in VO_2 max expressed relative to fat-free weight was r_{pb} = .607 (*95% CI* = 0.552-0.657, p < .001). When a random-effects model was employed the result was r_{pb} = .599 (*95% CI* = 0.524-0.664, p < .001). For both analyses the overall result was still significant. But the magnitude of sex difference was again smaller than when VO_2 max was expressed in absolute terms or relative to body weight. Taking into account body fatness reduced the sex difference in VO_2 max substantially. The test of effect heterogeneity across studies was non-significant (Q = 12.053, df = 8, p = .149) and low (I^2 = 33.624), indicating a reduced amount of true variance compared to the results above. In Figure D5 and D6 (Appendix D) the according forest plots of the fixed-effect (Figure D5) and random-effects analysis (Figure D6) can be found.

Additional analysis: Subgroup analysis

To explore whether the magnitude of sex difference in VO_2 max for trained men and women differed from that for untrained men and women, subgroup analyses were performed separately for the different expressions of VO_2 max. Computations in the original meta-analysis were based on a value called percentage difference, which was existent for each expression of VO_2 max for all 15 units of analysis. Subgroup analyses

though are based on effect size measures and in the present work the study was assumed to be the unit of analysis. Therefore, as in the analyses above, subgroup analyses for the three expressions of VO_2 max were based on differing numbers of effect size measures. Computations were based on 13 effect size measures except for computations concerning the expression of VO_2 max relative to fat-free weight, where the basis was nine effect size measures.

When the magnitude of sex difference in VO_2 max expressed in absolute terms was compared for trained and untrained groups, no significant differences (Q (total between) = 0.001, df = 1, p = .975) were found, when a fixed-effect model was employed. There were also no significant differences (Q (total between) = 0.001, df = 1, p = .974) found, when a mixed-effects model was employed.

Also, no significant differences could be found when the sex difference in VO_2 max expressed relative to body weight was compared for trained and untrained groups, when a fixed-effect model (Q (total between) = 1.408, df = 1, p = .235) was employed as well as when a mixed-effects model (Q (total between) = 0.002, df = 1, p = .964) was employed.

The magnitude of sex difference in VO_2 max when expressed relative to fat-free weight was significantly different for trained and untrained groups (Q (total between) = 4.507, df = 1, p = .034), when a fixed-effect model was employed. When a mixed-effects model was employed, there were no significant differences between the two groups (Q (total between) = 1.752, df = 1, p = .186).

Analyses concerning publication bias

The conduction of three separate meta-analyses for the three expressions of VO_2 max again required three separate analyses concerning publication bias. The funnel plots (Figures E1 to E6, Appendix E) were plotted with the correlation transformed to Fisher's Z on the X axis and the standard error plotted on the Y axis. They include observed studies as well as studies imputed by the Duval-Tweedie trim-and-fill method, which was once based on a fixed-effect model (Figures E1, E3 and E5), and the other time based on a random-effects model (Figures E2, E4 and E6).

Visual inspection of the funnel plots for VO_2 max expressed in absolute terms (Figures E1 and E2), suggested asymmetry. Begg and Mazumdar's rank correlation

method (p (2-tailed) = 1.000) and Egger's regression test (p (2-tailed) = .827) were non-significant, indicating no publication bias. These tests have lower power (Borenstein et al., 2009), therefore the results must be interpreted cautiously. Furthermore, it is important to keep in mind that the tests were based on only thirteen studies. The Duval-Tweedie trim-and-fill analysis based on a fixed-effect model indicated three missing studies. The adjusted point estimate was r_{pb} = .791 (*95% CI* = 0.765-0.814). The Duval-Tweedie trim-and-fill analysis based on a random-effects model indicated three missing studies as well. The adjusted point estimate was r_{pb} = .788 (*95% CI* = 0.742-0.826). In both cases the adjusted point estimate was reduced compared to the unadjusted value. Both analyses indicated three missing studies (almost one fourth of observed studies), which made the existence of publication bias rather plausible. The conclusion drawn from the above results was that there was evidence of publication bias.

Visual inspection of the funnel plots for VO_2 max expressed relative to body weight (Figures E3 and E4) suggested no asymmetry. Also Begg and Mazumdar's rank correlation method (p (2-tailed) = .903) and Egger's regression test (p (2-tailed) = .598) indicated no publication bias. The Duval-Tweedie trim-and-fill analysis based on a fixed-effect model indicated no missing studies. The Duval-Tweedie trim-and-fill analysis based on a random-effects model also indicated no missing studies. It was concluded that there was no evidence of publication bias.

Visual inspection of the funnel plot for VO_2 max expressed relative to fat-free weight (Figures E5 and E6) suggested asymmetry. Begg and Mazumdar's rank correlation method (p (2-tailed) = .175) and Egger's regression test (p (2-tailed) = .509) were non-significant, indicating no publication bias. Once again, the reason for non-significant results could be low power of the test and the small sample size (Borenstein et al., 2009). The Duval-Tweedie trim-and-fill analysis based on a fixed-effect model indicated three missing studies. The adjusted point estimate was r_{pb} = .520 (*95% CI* = 0.471-0.566). The Duval-Tweedie trim-and-fill analysis based on a random-effects model indicated three missing studies as well. The adjusted point estimate was r_{pb} = .534 (*95% CI* = 0.441-0.615). In both cases the adjusted point estimate was reduced compared to the unadjusted value. Both analyses indicated three missing studies (almost one fourth of observed studies), which made the existence of publication bias rather plausible. Thus existence of publication bias was assumed.

Retrospective power analysis

Retrospective power analysis was conducted for each of the three meta-analyses separately. The basis for the calculations in each case was once the summary effect resulting from the fixed-effect analysis, and the other time the summary effect resulting from the random-effects analysis. The assumption for power calculations was in all cases a significance level of .05 (two-tailed). Furthermore, given sample sizes were used. As Sparling (1980) is a health-related meta-analysis, the assumed power criterion was .8, which was lowered to .7 (or even .6) when only one study met a criterion of .8.

The summary effect of sex differences in VO_2 max expressed in absolute terms when a fixed-effect model was employed was r_{pb} = .809 (*95% CI* = 0.784-0.832). The mean power of the studies to detect this effect size was .716 with a standard deviation of 0.260 and a minimum of .240 and a maximum of .999. Retrospective power analysis yielded a summary effect of r_{pb} = .801 (*95% CI* = 0.768-0.829). The mean power of the included studies to detect this new effect size was .923 with a standard deviation of 0.047 and a minimum of .862 and a maximum of .999. When a random-effects model was employed the summary effect was r_{pb} = .810 (*95% CI* = 0.771-0.843), which is nearly identical to the above result. The mean power of the studies to detect this effect size was .717 with a standard deviation of 0.259 and a minimum of .240 and a maximum of .999. Retrospective power analysis yielded a summary effect of r_{pb} = .798 (*95% CI* = 0.722-0.854). The mean power of the included studies to detect this new effect size was .922 with a standard deviation of 0.048 and a minimum of .859 and a maximum of .999.

The summary effect of sex differences in VO_2 max expressed relative to body weight when a fixed-effect model was employed was r_{pb} = .729 (*95% CI* = 0.695-0.761). The mean power of the studies to detect this effect size was .654 with a standard deviation of 0.261 and a minimum of .204 and a maximum of 0.998. Retrospective power analysis yielded a summary effect of r_{pb} = .718 (*95% CI* = 0.671-0.759). The mean power of the included studies to detect this new effect size was .884 with a standard deviation of 0.070 and a minimum of .831 and a maximum of .998. When a random-effects model was employed the summary effect was r_{pb} = .720 (*95% CI* = 0.641-0.784). The mean power of the studies to detect this effect size was .646 with a standard deviation of 0.261 and a minimum of .200 and a maximum of .998. Retrospective power analysis yielded a summary effect of r_{pb} = .681 (*95% CI* = 0.502-0.804). The mean power of the included

studies to detect this new effect size was .893 with a standard deviation of 0.094 and a minimum of .810 and a maximum of .996.

The summary effect of sex differences in VO_2 max expressed relative to fat-free weight when a fixed-effect model was employed was r_{pb} = .607(*95% CI* = 0.552-0.657). The mean power of the studies to detect this effect size was .538 with a standard deviation of 0.249 and a minimum of .156 and a maximum of .984. Retrospective power analysis yielded a summary effect of r_{pb} = .594 (*95% CI* = 0.524-0.655). The mean power of the included studies to detect this new effect size was .743 with a standard deviation of 0.161 and a minimum of .619 and a maximum of .980. When a random-effects model was employed the summary effect was r_{pb} = .599 (*95% CI* = 0.524-0.664). The mean power of the studies to detect this effect size was .529 with a standard deviation of 0.248 and a minimum of .153 and a maximum of .981. Retrospective power analysis yielded a summary effect of r_{pb} = .560 (*95% CI* = 0.344-0.720). The mean power of the included studies to detect this new effect size was .745 with a standard deviation of 0.191 and a minimum of .624 and a maximum of .965.

Summary of reanalysis

Results of the reanalysis suggested that the magnitude of sex difference in VO_2 max was largest when VO_2 max was expressed in absolute terms. Taking into account variability in aerobic capacity due to disparities in body size and body fatness, the magnitude of sex difference in VO_2 max was substantially reduced. Fixed-effect and random-effects analyses both showed very similar results. The magnitude of sex difference in VO_2 max differed significantly between trained and untrained groups only, when VO_2 max was expressed relative to fat-free weight and a fixed-effect model was employed. Analyses concerning publication bias performed for each of the three meta-analyses separately showed evidence of publication bias for meta-analyses where VO_2 max was expressed in absolute terms and relative to fat-free weight; whereas no publication bias was found for the meta-analysis where VO_2 max was expressed relative to body weight. Interestingly, the respective adjusted point estimates (for when VO_2 max was expressed in absolute terms and relative to fat-free weight) suggested a smaller magnitude of sex difference compared to the unadjusted estimates. However, this did not cause a change of the overall conclusion that taking into account body size and body fatness substantially reduced the magnitude of sex difference in VO_2 max. Applying retrospective power

analysis attenuated the originally obtained summary effects (with a minimum reduction of 0.008 and a maximum reduction of 0.039), regardless of whether fixed-effect or random-effects analysis was employed and regardless of the expression of VO_2 max; yet the overall conclusion did not change.

In Appendix C, Table C1, the results of the reanalysis and the corresponding original meta-analysis conducted by Sparling (1980) were contrasted.

3.2.3 Reanalysis of Desilva, Hennekens, Lown, and Casscells (1981)

Summary of the original meta-analysis:

Desilva et al. (1981) conducted a meta-analysis to explore whether lignocaine prevented ventricular fibrillation (VF) during acute myocardial infarction. A total of six randomised clinical trials met the criteria for inclusion, which were the following. First, acute myocardial infarction was present, second, a loading dose of at least 50 mg lignocaine had to be given intravenously and third, for at least 24 hours an infusion of lignocaine of not less than 1 mg/min had to be administered. Further characteristics of the included trials can be found in Table B3, Appendix B, where the original data set that was presented in the article of Desilva et al. (1981) was reproduced. The summary measure in this meta-analysis was a risk ratio (*RR*). To statistically integrate the results of the trials, a summary relative risk and corresponding 95% confidence limits were calculated applying the Mantel-Haenszel method. Pooling the results of all six trials, the summary effect was *RR* = 0.53 (*95% CI* = 0.28-0.98). Two trials (Mogensen, 1970; O'Brien, Taylor, & Croxson, 1973) treated patients with left ventricular failure (heart failure) and shock, whereas the other four trials excluded those patients. When the two mentioned trials were excluded, the summary effect was *RR* = 0.22 (*95% CI* = 0.09-0.55). The results indicated a significant treatment effect of lignocaine in preventing ventricular fibrillation. This effect was even greater and again significant when the two mentioned trials were excluded (Desilva et al., 1981).

The original data set presented in the article of Desilva et al. (1982) is reproduced in Table B3, which can be found in Appendix B. Reanalysis of the original meta-analysis yielded the following results.

Reanalysis

Synthesis of results

When all six trials were pooled using a fixed-effect model, the summary effect was *RR* = 0.780 (*95% CI* = 0.409-1.486, *p* = .449). The closer the result is to zero, the greater is the benefit of lignocaine treatment in preventing VF compared to no treatment. Therefore, the result favoured the treatment group receiving lignocaine, but not significantly. When a random-effects model was employed, the summary effect *RR* = 0.704 (*95% CI* = 0.316-1.567, *p* = .390) even more favoured the treatment group, but again was not significant.

Measures of heterogeneity showed a non-significant (*Q* = 6.608, *df* = 5, *p* = .251) and low (I^2 = 24.340) effect heterogeneity across studies. In Figure D7 and D8 (Appendix D) the according forest plots of the fixed-effect (Figure D7) and random-effects analysis (Figure D8) can be found.

Additional analysis: Sensitivity analysis

In the following the robustness of the finding was assessed. Thus a sensitivity analysis was conducted addressing the same issue as the original meta-analysis.

The exclusion of the two trials which treated patients with left ventricular failure (heart failure) and shock yielded a summary effect of *RR* = 0.630 (*95% CI* = 0.280-1.419, *p* = .265) when a fixed-effect model was employed. Again the summary effect was not significant, but in this case more substantial than when all six trials had been included. It was even more substantial and therefore even more favouring the treatment group when a random-effects model was employed (*RR* = 0.511, *95% CI* = 0.154-1.694, *p* = .272), but as before the result did not reach statistical significance. More substantial in this context means that the risk ratio was closer to zero. The closer the result is to zero, the greater is the benefit of lignocaine treatment in preventing VF compared to no treatment.

Between-study effect heterogeneity was non-significant (*Q* = 5.123, *df* = 3, *p* = .163) and moderate (I^2 = 41.441).

Analyses concerning publication bias

Visual inspection of the funnel plots (Figures E7 and E8, Appendix E), with the risk ratio (in log units) plotted on the X axis and the standard error plotted on the Y axis,

suggested symmetry, as far as only six data points allow for a valid statement. They include observed studies as well as studies imputed by the Duval-Tweedie trim-and-fill method, which was once based on a fixed-effect model (Figure E7), and the other time based on a random-effects model (Figure E8). Begg and Mazumdar's rank correlation method (*p* (2-tailed) = .452) and Egger's regression test (*p* (2-tailed) = .080) were non-significant, indicating no publication bias. These tests have lower power (Borenstein et al., 2009), and therefore the results must be interpreted cautiously. It is further important to keep in mind that the tests were based on only six studies. The Duval-Tweedie trim-and-fill analysis based on a fixed-effect model indicated no missing studies. The Duval-Tweedie trim-and-fill analysis based on a random-effects model also indicated no missing studies. The conclusion drawn from the above was that there was no evidence of publication bias.

Retrospective power analysis

It was intended to calculate a retrospective power analysis with the following assumptions: The basis for the calculations in each case was once the summary effect resulting from the fixed-effect analysis, and the other time the summary effect resulting from the random-effects analysis. The assumption for power calculations was a significance level of .05 (two-tailed). Furthermore, given sample sizes were used. As Desilva et al. (1981) is a health-related meta-analysis the assumed power criterion was .8.

When all six trials were pooled using a fixed-effect model, the summary effect was *RR* = 0.780 (*95% CI* = 0.409-1.486). The mean power of the studies to detect this effect size was .170 with a standard deviation of 0.071 and a minimum of .103 and a maximum of .264. When a random-effects model was employed, the summary effect was *RR* = 0.704 (*95% CI* = 0.316-1.567, *p* = .390). The mean power of the studies to detect this effect size was .291 with a standard deviation of 0.139 and a minimum of .160 and a maximum of .471.

As can be seen from the above results, the maximum power for the fixed-effect and random-effects analysis was .264 and .471 respectively, meaning that not a single study met the power criterion of .8. All studies in the meta-analysis were so weak that no further analysis could be undertaken.

Summary of reanalysis

Results of the above reanalysis suggested that there was a benefit of lignocaine treatment in preventing ventricular fibrillation. This effect was even greater when the two studies that treated patients with left ventricular failure (heart failure) and shock were excluded. Generally random-effects analyses yielded summary effects more substantial (closer to zero) than the respective fixed-effect analyses. It is important to add that none of the results reached statistical significance. Power analysis found that all studies in the meta-analysis had low statistical power to detect the respective summary effects. Furthermore, there was no evidence of publication bias.

In Appendix C, Table C2, the results of the reanalysis and the corresponding original meta-analysis conducted by Desilva et al. (1981) were contrasted.

3.2.4 Reanalysis of Stampfer, Goldhaber, Yusuf, Peto, and Hennekens (1982)

Summary of the original meta-analysis

The meta-analysis conducted by Stampfer et al. (1982) was concerned with the "effect of intravenous streptokinase on acute myocardial infarction" (p. 1180). Mortality results from a total of eight published randomised trials were therefore examined. Participants were randomised to treatment or control group. The treatment group received intravenous streptokinase whereas the control group received either placebo or anticoagulation. The end point was mortality with a follow-up period of 40 days. Moreover, trials had to have a similar treatment protocol. The loading dose had to be uniform, which had to be followed by a continuous infusion therapy and therapy had to be initiated within 24 hours of onset of symptoms. Further characteristics of the included trials can be found in Table B4, Appendix B, where the original data set that was presented in the article of Stampfer et al. (1982) was reproduced. The summary measure in this meta-analysis was a risk ratio (*RR*), defined by Stampfer et al. (1982) as follows: "The proportion of deaths in streptokinase-treated patients divided by that among the controls" (p. 1180). To integrate the results, a weighted average of the risk ratios of the respective trials was calculated. The weight assigned to each trial was the inverse of the variance of

the respective risk ratio. Furthermore, a χ^2 test of heterogeneity was applied. The result was not statistically significant (p = .20), therefore a uniform effect was assumed. The weighted average of the risk ratios derived from the included trials was *RR* = 0.80 (*95% CI* = 0.68-0.95, p = .01). When two trials (Amery, Roeber, Vermeulen, & Verstraete, 1969; Heikinheimo et al., 1971) were excluded based on protocols differing from the others, the weighted average of the risk ratios of the remaining six trials was *RR* = 0.74 (*95% CI* = 0.62-0.89, p = .001). Participants of four trials were from coronary-care units (CCU; Aber et al., 1976; Bett et al., 1973; Dioguardi et al., 1971; European Cooperative Study Group For Streptokinase Treatment In Acute Myocardial Infarction, 1979). The trials reported not only risk ratios for the early weeks, but three (Aber et al., 1976; Bett et al., 1973; European Cooperative Study Group For Streptokinase Treatment In Acute Myocardial Infarction, 1979) out of the four trials also reported risk ratios at six (three in case of Bett et al.,1973) months. When risk ratios reported for the early weeks from all of the four trials were pooled, the result was *RR* = 0.85 (*95% CI* = 0.66-1.10, p = .23). When risk ratios reported for longer follow-up periods were pooled, the result was *RR* = 0.71 (*95% CI* = 0.56-0.91, p = .008). Generally the results suggested that mortality was reduced when intravenous streptokinase therapy was applied after acute myocardial infarction (Stampfer et al., 1982).

The original data set presented in the article of Stampfer et al. (1982) is reproduced in Table B4, which can be found in Appendix B. Reanalysis of the original meta-analysis yielded the following results.

Reanalysis

Synthesis of results

The statistical integration of all eight trials employing a fixed-effect model yielded a summary effect of *RR* = 0.805 (*95% CI* = 0.684-0.947, p = .009), significantly favouring the treatment group receiving streptokinase. When a random-effects model was employed, the summary effect was attenuated and lost significance (*RR* = 0.828, *95% CI* = 0.661-1.037, p = .100). The test of effect heterogeneity across studies was non-significant (Q = 11.775, df = 7, p = .108) and the I^2 index (I^2 = 40.55) was moderate. In Figure D9 and D10 (Appendix D) the forest plots for the fixed-effect (Figure D9) and random-effects analysis (Figure D10) can be found.

Additional analyses: Sensitivity analyses

In the following the robustness of the findings was assessed. Thus three sensitivity analyses were conducted addressing the same issues as the original meta-analysis.

When two trials (Amery, Roeber, Vermeulen, & Verstraete, 1969; Heikinheimo et al., 1971) were excluded because of differing protocols, the statistical integration of the remaining six trials employing a fixed-effect model yielded *RR* = 0.743 (*95% CI* = 0.623-0.886, *p* = .001). When a random-effects model was employed, the summary effect was *RR* = 0.745 (*95% CI* = 0.605-0.916, *p* = .005). Between-study effect heterogeneity was non-significant (Q = 6.293, *df* = 5, *p* = .279) and lower (I^2 = 20.547) than when all eight trials had been included. The obtained summary effects were significant in both analyses and even more favouring the treatment group than when all eight trials had been included. Moreover, exclusion of the two differing trials diminished effect heterogeneity across studies.

When risk ratios reported for the early weeks from the four trials that included only participants from coronary-care units (CCU) were pooled, employing a fixed-effect model, the result was *RR* = 0.857 (*95% CI* = 0.667-1.100, *p* = .226). When a random-effects model was employed, the result was identical to the result obtained when a fixed-effect model had been employed. The fixed-effect model is mathematically a special case of the random-effects model and if T^2, the between-studies variance, is zero, the models yield identical estimates (Borenstein et al., 2010). The between-studies variance (T^2 = 0.00) actually was zero in this case. Also the I^2 index (I^2 = 0.00) and *Q statistic* (Q = 2.135, *df* = 3, *p* = .545) conformed to the above. The obtained results showed that the four trials were homogenous as indicated by the I^2 index and between-studies variance. The treatment group was still favoured, but less than when all eight trials had been included; additionally the obtained result was non-significant.

When risk ratios reported for longer follow-up periods from three out of the four CCU-trials were pooled, employing a fixed-effect model, the result was *RR* = 0.722 (*95% CI* = 0.571-0.913, *p* = .006). When a random-effects model was employed, the result was *RR* = 0.714 (*95% CI* = 0.502-1.015, *p* = .061). Measures of heterogeneity showed a non-significant (Q = 4.360, *df* = 2, *p* = .113) and moderate (I^2 = 54.133) effect heterogeneity across studies. The fixed-effect analysis yielded a summary effect significantly favouring the treatment group, which was more substantial than all summary effects obtained before.

The random-effects analysis yielded an even more substantial summary effect favouring the treatment group, but this effect was non-significant. When compared to the analyses above, the major distinction to the present analysis was the length of the follow-up period.

Analyses concerning publication bias

Visual inspection of the funnel plots (Figures E9 and E10, Appendix E), with the risk ratio (in log units) plotted on the X axis and the standard error plotted on the Y axis, suggested asymmetry, as far as eight data points allow for a valid statement. They include observed studies as well as studies imputed by the Duval-Tweedie trim-and-fill method, which was once based on a fixed-effect model (Figure E9), and the other time based on a random-effects model (Figure E10). Begg and Mazumdar's rank correlation method (*p* (2-tailed) = .386) and Egger's regression test (*p* (2-tailed) = .423) were non-significant, indicating no publication bias. These tests have lower power (Borenstein et al., 2009), therefore the results must be interpreted cautiously. Furthermore, it is important to keep in mind that the tests were based on only eight studies. The Duval-Tweedie trim-and-fill analysis based on a fixed-effect model indicated two missing studies. The adjusted point estimate was *RR* = 0.744 (*95% CI* = 0.639-0.866). The Duval-Tweedie trim-and-fill analysis based on a random-effects model indicated two missing studies as well. The adjusted point estimate was *RR* = 0.746 (*95% CI* = 0.591-0.941). In both cases the adjusted point estimate was reduced (more favouring the treatment group) compared to the unadjusted value. Both analyses indicated two missing studies (one fourth of observed studies), which made the existence of publication bias rather plausible. The conclusion drawn from the above was that there was evidence of publication bias.

Retrospective power analysis

It was intended to calculate a retrospective power analysis with the following assumptions: The basis for the calculations in each case was once the summary effect resulting from the fixed-effect analysis, and the other time the summary effect resulting from the random-effects analysis. The assumption for power calculations was a significance level of .05 (two-tailed). Furthermore, given sample sizes were used. As Stampfer et al. (1982) is a health-related meta-analysis the assumed power criterion was .8.

When all eight trials were pooled using a fixed-effect model, the summary effect was *RR* = 0.805 (*95% CI* = 0.684-0.947). The mean power of the studies to detect this effect size was .279 with a standard deviation of 0.108 and a minimum of .142 and a maximum of .453. When a random-effects model was employed, the summary effect was *RR* = 0.828 (*95% CI* = 0.661-1.037). The mean power of the studies to detect this effect size was .233 with a standard deviation of 0.09 and a minimum of .123 and a maximum of .376.

As can be seen from the above results, the maximum power for the fixed-effect and random-effects analysis was .453 and .376 respectively, meaning that not a single study met the power criterion of .8. All studies in the meta-analysis were so weak that no further analysis could be undertaken.

Summary of reanalysis

Generally, the results favoured the treatment group receiving streptokinase and thus suggested benefitting effects of intravenous streptokinase therapy applied after acute myocardial infarction. Sensitivity analyses showed that two trials with differing protocols had a great impact on the overall result. Exclusion of these two trials yielded a more substantial summary effect (*RR* closer to zero). Furthermore, statistical integration of the results from the four CCU-trials reported for the early weeks yielded a summary effect less favouring the treatment group than when all eight trials had been included. However, when results reported for longer follow-up periods from three CCU-trials were integrated, the treatment group was even more favoured than indicated by any other result. Overall, fixed-effect and random-effects analyses yielded very similar summary effects, with two exceptions. When all eight trails were pooled, the fixed-effects analysis yielded a more substantial, and significant, summary effect (*RR* closer to zero). When risk ratios reported for longer follow-up periods from three out of the four CCU-trials were pooled, the fixed-effect analysis yielded a significant, but slightly less substantial summary effect than the random-effects analysis, whose summary effect was thus more substantial (*RR* closer to zero), but non-significant. Assessment of publication bias provided evidence of publication bias. Interestingly, the adjusted point estimates suggested a more substantial summary effect (*RR* closer to zero). Power analysis found that all studies in the meta-analysis had low statistical power to detect the respective summary effects.

In Appendix C, Table C3, the results of the reanalysis and the corresponding original meta-analysis conducted by Stampfer et al. (1982) were contrasted.

Discussion

4 Discussion

The present work sought to clarify whether it is possible to reanalyze meta-analyses published between 1977 and 1982 using state-of-the-art methods. A literature search for classic meta-analyses reported in English language and published in peer-reviewed journals between 1977 and 1982 yielded a total of 102 articles. Requirements needed to be fulfilled for reanalysis using modern methods were established; four criteria were determined a priori and an additional fifth criterion was determined after all articles had been screened. A total of 99 articles had to be excluded as a result of not meeting all five criteria. Thus three classic meta-analyses remained and were then reanalyzed with the intention of answering the original research questions using state-of-the-art methods.

4.1 Comparison of reanalyses and original meta-analyses

Reanalysis of the meta-analysis conducted by Sparling (1980) yielded results only partially consistent with the original meta-analysis. Based on summary effects that were very similar to the ones obtained in the original meta-analysis, it was likewise concluded that sex difference in VO_2 max could be substantially reduced when taking into account variability in aerobic capacity due to disparities in body size and body fatness. This conclusion could be drawn regardless of whether a fixed-effect or random-effects model was employed. In contrast to the original meta-analysis was the conclusion that was drawn concerning the comparison of sex differences in trained and untrained groups. The original meta-analysis concluded that the magnitude of sex differences was similar for trained and untrained groups when VO_2 max was expressed relative to fat-free weight, but found that the magnitude of sex differences was smaller for trained than untrained groups when VO_2 max was expressed in absolute terms or relative to body weight. Reanalysis on the other hand showed that the magnitude of sex differences in VO_2 max differed significantly between trained and untrained groups only when VO_2 max was expressed relative to fat-free weight (but only when employing a fixed-effect model). No significant differences between trained and untrained groups were found when VO_2 max was expressed in absolute terms or relative to body weight. Methods to assess differences between trained and untrained groups were not the same in the original meta-analysis and reanalysis.

Additional findings concern publication bias, which was not examined by the original meta-analysis. Evidence of publication bias was found for the meta-analyses where VO_2 max was expressed in absolute terms and relative to fat-free weight, whereas no evidence of publication bias was found for the meta-analysis where VO_2 max was expressed relative to body weight. Even when considering the adjusted point estimates, which were all slightly smaller than the unadjusted values, the overall conclusion did not change. In addition, applying retrospective power analysis did not change the overall conclusion. Therefore the results of the reanalysis suggested the same conclusion as drawn from the original meta-analysis concerning the reduction of sex difference in VO_2 max when taking into account body weight and body fatness. A conclusion contrary to the original meta-analysis was suggested by the present reanalysis when it comes to the comparison of sex differences in VO_2 max between trained and untrained groups.

The meta-analysis conducted by Desilva et al. (1981) and its reanalysis yielded consistent conclusions with important constrictions. Summary effects of both the original meta-analysis and the reanalysis suggested that there was a benefit of lignocaine treatment in preventing ventricular fibrillation. Both analyses agreed that this effect was even greater when the two studies that treated patients with left ventricular failure (heart failure) and shock were excluded. Still, there were two important differences between the original meta-analysis and its reanalysis. The summary effects calculated by Desilva et al. (1981) were favouring the treatment group to a greater extent than the summary effects calculated in the present reanalysis, regardless of whether a fixed-effect or random-effects model was employed. Furthermore, Desilva et al. (1981) found these summary effects to be significant, whereas in the present reanalysis no summary effect was found to be significant. Therefore, the results of the present reanalysis suggest treatment effects that are smaller and moreover non-significant compared to the results of the original meta-analysis. Power analysis in the course of reanalysis revealed that all studies in the meta-analysis had low statistical power to detect the respective summary effects. Moreover, the present reanalysis found no evidence of publication bias.

Reanalysis of the meta-analysis conducted by Stampfer et al. (1982) yielded results largely consistent with the original meta-analysis. Reanalysis supported the overall conclusion of the original meta-analysis that mortality was reduced when intravenous streptokinase therapy was applied after acute myocardial infarction. Between-study effect heterogeneity of the eight trials was non-significant in both, the original meta-analysis and reanalysis. Statistical integration of these eight trials yielded a significant summary effect

(nearly) identical to the original meta-analysis when a fixed-effect analysis was employed. This effect was attenuated and lost significance when a random-effects model was employed. The three sensitivity analyses yielded summary effects that were very similar comparing the respective fixed-effect and random-effects analyses. Moreover, these summary effects were very similar to the respective summary effects of the original meta-analysis. Furthermore, if summary effects of the sensitivity analyses were significant in the original meta-analysis, the corresponding summary effects of the reanalysis were significant as well. The same applied to non-significant summary effects respectively. The only exception to that concerned the statistical integration of the three CCU-trials reporting results for longer follow-up periods. Here the summary effects were still very similar altogether, but significant in the original and fixed-effect analysis and non-significant in the random-effects analysis. Additional analyses concerning publication found evidence of publication bias. Nevertheless the adjusted point estimates, being even closer to zero than the unadjusted values, as well supported the conclusion of benefitting effects of intravenous streptokinase therapy applied after acute myocardial infarction. Power analysis in the course of reanalysis revealed that all studies in the meta-analysis had low statistical power to detect the respective summary effects.

4.2 Feasibility and results of reanalysing classic meta-analyses

The intention of the present work was to find out whether reanalysis of classic meta-analyses using state-of-the-art methods was at all possible, and further whether new or different conclusions could be drawn from the original data using modern methods. With the obtained results at hand, answering the posed research questions can now be approached.

Reanalysis of classic meta-analyses using state-of-the-art methods was possible if certain prerequisites were met. These concern the following criteria. First and foremost the study to be reanalyzed needed to be a (classic) meta-analysis and in the case of the present work, it needed to be a meta-analysis of effect sizes. Furthermore, an effect size needed to be reported, or had to be computable from presented information, for each included original study. Those reported effect sizes needed to be common today or at least possible to enter into spreadsheets of the software intended to execute calculations. In addition,

variance (or standard error) for each included study or information to calculate variance, had to be reported. Beyond that, the quality of reporting had a great impact on the decision whether reanalysis was possible or not.

Reanalysis of classic meta-analyses using state-of-the-art methods could yield results that were different and results that were new, but also results that were consistent with the results of the original meta-analysis. Therefore conclusions drawn from the original data using modern methods could extend, contradict or confirm conclusions drawn by the original meta-analysis. If a more general answer in terms of rates indicating if the majority of conclusions drawn by the original meta-analyses were confirmed or rather contradicted was intended, an amount of reanalyses by far larger than three would be required.

4.3 Relation of results to previous research

Only three classic meta-analyses could be reanalyzed using state-of-the-art methods. If only meta-analyses of effect sizes were considered, only three out of 78 classic meta-analyses were eligible for reanalysis, which is equivalent to about 4%. This small amount of meta-analyses appropriate for reanalysis using state-of-the-art methods was to be expected and is consistent with other research. Schmidt et al. (2009) reanalyzed five meta-analytic studies that had been published in *Psychological Bulletin* between 1988 and 2006 using state-of-the-art methods. One of the inclusion criteria was that meta-analyses presented data tables containing data on included original studies (effect sizes, sample size and other information necessary for coding of studies). Schmidt et al. (2009) were surprised to find only few meta-analyses meeting this criterion. Only around 3% (5 out of 169) of the meta-analyses screened for inclusion were finally identified as appropriate. Therefore, Schmidt et al. (2009) encountered the very same problem when trying to reanalyze meta-analyses as was encountered in the present work. It is remarkable that after the period of classic meta-analyses and apparently up until the year 2006 data on studies included for meta-analysis was still not entirely reported. There is evidence that the same problem occurred in another field other than psychology, as the study conducted by Jennions and Möller (2002) demonstrated. Jennions and Möller (2002) examined 44 meta-analyses of the field of biology to assess the relationship of magnitude of effect size and year of publication. Initially 81 meta-analyses seemed to be eligible for inclusion. Not

providing effect sizes for original studies was one of the exclusion criteria. A total of 37 meta-analyses was excluded, thus nearly every second meta-analysis (46%, 37 out of 81) had to be excluded.

Results obtained in the present work concerning the quality of reporting data for primary research studies included in a meta-analysis, and related to that the exclusion of a large amount of meta-analyses, seem to be aligned with other research. The development of reporting guidelines for reporting meta-analyses in recent years, such as the PRISMA Standards (Liberati et al., 2009) and MARS (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008), represents a crucial step to enhance reporting quality and thus to enable reanalysis. The present work as well as other research such as Schmidt et al. (2009) showed that reporting guidelines specifying the items of a meta-analysis that ought to be reported are of great importance.

Results obtained through reanalysis of each of the three classic meta-analyses showed, as already described above, that conclusions drawn from the original data using modern methods could extend, contradict or confirm conclusions drawn by the original meta-analysis. Other reanalyses or replication studies of classic meta-analyses as well showed results confirming, contradicting or extending the conclusions of the original meta-analysis. Landman and Dawes (1982) as well as Shapiro and Shapiro (1982) replicated the meta-analysis conducted by Smith and Glass (1977). Other than in the present work, meta-analytic calculations were in each case based on a sample of studies different to that of the original meta-analysis and moreover statistical procedures applied were similar to those of the original meta-analysis. Nevertheless both studies assessed the same research question as in the original meta-analysis besides additional questions. Both replication studies supported Smith and Glass's (1977) conclusion and thus confirmed it. Furthermore, Shapiro and Shapiro (1982) found that behavioral and cognitive methods were superior, whereas dynamic and humanistic methods were found to be inferior. Smith and Glass (1977) however did not find a difference in effectiveness between behavioral and nonbehavioral therapies. Hence, the result obtained by Shapiro and Shapiro (1982) is contrary to that of Smith and Glass (1977) and therefore contradicting it. Rosenthal and Rubin (1982) in turn reanalyzed the original data set of the meta-analysis conducted by Hyde (1982) applying advanced statistical techniques. Results extended the conclusions drawn by the original meta-analysis.

4.4 Limitations and future directions

Limitations of the present work were already indicated further up. Only three classic meta-analyses could be reanalyzed using state-of-the-art methods. This circumstance raises two questions and one problem. First, it is to be asked whether the applied search strategy was apt to discover all meta-analyses published in peer-reviewed journals between 1977 and 1982. The reason for so few classic meta-analyses meeting the criteria required for reanalysis could be seen in an incomplete list of all meta-analyses that could potentially have been taken into account. Second question concerns the criteria established to decide whether a study was eligible for reanalysis or not. Were criteria appropriate to filter out all classic meta-analyses that could have been reanalyzed using state-of-the-art methods? Attention was paid that criteria were just as strict as to exclude classic meta-analyses not meeting the basic requirements necessary for reanalysis. The problem mentioned above addresses the possibility of generalisation of findings. Since only three classic meta-analyses could be reanalyzed, no generalisation could be made concerning the question whether conclusions originally drawn by classic meta-analyses had been predominantly confirmed or rather contradicted through reanalysis using state-of-the-art methods. To generalise results, a larger (and more representative) amount of classic meta-analyses ought to be reanalyzed.

Extending the sample of meta-analyses eligible for reanalysis within the present circumstances is difficult because many classic meta-analyses did not report enough data to be reanalyzed. One possibility is to detect a crucial amount of published classic meta-analyses fulfilling requirements for reanalysis that has not yet discovered. Other than that, it might be suggested having a closer look at the definition of the period of classic meta-analyses. The end of the period of classic meta-analyses was in the present work set at the year 1982. One reason for that was the presentation of a random-effects model for the analysis of effect sizes by Hedges in the year 1983. Moreover, it was stated that a characteristic of classic meta-analyses was the absence of forest and funnel plot, which have not been employed until later. The funnel plot was introduced in 1984 by Light and Pillemer and a precursor of today's forest plot appeared in 1989 in a book of Hedges and Olkin. A suggestion for future research is to extend the period of classic meta-analyses until random-effects models were actually employed (and not only theoretically introduced) and forest and funnel plots were actually printed in published meta-analyses. Hence one can assume that more meta-analyses could be reanalyzed and more conclusive statements could be made about the obtained results.

Notwithstanding these additional considerations, the present work represents a first step that was taken to reanalyze early meta-analyses with the intention of finding out whether the application of today's statistical methods on reported data yielded new and different insights. Further research could continue the present work by extending the period of classic meta-analyses as suggested above and examine whether and, if so, when the quality of reporting meta-analyses changed significantly. Meta-analyses published during the extended period that meet requirements necessary for reanalysis could be reanalyzed using state-of-the-art methods. In terms of time the present work could also be continued in the opposed direction. Studies employing precursors of meta-analytical techniques published before 1977 could as well be examined concerning quality of reporting. If possible these studies could also be reanalyzed using state-of-the-art methods. Therefore the present work could be seen as a basis and starting point for a series of investigations along the lines of thought indicated here.

Appendices

Appendices

Appendix A: List of articles screened for inclusion

A list of all articles and the corresponding reasons for exclusion or inclusion in alphabetical order for every year of publication, starting chronologically backwards

Nr.	Study	Reasons for exclusion/**inclusion**
1	Chalmers, T. C., Matta, R. J., Smith, H., & Kunzler, A.-M. (1977). Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *New England Journal of Medicine, 297,* 1091-1096.	Meta-analysis using a different approach than that of effect sizes.
2	Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752-760.	Effect sizes were not reported or not computable from presented information for each included original study.
3	Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85,* 845-857.	Missing information concerning variance.
4	Kennedy, M. M. (1978). Findings from the follow through planned variation study. *Educational Researcher, 7,* 3-11.	No meta-analysis was conducted.
5	Rosenthal. R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3,* 377-415.	Effect sizes were not reported or not computable from presented information for each included original study.
6	Tittle, C. R., Villemez, W. J., & Smith, D. A. (1978). The myth of social class and criminality: An empirical assessment of the empirical evidence. *American Sociological Review, 43,* 643-656.	Effect sizes were not reported or not computable from presented information for each included original study.
7	Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology, 37,* 131-146.	Meta-analysis using a different approach than that of effect sizes.
8	Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis, 1,* 2-16.	Specific problems obstructing reanalysis (illegible version of the related report containing the required data).
9	Hereford, S. M. (1979). The Keller Plan within a conventional academic environment: An empirical "meta-analytic" study. *Engineering Education, 70,* 250-260.	No meta-analysis was conducted.

10	Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721-735.	Meta-analysis using a different approach than that of effect sizes.
11	Kazrin, A., Durac, J., & Agteros, T. (1979). Meta-meta analysis: A new method for evaluating therapy outcome. *Behaviour Research and Therapy, 17,* 397-399.	No meta-analysis was conducted.
12	Kulik, J. A., Kulik, C.-L. C., & Cohen, P. A. (1979). A meta-analysis of outcome studies of Keller's personalized system of instruction. *American Psychologist, 34,* 307-318.	Effect sizes were not reported or not computable from presented information for each included original study.
13	Kulik, J. A., Kulik, C.-L. C., & Cohen, P. A. (1979). Research on audio-tutorial instruction: A meta-analysis of comparative studies. *Research in Higher Education, 11,* 321-341.	Effect sizes were not reported or not computable from presented information for each included original study.
14	Schmidt, F. L., Hunter, J. E., Pearlman, K., & Shane, G. S. (1979). Further tests of the Schmidt-Hunter bayesian validity generalization procedure. *Personnel Psychology, 32,* 257–281.	Meta-analysis using a different approach than that of effect sizes.
15	Schwab, D. P., Olian-Gottlieb, J. D., & Heneman, H. G. (1979). Between-subjects expectancy theory research: A statistical review of studies predicting effort and performance. *Psychological Bulletin, 86,* 139-147.	Effect sizes were not reported or not computable from presented information for each included original study.
16	Uguroglu, M. E., & Walberg, H. J. (1979). Motivation and achievement: A quantitative synthesis. *American Educational Research Journal, 16,* 375-389.	Effect sizes were not reported or not computable from presented information for each included original study.
17	Andrews, G., Guitar, B., & Howie, P. (1980). Meta-analysis of the effects of stuttering treatment. *Journal of Speech & Hearing Disorders, 45,* 287-307.	Effect sizes were not reported or not computable from presented information for each included original study.
18	Anonymous (1980). Aspirin after myocardial infarction. *Lancet, 1,* 1172-1173.	Meta-analysis using a different approach than that of effect sizes.
19	Arkin, R. M., Cooper, H. M., & Kolditz, T. A. (1980). A statistical review of the literature concerning the self-serving attribution bias in interpersonal influence situations. *Journal of Personality, 48,* 435–448.	Missing information concerning variance.
20	Blanchard, E. B., Andrasik, F., Ahles, T. A., Teders, S. J., & O'Keefe, D. (1980). Migraine and tension headache: A meta-	Meta-analysis using a different approach than that of effect sizes.

	analytic review. *Behavior Therapy, 11,* 613-631.	
21	Bredderman, T. (1980). Process curricula in elementary school service. *Evaluation in Education, 4,* 43-44.	Effect sizes were not reported or not computable from presented information for each included original study.
22	Carlberg, C., & Kavale, K. (1980). The efficacy of special versus regular class placement for exceptional children: A meta-analysis. *Journal of Special Education, 14,* 295-309.	Effect sizes were not reported or not computable from presented information for each included original study.
23	Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13,* 321-341.	Effect sizes were not reported or not computable from presented information for each included original study.
24	Haertel, G. D., Walberg, H. J., & Haertel, E. H.(1980). Classroom socio-psychological environment. *Evaluation in Education, 4,* 113-114.	Effect sizes were not reported or not computable from presented information for each included original study.
25	Hartley, S. S. (1980). Instruction in mathematics. *Evaluation in Education, 4,* 56-57.	Effect sizes were not reported or not computable from presented information for each included original study.
26	Hearold, S. L. (1980). Television and social behaviour. *Evaluation in Education, 4,* 94-95.	Effect sizes were not reported or not computable from presented information for each included original study.
27	Kavale, K. (1980). Auditory-visual integration and its relationship to reading achievement: A meta-analysis. *Perceptual and Motor Skills, 51,* 947-955.	Effect sizes were not reported or not computable from presented information for each included original study.
28	Kulik, C. -L. C., Kulik, J. A., & Cohen, P. A. (1980). Instructional technology and college teaching. *Teaching of Psychology, 7,* 199-205.	Effect sizes were not reported or not computable from presented information for each included original study.
29	Kulik, J. A., Cohen, P. A., & Ebeling, B. J. (1980). Effectiveness of programmed instruction in higher education: A meta-analysis of findings. *Educational Evaluation and Policy Analysis, 2,* 51-64.	Effect sizes were not reported or not computable from presented information for each included original study.
30	Kulik, J. A., Kulik, C.-L. C., & Cohen, P. A. (1980). Effectiveness of computer-based college teaching: A meta-analysis of findings. *Review of Educational Research, 50,* 525-544.	Effect sizes were not reported or not computable from presented information for each included original study.

31	Ladas, H. (1980). Summarizing research: A case study. *Review of Educational Research, 50,* 597-624.	No meta-analysis was conducted.
32	Levy, S. R., Iverson, B. K., & Walberg, H. J. (1980). Nutrition-education research: An interdisciplinary evaluation and review. *Health Education quarterly, 7,* 107-126.	An effect size not common today was employed.
33	Luiten, J., Ames, W., & Ackerson, G. (1980). A meta-analysis of the effects of advance organizers on learning and retention. *American Educational Research Journal, 17,* 211-218.	Effect sizes were not reported or not computable from presented information for each included original study.
34	Maccoby, E. E., & Jacklin, C. N. (1980). Sex differences in aggression: A rejoinder and reprise. *Child Development, 51,* 964-980.	Meta-analysis using a different approach than that of effect sizes.
35	Miller, T. I. (1980). Drug therapy for psychological disorders. *Evaluation in Education, 4,* 96-97.	Effect sizes were not reported or not computable from presented information for each included original study.
36	Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *Journal of Applied Psychology, 65,* 373-406.	Meta-analysis using a different approach than that of effect sizes.
37	Peterson, P. L. (1980). Open versus traditional classrooms. *Evaluation in Education, 4,* 58-60.	Effect sizes were not reported or not computable from presented information for each included original study.
38	Pflaum, S. W., Walberg, H. J., Karegianes, M. L., & Rasher, S. P. (1980). Reading instruction: A quantitative analysis. *Educational Researcher, 9,* 12-18.	Effect sizes were not reported or not computable from presented information for each included original study.
39	Posavac, E. J. (1980). Evaluations of patient education programs: A meta-analysis. *Evaluation & the Health Professions, 3,* 47-62.	Effect sizes were not reported or not computable from presented information for each included original study.
40	Schmidt, F. L., Gast-Rosenberg, I., & Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology, 65,* 643-661.	Meta-analysis using a different approach than that of effect sizes.
41	Smith, M. L. (1980). Sex bias in counseling and psychotherapy, *Psychological Bulletin, 87,* 392-407.	Missing information concerning variance.
42	Smith, M. L., & Glass, G. V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal,*	Effect sizes were not reported or not computable from presented information for each included original study.

	17, 419-433.	
43	Smith, D. A., & Visher, C. A. (1980). Sex and involvement in deviance/crime: A quantitative review of the empirical literature. *American Sociological Review, 45,* 691-701.	Effect sizes were not reported or not computable from presented information for each included original study.
44	Sparling, P. B. (1980). A meta-analysis of studies comparing maximal oxygen uptake in men and women. *Research Quarterly for Exercise and Sport, 51,* 542-552.	**All requirements met.**
45	Andrews, G., & Harvey, R. (1981). Does psychotherapy benefit neurotic patients? A reanalysis of the Smith, Glass, and Miller data. *Archives of General Psychiatry, 38,* 1203-1208.	Effect sizes were not reported or not computable from presented information for each included original study.
46	Baum, M. L., Anish, D. S., Chalmers, T. C., Sacks, H. S., Smith, H., Jr., Fagerstrom, R. M. (1981). A survey of clinical trials of antibiotic prophylaxis in colon surgery: Evidence against further use of no-treatment controls. *New England Journal of Medicine, 305,* 795-799.	Effect sizes were not reported or not computable from presented information for each included original study.
47	Berk, A. A., & Chalmers, T. C. (1981). Cost and efficacy of the substitution of ambulatory for inpatient care. *New England Journal of Medicine, 304,* 393-397.	No meta-analysis was conducted.
48	Boulanger, F. D. (1981a). Ability and science learning: A quantitative synthesis. *Journal of Research in Science Teaching, 18,* 113-121.	Effect sizes were not reported or not computable from presented information for each included original study.
49	Boulanger, F. D. (1981b). Instruction and science learning: A quantitative synthesis. *Journal of Research in Science Teaching, 18,* 311–327.	Effect sizes were not reported or not computable from presented information for each included original study.
50	Burger, J. M. (1981). Motivational biases in the attribution of responsibility for an accident: A meta-analysis of the defensive-attribution hypothesis. *Psychological Bulletin, 90,* 496-512.	Missing information concerning variance.
51	Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51,* 281-309.	Effect sizes were not reported or not computable from presented information for each included original study.
52	Cohen, P. A., Ebeling, B. J., & Kulik, J. A. (1981). A meta-analysis of outcome studies of visual-based instruction. *Educational*	Effect sizes were not reported or not computable from presented information for each included

	Communication & Technology, 29, 26-36.	original study.
53	Cooper, H. M., Burger, J. M., & Good, T. L. (1981). Gender differences in the academic locus of control beliefs of young children. *Journal of Personality and Social Psychology, 40,* 562-572.	Effect sizes were not reported or not computable from presented information for each included original study.
54	Dekkers, J., & Donatti, S. (1981). The integration of research studies on the use of simulation as an instructional strategy. *Journal of Educational Research, 74,* 424-427.	Effect sizes were not reported or not computable from presented information for each included original study.
55	Desilva, R. A., Hennekens, C.H., Lown, B., & Casscells, W. (1981). Lignocaine prophylaxis in acute myocardial infarction: An evaluation of randomised trials. *Lancet, 2,* 855-858.	**All requirements met.**
56	Eagly, A. H., & Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies. *Psychological Bulletin, 90,* 1-20.	Effect sizes were not reported or not computable from presented information for each included original study.
57	Haertel, G. D., Walberg, H. J., & Haertel, E. H. (1981). Socio-psychological environments and learning: A quantitative synthesis. *British Educational Research Journal, 7,* 27-36.	Effect sizes were not reported or not computable from presented information for each included original study.
58	Horak, V. M. (1981). A meta-analysis of research findings on individualized instruction in mathematics. *Journal of Educational Research, 74,* 249-253.	Effect sizes were not reported or not computable from presented information for each included original study.
59	Hyde, J. S. (1981). How large are cognitive gender differences? *American Psychologist, 36,* 892-901.	Missing information concerning variance.
60	Ide, J. K., Parkerson, J., Haertel, G. D., & Walberg, H. J. (1981). Peer group influence on educational outcomes: A quantitative synthesis. *Journal of Educational Psychology, 73,* 472-484.	Effect sizes were not reported or not computable from presented information for each included original study.
61	Johnson, D. W., Maruyama, G., Johnson, R., Nelson, D., & Skon, L. (1981). Effects of cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis. *Psychological Bulletin, 89,* 47-62.	Effect sizes were not reported or not computable from presented information for each included original study.

62	Kavale, K. (1981a). Functions of the Illinois Test of Psycholinguistic Abilities (ITPA): Are they trainable? *Exceptional Children, 47,* 496-510.	An effect size not common today was employed.
63	Kavale, K. (1981b). The relationship between auditory perceptual skills and reading ability: A meta-analysis. *Journal of Learning Disabilities, 14,* 539-546.	Effect sizes were not reported or not computable from presented information for each included original study.
64	Kremer, B. K., & Walberg, H. J. (1981). A synthesis of social and psychological influences on science learning. *Science Education, 65,* 11–23.	Specific problems obstructing reanalysis (reporting).
65	Linn, R. L., Harnisch, D. L., Dunbar, S. B. (1981). Validity generalization and situational specificity: An analysis of the prediction of first-year grades in law school. *Applied Psychological Measurement, 5,* 281-289.	Meta-analysis using a different approach than that of effect sizes.
66	Lysakowski, R. S., & Walberg, H. J. (1981). Classroom reinforcement and learning: A quantitative synthesis. *Journal of Educational Research, 75,* 69-77.	Effect sizes were not reported or not computable from presented information for each included original study.
67	Readence, J. E., & Moore, D. W. (1981). A meta-analytic review of the effect of adjunct pictures on reading comprehension. *Psychology in the Schools, 18,* 218-224.	Effect sizes were not reported or not computable from presented information for each included original study.
68	Redfield, D. L., & Rousseau, E. W. (1981). A meta-analysis of experimental research on teacher questioning behavior. *Review of Educational Research, 51,* 237-245.	Effect sizes were not reported or not computable from presented information for each included original study.
69	Smith, L. R., & Land, M. L. (1981). Low-inference verbal behaviors related to teacher clarity. *Journal of Classroom Interaction, 17,* 37-42.	No meta-analysis was conducted.
70	Strube, M. J. (1981). Meta-analysis and cross-cultural comparison: Sex differences in child competitiveness. *Journal of Cross-Cultural Psychology, 12,* 3-20.	Meta-analysis using a different approach than that of effect sizes.
71	Strube, M. J., & Garcia, J. E. (1981). A meta-analytic investigation of Fiedler's contingency model of leadership effectiveness. *Psychological Bulletin, 90,* 307-321.	Meta-analysis using a different approach than that of effect sizes.
72	Bassoff, E. S., & Glass, G. V. (1982). The relationship between sex roles and mental health: A meta-analysis of twenty-six studies. *The Counseling Psychologist, 10,* 105-112.	Effect sizes were not reported or not computable from presented information for each included original study.

73	Cohen, P. A. (1982). Validity of student ratings in psychology courses: A research synthesis. *Teaching of Psychology, 9,* 78-82.	Effect sizes were not reported or not computable from presented information for each included original study.
74	Cohen, P. A., Kulik, J. A., & Kulik, C.-L. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19,* 237-248.	Effect sizes were not reported or not computable from presented information for each included original study.
75	Cotton, J. L., & Cook, M. S. (1982). Meta-analyses and the effects of various reward systems: Some different conclusions from Johnson et al. *Psychological Bulletin, 92,* 176-183.	No meta-analysis was conducted.
76	Giaconia, R. M., & Hedges, L. V. (1982). Identifying features of effective open education. *Review of Educational Research, 52,* 579-602.	Effect sizes were not reported or not computable from presented information for each included original study.
77	Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement/performance measures. *Review of Educational Research, 52,* 123-142.	Effect sizes were not reported or not computable from presented information for each included original study.
78	Hattie, J. A., & Hansford, B. C. (1982). Self measures and achievement: Comparing a traditional review of literature with a meta-analysis. *Australian Journal of Education, 26,* 71-75.	No meta-analysis was conducted.
79	Inglis, J., & Lawson, J. S. (1982). A meta-analysis of sex differences in the effects of unilateral brain damage on intelligence test results. *Canadian Journal of Psychology, 36,* 670-683.	Meta-analysis using a different approach than that of effect sizes.
80	Iverson, B. K., & Levy, S. R. (1982). Using meta analysis in health education research. *Journal of School Health, 52,* 234-239.	An effect size not common today was employed.
81	Iverson, B. K., & Walberg, H. J. (1982). Home environment and school learning: A quantitative synthesis. *Journal of Experimental Education, 50,* 144-151.	Effect sizes were not reported or not computable from presented information for each included original study.
82	Kavale, K. (1982a). Meta-analysis of the relationship between visual perceptual skills and reading achievement. *Journal of Learning Disabilities, 15,* 42-51.	Effect sizes were not reported or not computable from presented information for each included original study.
83	Kavale, K. (1982b). The efficacy of stimulant drug treatment for hyperactivity: A meta-analysis. *Journal of Learning Disabilities, 15,* 280-289.	Effect sizes were not reported or not computable from presented information for each included original study.
84	Kulik, C.-L. C., & Kulik, J. A. (1982). Effects of ability grouping on secondary school students: A meta-analysis of	Effect sizes were not reported or not computable from presented information for each included

	evaluation findings. *American Educational Research Journal, 19,* 415-428.	original study.
85	Kulik, C.-L. C., Schwalb, B. J., & Kulik, J. A. (1982). Programmed instruction in secondary education: A meta-analysis of evaluation findings. *Journal of Educational Research, 75,* 133-138.	Effect sizes were not reported or not computable from presented information for each included original study.
86	Landman, J. T., & Dawes, R. M. (1982). Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. *American Psychologist, 37,* 504-516.	Effect sizes were not reported or not computable from presented information for each included original study.
87	Lysakowski, R. S., & Walberg, H. J. (1982). Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. *American Educational Research Journal, 19,* 559-578.	Effect sizes were not reported or not computable from presented information for each included original study.
88	Mabe, P. A., West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67,* 280-296.	Meta-analysis using a different approach than that of effect sizes.
89	Mullen, B., & Suls, J. (1982). The effectiveness of attention and rejection as coping styles: A meta-analysis of temporal differences. *Journal of Psychosomatic Research, 26,* 43-49.	Meta-analysis using a different approach than that of effect sizes.
90	Mumford, E., Schlesinger, H. J., & Glass, G. V. (1982). The effects of psychological intervention on recovery from surgery and heart attacks: An analysis of the literature. *American Journal of Public Health, 72,* 141-151.	An effect size not common today was employed.
91	Remmer, A. M., & Jernstedt, G. C. (1982). Comparative effectiveness of simulation games in secondary and college level instruction: A meta-analysis. *Psychological Reports, 51,* 742.	Effect sizes were not reported or not computable from presented information for each included original study.
92	Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin, 92,* 581-604.	Effect sizes were not reported or not computable from presented information for each included original study.
93	Smith, M. L. (1982). What research says about the effectiveness of psychotherapy. *Hospital & Community Psychiatry, 33,* 457-461.	Effect sizes were not reported or not computable from presented information for each included original study.
94	Stampfer, M. J., Goldhaber, S. Z., Yusuf, S., Peto, R., & Hennekens, C. H. (1982). Effect of intravenous streptokinase on acute myocardial infarction: Pooled results from randomized trials. *New England Journal of Medicine, 307,* 1180-1182.	**All requirements met.**

95	Tornatzky, L. G., & Klein, K. J. (1982). Innovation characteristics and innovation adoption- implementation: A meta-analysis of findings. *IEEE Transactions on Engineering Management, 29,* 28-45.	Meta-analysis using a different approach than that of effect sizes.
96	Tran, Z. V., Weltman, A., Glass, G. V, & Mood, D. P. (1982). The effects of exercise on blood lipids and lipoproteins: A meta-analysis of studies. *Medicine and Science in Sports and Exercise, 14,* 110-110.	Effect sizes were not reported or not computable from presented information for each included original study.
97	Wampler, K. S. (1982). Bringing the review of literature into the age of quantification: Meta-analysis as a strategy for integrating research findings in family studies. *Journal of Marriage and the Family, 44,* 1009-1023.	An effect size not common today was employed.
98	Weinstein, T., Boulanger, F. D., & Walberg, H. J. (1982). Science curriculum effects in high school: A quantitative synthesis. *Journal of Research in Science Teaching, 19,* 511-522.	Effect sizes were not reported or not computable from presented information for each included original study.
99	White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91,* 461-481.	Effect sizes were not reported or not computable from presented information for each included original study.
100	Williams, P. A., Haertel, E. H., Haertel, G. D., & Walberg, H. J. (1982). The impact of leisure-time television on school learning: A research synthesis. *American Educational Research Journal, 19,* 19-50.	Specific problems obstructing reanalysis (reporting).
101	Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal, 19,* 249-258.	Effect sizes were not reported or not computable from presented information for each included original study.
102	Wilson, L. B., & Simson, S. (1982). The status of preventive care for the aged: A meta-analysis. *Gerontologist, 22,* 74-74.	Effect sizes were not reported or not computable from presented information for each included original study.

Appendix B: Datasets of the original meta-analyses

Reproduced datasets that had been presented in the articles of the classic meta-analyses

Table B1

Dataset 1 presented in Sparling (1980)

Study	Sex	n	Activity Status	Age years	Height cm	Weight kg	Fat %
Von Dobeln (1956)[1]	M	35	Physical education, students and teachers, Swedish	26.1 ± 4.7	177.9 ± 6.9	69.4 ± 8.2	10.6
	F	34	Physical education, students and teachers, Swedish	22.6 ± 3.3	169.6 ± 3.8	62.8 ± 6.5	20.3
Hermansen and Andersen (1965)[2]	M	14	National-level cross-country skiers, Norwegian	27.7 ± 3.1	174.8 ± 6.3	66.7 ± 5.0	10
	F	5	National-level cross-country skiers, Norwegian	25.1 ± 5.9	169.0 ± 5.7	61.6 ± 6.2	20.9
Cotes, Davies, Edholm, Healy, and Tanner (1969)[3]	M	23	Factory workers, fairly heavy work, British	25	176	71.4	13.9
	F	20	Factory workers, fairly light work, British	23.7	162	55	27.2
MacNab, Conger, and Taylor (1969)[4]	M	24	Physical education/recreation students, Canadian	20.0 ± 1.2	179.3 ± 6.1	76.1 ± 8.8	12.7
	F	24	Physical education/recreation students, Canadian	18.7 ± 0.6	165.8 ± 5.3	59.2 ± 5.9	23.4
Dill, Myhre, Greer, Richardson, and Singleton (1972)[5]	M	11	High school students, American	17.9 ±1.5	181.3 ± 7.6	73.1 ± 16.4	14.5
	F	10	High school students, American	16.7 ± 1.1	166.5 ± 6.3	53.8 ± 8.2	21.7
Davies, Mbelwa, Crockford, and Weiner (1973)[6]	M	62	Activity level not stated, African	22.7 ± 2.7	165.9 ± 6.7	58.0 ± 5.7	11.6
	F	32	Activity level not stated, African	27.0 ±9.5	153.5 ± 5.7	50.1 ± 7.3	26.1

[1] Von Dobeln, W. (1956). Human standard and maximal metabolic rate in relation to fatfree body mass. *Acta Physiologica Scandinavica Supplementum, 37*, 1-79.

[2] Hermansen, L., & Andersen, K. L. (1965). Aerobic work capacity in young Norwegian men and women. *Journal of Applied Physiology, 20*, 425-431.

[3] Cotes, J. E., Davies, C. T. M., Edholm, O. G., Healy, M. J. R., & Tanner, J. M. (1969). Factors relating to the aerobic capacity of 46 healthy British males and females, ages 18 to 28 years. *Proceedings of the Royal Society of London Britain, 174*, 91-114.

[4] MacNab, R. B. J., Conger, P. R., & Taylor, P. S. (1969). Differences in maximal and submaximal work capacity in men and women. *Journal of Applied Physiology, 27*, 644-648.

[5] Dill, D. B., Myhre, L. G., Greer, S. M., Richardson, J. C., & Singleton, K. J. (1972). Body composition and aerobic capacity of youth of both sexes. *Medicine and Science in Sports, 4*, 198-204.

[6] Davies, C. T. M., Mbelwa, D., Crockford, G., & Weiner, J. S. (1973). Exercise tolerance and body composition of male and female Africans aged 18 to 30 years. *Human Biology, 45*, 31-40.

Mayhew (1976)[7]	M	24	High school track athletes, American	16.7 ± 0.9	176.0 ± 4.9	61.2 ± 7.6	8.2
	F	21	High school track athletes, American	16.5 ± 0.8	168. 9 ± 5.9	56.4 ± 5.7	17.2
Dill, Soholt, McLean, Drost, and Loughran (1977)[8]	M	14	High school athletes, American	-	178. 0 ± 5.4	67.3 ± 9.4	11.8
	F	12	High school athletes, American	-	166.0 ± 7.8	57.7 ± 6.7	27.4
Kitagawa, Miyashita, and Yamamoto (1977)[9]	M	39	University students, Japanese	19.3 ± 0.8	172.1 ± 5.0	62.0 ± 6.7	13.1
	F	33	University students, Japanese	18.7 ± 0.3	157.0 ± 4.5	53.2 ± 5.1	21.6
Diaz, Hagan, Wright, and Horvath (1978)[10]	M	7	Activity level not stated, American	28.6	177	74. 4	12
	F	5	Activity level not stated, American	29	163	59.2	22.6
Daniels, Vogel, and Kowal (1978)[11]	M	30	First-year West Point cadets, American	-	-	70.6 ± 7.6	13.1
	F	30	First-year West Point cadets, American	-	-	57.7 ± 6.0	23.8
Vogel and Patton (1978)[12]	M	92	Untrained Army recruits, American	21.0 ± 4.0	-	72.0 ± 11.0	15.8
	F	92	Untrained Army recruits, American	20.0 ± 2.0	-	59.0 ± 7.0	27.9
Sparling (1979)[13]	M	34	Trained runners, American	26.6 ± 4.0	180.5 ± 6.8	70.3 ± 6.8	10.8
	F	34	Trained runners, American	25.0 ± 4.6	162.5 ± 6.9	52.9 ± 6.8	19.8

Note. Adapted from "A meta-analysis of studies comparing maximal oxygen uptake in men and women," by P. B. Sparling, 1980, *Research Quarterly for Exercise and Sport, 51*, p. 545.

[7] Mayhew, J. L. (1976). *Relative contributions of body composition, selected hematological parameters and aerobic capacity to endurance running performance of male and female adolescent track athletes* (Doctoral dissertation, University of Illinois).

[8] Dill, D. B., Soholt, L. F., McLean, D. C., Drost, T. F., & Loughran, M. T. (1977). Capacity of young males and females for running in desert heat. *Medicine and Science in Sports, 9*, 137-142.

[9] Kitagawa, K., Miyashita, M., & Yamamoto, K. (1977). Maximal oxygen uptake, body composition, and running performance in young Japanese adults of both sexes. *Japanese Journal of Physical Education, 21*, 335-340.

[10] Diaz, F. J., Hagan, R. D., Wright, J. E., & Horvath, S. M. (1978). Maximal and submaximal exercise in different positions. *Medicine and Science in Sports, 10*, 214-217.

[11] Daniels, W. L., Vogel, J. A., & Kowal, D. M. (1978). *Fitness levels and response to training of women in U.S. Army*. Toronto, Canada: Defence and Civil Institute of Environmental Medicine.

[12] Vogel, J. A., & Patton, J. F. (1978). Evaluation of fitness in U.S. Army. Toronto, Canada: Defence and Civil Institute of Environmental Medicine.

[13] Sparling, P. B. (1979). *Biological determinants of the sex difference in distance run performance among trained runners* (Doctoral dissertation, University of Georgia).

Table B2

Dataset 2 presented in Sparling (1980)

Study	Test	Units	Male	Female	% Difference (M-F)/F x 100	t	p	r	r^2
Von Dobeln (1956)[14]	Bicycle	l/min	3.90 ± 0.56	3.04 ± 0.54	28	6.5	0.001	0.62	0.39
	Ergometer	ml/min·kg BW	56.5 ± 6.9	48.7 ± 8.8	16	4.1	0.001	0.45	0.2
		ml/min·kg FFW	63.3 ± 6.6	60.6 ± 10.2	4	1.3	ns	-	-
Hermansen and Andersen (1965)[15]	Bicycle	l/min	4.8 ± 0.53	3.3 ± 0.43	45	5.2	0.001	0.78	0.61
	Ergometer	ml/min·kg BW	71.0 ± 6.8	55.0 ± 3.1	29	4.8	0.001	0.76	0.58
		ml/min·kg FFW	80.7 ± 7.5	67.8 ± 3.4	18	3.3	0.01	0.63	0.39
Cotes, Davies, Edholm, Healy, and Tanner (1969)[16]	Bycicle	l/min	3.43 ± 0.53	2.14 ± 0.38	60	8.9	0.001	0.81	0.66
	Ergometer	ml/min·kg BW	48.5 ± 7.9	39.2 ± 6.5	24	4.1	0.001	0.54	0.29
		ml/min·kg FFW	55.8 ± 8.7	53.5 ± 7.4	4	<1.0	ns	-	-
MacNab, Conger, and Taylor (1969)[17]	Treadmill	l/min	3.92 ± 0.58	2.32 ± 0.41	69	11	0.001	0.85	0.73
		ml/min·kg BW	51.7 ± 5.1	39.1 ± 5.1	32	8.6	0.001	0.78	0.62
		ml/min·kg FFW	59.4 ± 5.9	50.4 ± 6.0	18	7.1	0.001	0.72	0.52
	Bicycle	l/min	3.52 ± 0.61	2.12 ± 0.41	66	9.3	0.001	0.81	0.66
	Ergometer	ml/min·kg BW	46.5 ± 6.3	35.7 ± 5.6	30	6.3	0.001	0.68	0.46
		ml/min·kg FFW	53.3 ± 6.6	46.9 ± 7.2	14	5.1	0.001	0.6	0.36

[14] Von Dobeln, W. (1956). Human standard and maximal metabolic rate in relation to fatfree body mass. *Acta Physiologica Scandinavica Supplementum, 37*, 1-79.

[15] Hermansen, L., & Andersen, K. L. (1965). Aerobic work capacity in young Norwegian men and women. *Journal of Applied Physiology, 20*, 425-431.

[16] Cotes, J. E., Davies, C. T. M., Edholm, O. G., Healy, M. J. R., & Tanner, J. M. (1969). Factors relating to the aerobic capacity of 46 healthy British males and females, ages 18 to 28 years. *Proceedings of the Royal Society of London Britain, 174*, 91-114.

[17] MacNab, R. B. J., Conger, P. R., & Taylor, P. S. (1969). Differences in maximal and submaximal work capacity in men and women. *Journal of Applied Physiology, 27*, 644-648.

Dill, Myhre, Greer, Richardson, and Singleton (1972)[18]	Bicycle Ergometer	l/min	3.21 ± 0.49	1.92 ± 0.27	67	7.4	0.001	0.86	0.74
		ml/min·kg BW	45.2 ± 6.4	35.9 ± 3.3	26	4.1	0.001	0.81	0.65
		ml/min·kg FFW	52.9 ± 5.0	46.0 ± 4.7	15	3.3	0.01	0.6	0.36
Davies, Mbelwa, Crockford, and Weiner (1973)[19]	Bicycle Ergometer	l/min	2.76 ± 0.39	2.00 ± 2.4	38	10	0.001	0.72	0.52
		ml/min·kg BW	47.0 ± 5.2	40.2 ± 4.8	17	6.1	0.001	0.54	0.29
		ml/min·kg FFW	53.4 ±5.6	52.8 ± 6.0	1	<1.0	ns	-	-
Mayhew (1976)[20]	Treadmill	l/min	3.89 ± 0.45	2.70 ± 0.34	44	9.9	0.001	0.83	0.69
		ml/min·kg BW	63.8 ± 5.7	47.5 ± 4.1	34	10.8	0.001	0.85	0.73
		ml/min·kg FFW	69.8 ±6.1	59.1 ± 7.4	18	5.4	0.001	0.63	0.4
Dill, Soholt, McLean, Drost, and Loughran (1977)[21]	Treadmill	l/min	3.63 ± 0.60	2.13 ± 0.50	70	6.6	0.001	0.8	0.64
		ml/min·kg BW	54.0 ± 8.7	36.9 ± 4.1	46	6	0.001	0.77	0.6
		ml/min·kg FFW	61.0 ± 8.3	51.1 ± 6.9	19	3.1	0.001	0.53	0.29
Kitagawa, Miyashita, and Yamamoto (1977)[22]	Treadmill	l/min	3.22 ± 0.56	2.08 ± 0.21	55	10.9	0.001	0.79	0.63
		ml/min·kg BW	51.8 ± 6.6	39.2 ± 3.0	32	10	0.001	0.77	0.59
		ml/min·kg FFW	59.7 ± 6.9	50.0 ± 3.9	19	7.1	0.001	0.65	0.42

[18] Dill, D. B., Myhre, L. G., Greer, S. M., Richardson, J. C., & Singleton, K. J. (1972). Body composition and aerobic capacity of youth of both sexes. *Medicine and Science in Sports, 4,* 198-204.

[19] Davies, C. T. M., Mbelwa, D., Crockford, G., & Weiner, J. S. (1973). Exercise tolerance and body composition of male and female Africans aged 18 to 30 years. *Human Biology, 45,* 31-40.

[20] Mayhew, J. L. (1976). *Relative contributions of body composition, selected hematological parameters and aerobic capacity to endurance running performance of male and female adolescent track athletes* (Doctoral dissertation, University of Illinois).

[21] Dill, D. B., Soholt, L. F., McLean, D. C., Drost, T. F., & Loughran, M. T. (1977). Capacity of young males and females for running in desert heat. *Medicine and Science in Sports, 9,* 137-142.

[22] Kitagawa, K., Miyashita, M., & Yamamoto, K. (1977). Maximal oxygen uptake, body composition, and running performance in young Japanese adults of both sexes. *Japanese Journal of Physical Education, 21,* 335-340.

Diaz, Hagan, Wright, and Horvath (1978)[23]	Treadmill	l/min	3.78 ± 0.37	2.41 ± 4.7	57	5.2	0.001	0.85	0.73
		ml/min·kg BW	50.7 ± 4.2	40.5 ± 8.7	25	2.5	0.05	0.62	0.38
		ml/min·kg FFW	57.5 ± 4.2	52.3 ± 10.8	10	1.1	ns	-	-
	Bicycle	l/min	3.68 ± 0.42	2.21 ± 0.38	67	5.7	0.001	0.87	0.76
	Ergometer	ml/min·kg BW	49.8 ± 4.8	37.7 ± 7.8	32	3	0.05	0.69	0.47
		ml/min·kg FFW	56.4 ± 4.8	48.6 ± 9.2	16	1.7	ns	-	-
Daniels, Vogel, and Kowal (1978)[24]	Treadmill	l/min	4.19 ± 0.54	2.64 ± 0.31	58	13.4	0.001	0.87	0.75
		ml/min·kg BW	59.4 ± 5.9	46.0 ± 5.1	29	9.2	0.001	0.77	0.59
		ml/min·kg FFW	68.3 ± 5.7	60.0 ± 5.2	14	5.7	0.001	0.6	0.36
Vogel and Patton (1978)[25]	Treadmill	l/min	3.36 ± 0.48	2.25 ± 0.32	63	20.1	0.001	0.83	0.69
		ml/min·kg BW	50.8 ± 3.5	38.1 ± 3.5	33	20.1	0.001	0.83	0.69
		ml/min·kg FFW	60.4 ± 3.7	52.9 ± 3.5	14	11.5	0.001	0.65	0.42
Sparling (1979)[26]	Treadmill	l/min	4.29 ± 0.47	2.75 ± 0.40	56	14.6	0.001	0.87	0.76
		ml/min·kg BW	61.0 ± 4.9	51.9 ± 5.1	18	7.5	0.001	0.68	0.46
		ml/min·kg FFW	68.6 ± 5.5	65. 1 ± 5.6	5	2.6	0.05	0.31	0.09

Note. Adapted from "A meta-analysis of studies comparing maximal oxygen uptake in men and women," by P. B. Sparling, 1980, *Research Quarterly for Exercise and Sport, 51*, p. 546.

[23] Diaz, F. J., Hagan, R. D., Wright, J. E., & Horvath, S. M. (1978). Maximal and submaximal exercise in different positions. *Medicine and Science in Sports, 10*, 214-217.

[24] Daniels, W. L., Vogel, J. A., & Kowal, D. M. (1978). *Fitness levels and response to training of women in U.S. Army*. Toronto, Canada: Defence and Civil Institute of Environmental Medicine.

[25] Vogel, J. A., & Patton, J. F. (1978). Evaluation of fitness in U.S. Army. Toronto, Canada: Defence and Civil Institute of Environmental Medicine.

[26] Sparling, P. B. (1979). *Biological determinants of the sex difference in distance run performance among trained runners* (Doctoral dissertation, University of Georgia).

Table B3

Dataset presented in Desilva et al. (1981)

Study	Exclusions	% older than 70 years (untreated, lignocaine)	Mean age (untreated, lignocaine)	Blinded	Onset-admission interval	Bolus	Infusion	Duration	Cross-overs	Toxic effects	Lignocaine level (mg/ml)	VF incidence in untreated group	VF incidence in lignocaine group	Statistical significance
Bleifeld, Merx, Heinrich, and Effert (1973)[27]	AV block Shock VT,VF Severe LV failure	Not stated	60.1 (61.0; 59.0)	Not stated	< 5h -34%; <24h - 63%; <48h - 82%; (unknown - 18%)	100 mg i.v.	14-42mg/kg/min	120h	No	7% slight CNS; more 2nd degree AVB on days 2 and 3 in lignocaine group	Not determined	2/48 (4.2%)	0/41 (0%)	NS
Bennett, Wilner, and Pentecost (1970)[28]	AV block VT,VF Shock HR<50 Severe LV failure	8.2 (8.0; 8.4)	57.0 (56.8; 57.2)	No	≤ 3h - 36%; ≤12h - 72%; ≤24h - 100%	60 mg i.v.	1mg/min	48h	Yes-31%	None	Not determined	7/125 (5.6%)	5/131 (3.8%)	NS
Mogensen (1970)[29]	Shock AV block	27 (24; 29)	63.7 (63; 64.3)	Not stated	≤6h - 58%; ≤12h - 72%; >12h - 28%	75 mg i.v.	2mg/min	24h	Yes-65%	39% slight CNS	1,0-5,6 at 1h	1/37 (2.7%)	0/42 (0%)	NS

[27] Bleifeld, W., Merx, W., Heinrich, K. W., & Effert, S. (1973). Controlled trial of prophylactic treatment with lidocaine in myocardial infarction. *European Journal of Clinical Pharmacology, 6,* 119-126.

[28] Bennett, M. A., Wilner, J. M., & Pentecost, B. L. (1970). Controlled trial of lignocaine in prophylaxis of ventricular arrhythmias complicating myocardial infarction. *Lancet, 2,* 909-911.

[29] Mogensen, L. (1970). Ventricular tachyarrhytmias and lignocaine prophylaxis in acute myocardial infarction. *Acta Medica Scandinavica Supplementum, 513,* 1-80.

O'Brien, Taylor, and Croxson (1973)[30]	VT,VF Other cardiac arrest	Not stated	Not stated	Double	Not stated	75 mg i.v.	2.5mg/min	48h	No	CNS: 36% asystole (duration unstated) in 7 lignocaine patients and 2 untreated	4.0 at 24h; 5.5 at 48h	5/146 (3.4%)	7/154 (4.5%)	Not stated
Lie, Wellens, Van Capelle, and Durrer (1974)[31]	AV block Shock VT,VF HR<50 Age>70 LV failure	0	58.5 (59.0; 58.1)	Double	<2h - 47%; <6h - 100%	100 mg i.v.	3mg/min	48h	No	15% slight CNS	1.5-6.4	11/105 (10.5%) Transient VF-2 patients	0/107 (0%)	p< .03
Church and Biern (1972)[32]	LV failure Arrhythmias AV block Shock	Not stated	Not stated	Single	<4h 64% lignocaine 68% untreated	50,75 mg i.v.	2mg/min	48h	Yes-23%	21% brady-cardia or hypo-tension	Not stated	3/44 (6.8%)	4/42 (9.5%)	Not stated

Note. Adapted from "Lignocaine prophylaxis in acute myocardial infarction: An evaluation of randomised trials," by R. A. Desilva,C. H. Hennekens, B. Lown, and W. Casscells, 1981, *Lancet, 2*, p. 857.

[30] O'Brien, K. P., Taylor, P. M., Croxson, R. S. (1973). Prophylactic lignocaine in hospitalized patients with acute myocardial infarction. *Medical Journal of Australia Supplementum, 2,* 36-37.

[31] Lie, K. I., Wellens, H. J., Van Capelle, F. J., & Durrer, D. (1974). Lidocaine in the prevention of primary ventricular fibrillation. *New England Journal of Medicine, 291,* 1324-1326.

[32] Church, G., & Biern, R. (1972). Prophylactic lidocaine in acute myocardial infarction. *Circulation, 45-46,* 11-139.

Table B4

Dataset presented in Stampfer et al. (1982)

Trial	**Duration of symptoms (hours)**	**Loading dose (thousands of IU)**	**Infusion dose (thousands of IU)**	**Period (hours)**	**Placebo controls?**	**Follow-up period (days)**	**Mortality - drug group (no. dead/total)**	**Mortality - controls (no. dead/total)**	**Risk Ratio (95% Confidence Limits)**	**Two-tailed p value**
1st European trial, 1969[33]	72	1250	104	72	No	HS (Hospital stay)	20/83 (24.1%)	15/84 (17.9%)	1.35 (0.74-2.45)	.32
2nd European trial, 1971[34]	24	250	100	24	No	HS	69/373 (18.5%)	94/357 (26.3%)	0.70 (0.53-0.92)	.01
Finnish study, 1971[35]	72	600	Varied	Varied	No	42	22/219 (10.0%)	17/207 (8.2%)	1.22 (0.67-2.24)	.51
Italian study (CCU), 1971[36]	12	250	150	12	No	40	19/164 (11.6%)	18/157 (11.5%)	1.01 (0.55-1.85)	.97
2nd Frankfurt study, 1972[37]	12	250	200	2,5	Yes	HS	13/102 (12.7%)	29/104 (27.9%)	0.46 (0.26-0.81)	.007

[33] Amery, A., Roeber, G.,Vermeulen, H. J., & Verstraete, M. (1969). Single-blind randomised multicenter trial comparing heparin and streptokinase treatment in recent myocardial infarction. *Acta Medica Scandinavica Supplementum, 505,* 5-35.

[34] European working party. (1971). Streptokinase in recent myocardial infarction: A controlled multicenter trial. *British Medical Journal, 3,* 325-331.

[35] Heikinheimo, R., Ahrenberg, P., Honkapohja, H., Iisalo, E., Kallio, V., Konttinen, Y., … Siitonen, L. (1971). Fibrinolytic treatment in acute myocardial infarction. *Acta Medica Scandinavica, 189,* 7-13.

[36] Dioguardi, N., Lotto, A., Levi, G. F., Rota, M., Proto, C., Mannucci, P. M., … Agostoni, A. (1971). Controlled trial of streptokinase and heparin in acute myocardial infarction. *Lancet, 2,* 891-895.

[37] Breddin, K., Ehrly, A. M. Fechler, L., Frick, D., König, H., Kraft, H., … Wylicil, P. (1973). Die Kurzzeitfibrinolyse beim akuten Myokardinfarkt. *Deutsche Medizinische Wochenschrift, 98,* 861-873.

Australian trial (CCU), 1973[38]	24	250	100	17	No	40	21/264 (8.0%)	23/253 (9.1%)	0.88 (0.50-1.54)	.64
Australian trial (CCU), 1973	24	250	100	17	No	90	26/264 (9.8%)	32/253 (12.6%)	0.78 (0.48-1.27)	.31
British study (CCU), 1976[39]	24	250	100	24	Yes	42	43/302 (14.2%)	44/293 (15.0%)	0.95 (0.64-1.40)	.79
British study (CCU), 1976	24	250	100	24	Yes	6mo	48/302 (15.9%)	52/293 (17.7%)	0.90 (0.63-1.28)	.55
European Study Group (CCU), 1979[40]	12	250	100	24	Yes	21	18/156 (11.5%)	30/159 (18.9%)	0.61 (0.36-1.04)	.07
European Study Group (CCU), 1979	12	250	100	24	Yes	6mo	25/156 (16.0%)	50/159 (31.4%)	0.51 (0.34-0.77)	.001

Note. Adapted from "Effect of intravenous streptokinase on acute myocardial infarction: Pooled results from randomized trials," by M. J. Stampfer, S. Z. Goldhaber, S. Yusuf, R. Peto, and C. H. Hennekens, 1982, *New England Journal of Medicine, 307,* p. 1181.

[38] Bett, J. H. N., Castaldi, P. A., Hale, G. S., Isbister, J. P., Mclean, K. H., O'Sullivan, E. F., … Rosenbaum, M. (1973). Australian multicenter trial of streptokinase in acute myocardial infarction. *Lancet, 1,* 57-60.

[39] Aber, C. P., Bass, N. M., Berry, C. L., Carson, P. H. M., Dobbs, R. J., Fox, K. M., … Stock, J. P. P. (1976). Streptokinase in acute myocardial infarction: A controlled multicenter study in the United Kingdom. *British Medical Journal, 2,* 1100-1104.

[40] European cooperative study group for streptokinase treatment in acute myocardial infarction. (1979). Streptokinase in acute myocardial infarction. *New England Journal of Medicine, 301,* 797-802.

Appendix C: Reanalyses contrasted to original meta-analyses

Table C1

Results of Sparling (1980) contrasted to results of reanalysis

		Sparling (1980)	**Reanalysis**
Synthesis of results Sex differences in VO_2 max expressed in absolute terms (liters/minute)	summary effect	r_{pb} = .81	*FEM* r_{pb} = .809 (*95% CI* = 0.784-0.832, p < .001)
			REM r_{pb} = .810 (*95% CI* = 0.771-0.843, p < .001)
	heterogeneity	-	Q = 22,773, df = 12, p = .03
			I^2 = 47.305
Synthesis of results Sex differences in VO_2 max expressed relative to body weight (ml/minute*kg BW)	summary effect	r_{pb} = .70	*FEM* r_{pb} = .729 (*95% CI* = 0.695-0.761, p < .001)
			REM r_{pb} = .720 (*95% CI* = 0.641-0.784, p < .001)
	heterogeneity	-	Q = 46.541, df = 12, p < .001
			I^2 = 74.216
Synthesis of results Sex differences in VO_2 max expressed relative to fat-free weight (ml/minute*kg FFW)	summary effect	r_{pb} = .59	*FEM* r_{pb} = .607 (*95% CI* = 0.552-0.657, p < .001)
			REM r_{pb} = .599 (*95% CI* = 0.524-0.664, p < .001)
	heterogeneity	-	Q = 12.053, df = 8, p = .149
			I^2 = 33.624
Additional analysis Subgroup analysis	Trained men vs. trained women	Percentage difference: 46%	*FEM*: r_{pb} = .810 (*95% CI* = 0.762-0.849, p < .001)

(liters/minute)			*MEM*: r_{pb} = .811 (*95% CI* = 0.701-0.883, $p < .001$)
	Untrained men vs. untrained women	Percentage difference: 60%	*FEM*: r_{pb} = .809 (*95% CI* = 0.777-0.836, $p < .001$)
			MEM: r_{pb} = .809 (*95% CI* = 0.777-0.836, $p < .001$)
	Comparison/ heterogeneity	46% versus 60%	*FEM*: (*Q* (total between) = 0.001, *df* = 1, $p = .975$)
			MEM: (*Q* (total between) = 0.001, *df* = 1, $p = .974$)
Additional analysis <u>Subgroup analysis</u> (ml/minute*kg BW)	Trained men vs. trained women	Percentage difference: 25%	*FEM*: r_{pb} = .699 (*95% CI* = 0.629-0.757, $p < .001$)
			MEM: r_{pb} = .718 (*95% CI* = 0.559-0.826, $p < .001$)
	Untrained men vs. untrained women	Percentage difference: 30%	*FEM*: r_{pb} = .743 (*95% CI* = 0.702-0.778, $p < .001$)
			MEM: r_{pb} = .722 (*95% CI* = 0.621-0.799, $p < .001$)
	Comparison/ heterogeneity	25% versus 30%	*FEM*: (*Q* (total between) = 1.408, *df* = 1, $p = .235$)
			MEM: (*Q* (total between) = 0.002, *df* = 1, $p = .964$)
Additional analysis <u>Subgroup analysis</u> (ml/minute*kg FFW)	Trained men vs. trained women	Percentage difference: 12%	*FEM*: r_{pb} = .518 (*95% CI* = 0.404-0.617, $p < .001$)
			MEM: r_{pb} = .536 (*95% CI* = 0.356-0.678, $p < .001$)
	Untrained men vs. untrained women	Percentage difference: 13%	*FEM*: r_{pb} = .645 (*95% CI* = 0.582-0.700, $p < .001$)

			MEM: r_{pb} = .645 (*95% CI* = 0.582-0.700, $p < .001$)
	Comparison/ heterogeneity	12% versus 13%	*FEM*: (*Q* (total between) = 4.507, *df* = 1, *p* = .034)
			MEM: (*Q* (total between) = 1.752, *df* = 1, *p* = .186)
Analyses concerning publication bias (liters/minute)		-	Visual inspection of the funnel plots suggested asymmetry.
			Begg and Mazumdar's rank correlation method (*p* (2-tailed) = 1.000)
			Egger's regression test (*p* (2-tailed) = .827)
			Duval-Tweedie trim-and-fill method (based on *FEM*): Three missing studies, adjusted point estimate r_{pb} = .791 (*95% CI* = 0.765-0.814)
			Duval-Tweedie trim-and-fill method (based on *REM*): Three missing studies, adjusted point estimate r_{pb} = .788 (*95% CI* = 0.742-0.826)
Analyses concerning publication bias (ml/minute*kg BW)		-	Visual inspection of the funnel plots suggested no asymmetry.
			Begg and Mazumdar's rank correlation method (*p* (2-tailed) = .903)
			Egger's regression test (*p* (2-tailed) = .598)
			Duval-Tweedie trim-and-fill method (based on *FEM*): No missing studies

			Duval-Tweedie trim-and-fill method (based on *REM*): No missing studies
Analyses concerning publication bias (ml/minute*kg FFW)		-	Visual inspection of the funnel plots suggested asymmetry.
			Begg and Mazumdar's rank correlation method (*p* (2-tailed) = .175)
			Egger's regression test (*p* (2-tailed) = .509)
			Duval-Tweedie trim-and-fill method (based on *FEM*): Three missing studies, adjusted point estimate r_{pb} = .520 (*95% CI* = 0.471-0.566)
			Duval-Tweedie trim-and-fill method (based on *REM*): Three missing studies, adjusted point estimate r_{pb} = .534 (*95% CI* = 0.441-0.615)
Retrospective power analysis (liters/minute)		-	*FEM*: r_{pb} = .801 (*95% CI* = 0.768-0.829) The mean power of the included studies to detect this new effect size was .923 with a standard deviation of 0.047 and a minimum of .862 and a maximum of .999.
			REM: r_{pb} = .798 (*95% CI* = 0.722-0.854) The mean power of the included studies to detect this new effect size was .922 with a standard deviation of 0.048 and a minimum of .859 and a maximum of .999.
Retrospective power analysis (ml/minute*kg BW)		-	*FEM*: r_{pb} = .718 (*95% CI* = 0.671-0.759)

			The mean power of the included studies to detect this new effect size was .884 with a standard deviation of 0.070 and a minimum of .831 and a maximum of .998.
			REM: r_{pb} = .681 (*95% CI* = 0.502-0.804) The mean power of the included studies to detect this new effect size was .893 with a standard deviation of 0.094 and a minimum of .810 and a maximum of .996.
Retrospective power analysis (ml/minute*kg FFW)		-	*FEM*: r_{pb} = .594 (*95% CI* = 0.524-0.655) The mean power of the included studies to detect this new effect size was .743 with a standard deviation of 0.161 and a minimum of .619 and a maximum of .980.
			REM: r_{pb} = .560 (*95% CI* = 0.344-0.720) The mean power of the included studies to detect this new effect size was .745 with a standard deviation of 0.191 and a minimum of .624 and a maximum of .965.

Note. FEM = fixed-effect analysis; REM = random-effects analysis; MEM = mixed-effects analysis.

Table C2

Results of Desilva et al. (1981) contrasted to results of reanalysis

		Desilva et al. (1981)	**Reanalysis**
Synthesis of results All six trials included for meta-analysis	summary effect	*RR* = 0.53 (*95% CI* = 0.28-0.98)	*FEM* *RR* = 0.780 (*95% CI* = 0.409-1.486, *p* = .449)
			REM *RR* = 0.704 (*95% CI* = 0.316-1.567, *p* = .390)
	heterogeneity	-	*Q* = 6.608, *df* =5, *p* = .251
			I^2 = 24.340
Additional analysis <u>Sensitivity analysis</u> Four trials included for meta-analysis (excluded: Mogensen, 1970; O'Brien, Taylor, & Croxson, 1973)	summary effect	*RR* = 0.22 (*95% CI* = 0.09-0.55)	*FEM* *RR* = 0.630 (*95% CI* = 0.280-1.419, *p* = .265)
			REM *RR* = 0.511 (*95% CI* = 0.154-1.694, *p* = .272)
	heterogeneity	-	*Q* = 5.123, *df* = 3, *p* = .163
			I^2 = 41.441
Analyses concerning publication bias			Visual inspection of the funnel plots suggested symmetry
			Begg and Mazumdar's rank correlation method (*p* (2-tailed) = .452)
			Egger's regression test (*p* (2-tailed) = .080)
			Duval-Tweedie trim-and-fill method (based on *FEM*): No missing studies
			Duval-Tweedie trim-and-fill method (based on *REM*): No missing studies

Retrospective power analysis		-	All studies in the meta-analysis were so weak (maximum power for the fixed-effect and random-effects analysis was 0.264 and 0.471 respectively) that no further analysis could be undertaken.

Note. FEM = fixed-effect analysis; REM = random-effects analysis; RR = risk ratio.

Table C3

Results of Stampfer et al. (1982) contrasted to results of reanalysis

		Stampfer et al. (1982)	**Reanalysis**
Synthesis of results All eight trials included for meta-analysis	summary effect	*RR* = 0.80 (*95% CI* = 0.68-0.95, *p* = .01)	*FEM* *RR* = 0.805 (*95% CI* = 0.684-0.947, *p* = .009)
			REM *RR* = 0.828 (*95% CI* = 0.661-1.037, *p* = .100)
	heterogeneity	Chi-square test of heterogeneity *p* = .20	*Q* = 11.775, *df* = 7, *p* = .108
			I^2 = 40.55
Additional analysis <u>Sensitivity analysis</u> Six trials included for meta-analysis (excluded: Amery, Roeber, Vermeulen, & Verstraete, 1969; Heikinheimo et al., 1971)	summary effect	*RR* = 0.74 (*95% CI* = 0.62-0.89, *p* = .001)	*FEM* *RR* = 0.743 (*95% CI* = 0.623-0.886, *p* = .001)
			REM *RR* = 0.745 (*95% CI* = 0.605-0.916, *p* = .005)
	heterogeneity	-	*Q* = 6.293, *df* = 5, *p* = .279
			I^2 = 20.547
Additional analysis <u>Sensitivity analysis</u> Four CCU-trials included for meta-analysis (only risk-ratios reported for early weeks were pooled)	summary effect	*RR* = 0.85 (*95% CI* = 0.66-1.10, *p* = .23)	*FEM* *RR*= 0.857 (*95% CI* = 0.667-1.100, *p* = .226)
			REM *RR* = 0.857 (*95% CI* = 0.667-1.100, *p* = .226)
	heterogeneity	-	*Q* = 2.135, *df* = 3, *p* = .545
			I^2 = 0.00
			(T^2 = 0.00)
Additional analysis	summary	*RR* = 0.71 (*95% CI* =	*FEM*

Sensitivity analysis Three of the four CCU-trials included for meta-analysis (only risk ratios reported for longer follow-up periods were pooled)	effect	0.56-0.91, *p* = .008)	*RR* = 0.722 (*95% CI* = 0.571-0.913, *p* = .006)
			REM *RR* = 0.714 (*95% CI* = 0.502-1.015, *p* = .061)
	heterogeneity		*Q* = 4.360, *df* = 2, *p* = .113
			I^2 = 54.133
Analyses concerning publication bias			Visual inspection of the funnel plots suggested asymmetry.
			Begg and Mazumdar's rank correlation method (*p* (2-tailed) = .386)
			Egger's regression test (*p* (2-tailed) = .423)
			Duval-Tweedie trim-and-fill method (based on *FEM*): Two missing studies, adjusted point estimate *RR* = 0.744 (*95% CI* = 0.639-0.866)
			Duval-Tweedie trim-and-fill method (based on *REM*): Two missing studies, adjusted point estimate *RR* = 0.746 (*95% CI* = 0.591-0.941)
Retrospective power analysis		-	All studies in the meta-analysis were so weak (maximum power for the fixed-effect and random-effects analysis was .453 and .376. respectively) that no further analysis could be undertaken.

Note. FEM = fixed-effect model; REM = random-effects model; RR = risk ratio.

Appendix D: Forest plots

Forest plots for the fixed-effect and random-effects analyses of the respective reanalyses

Figure D1

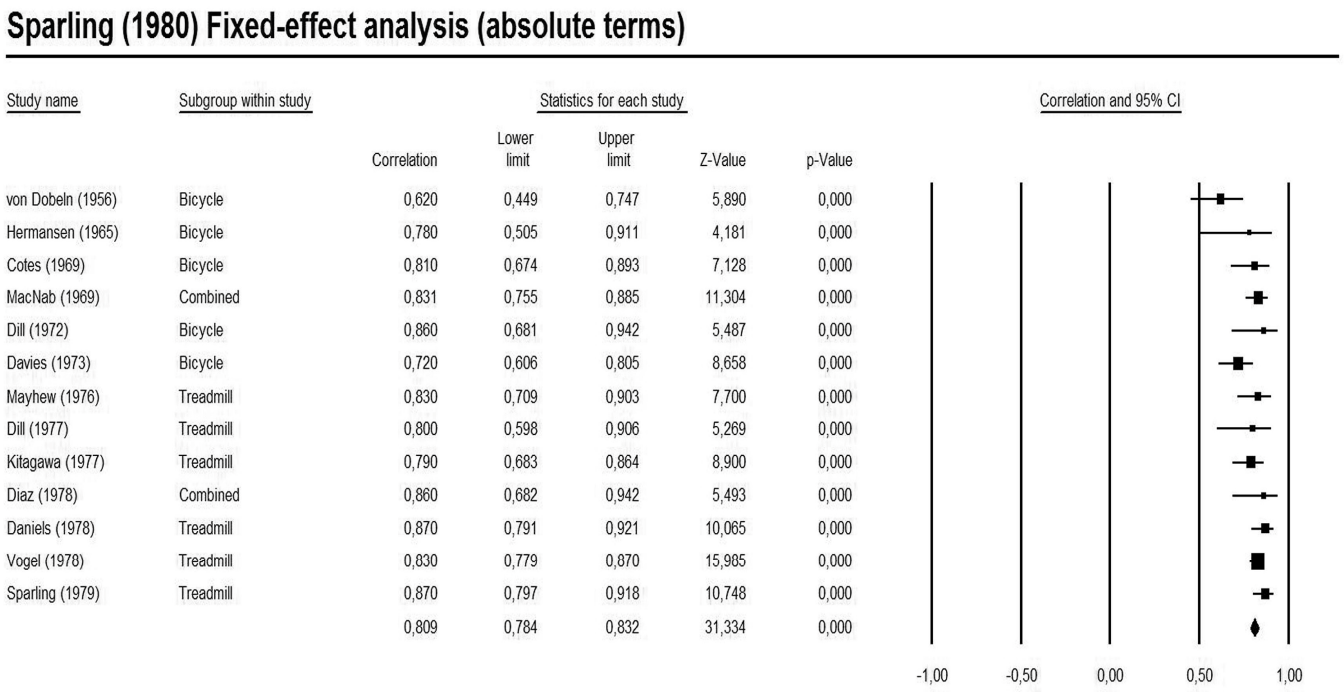


Figure D2

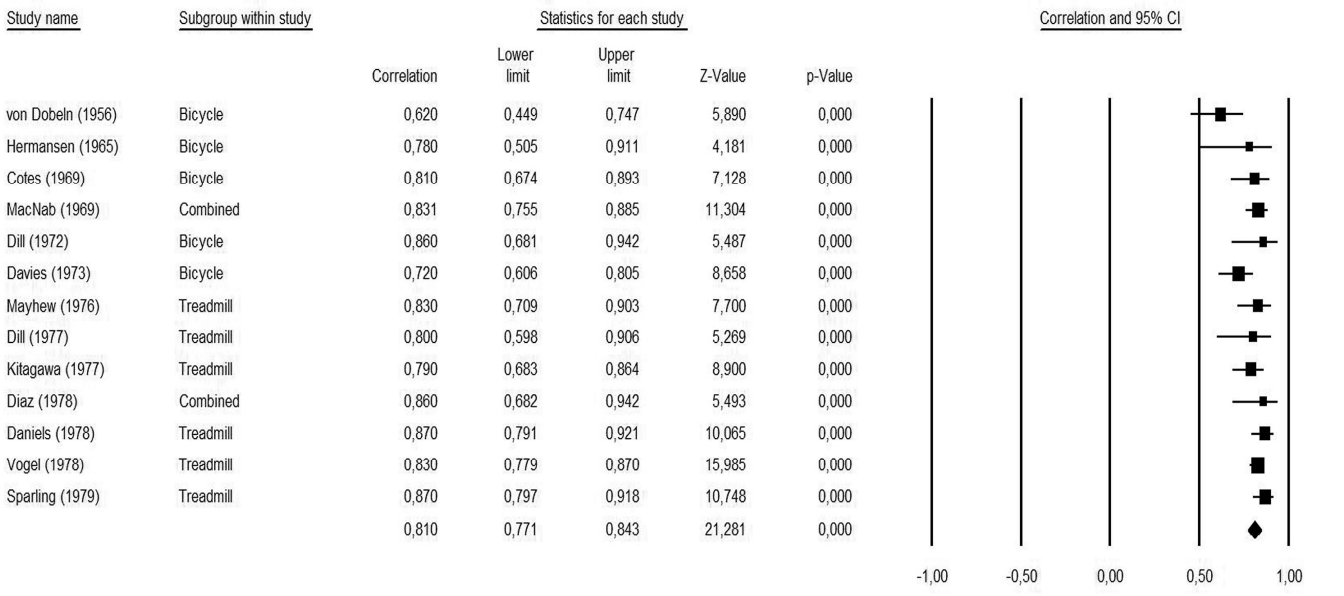


Figure D3

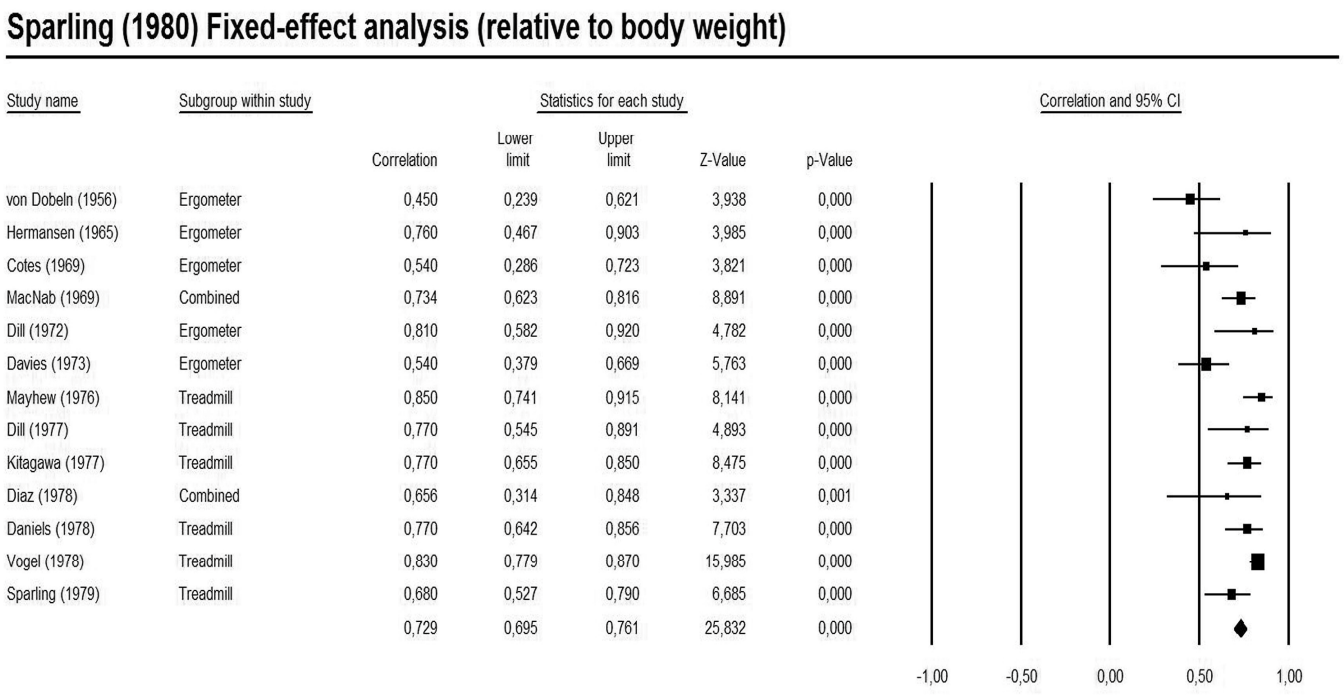


Figure D4

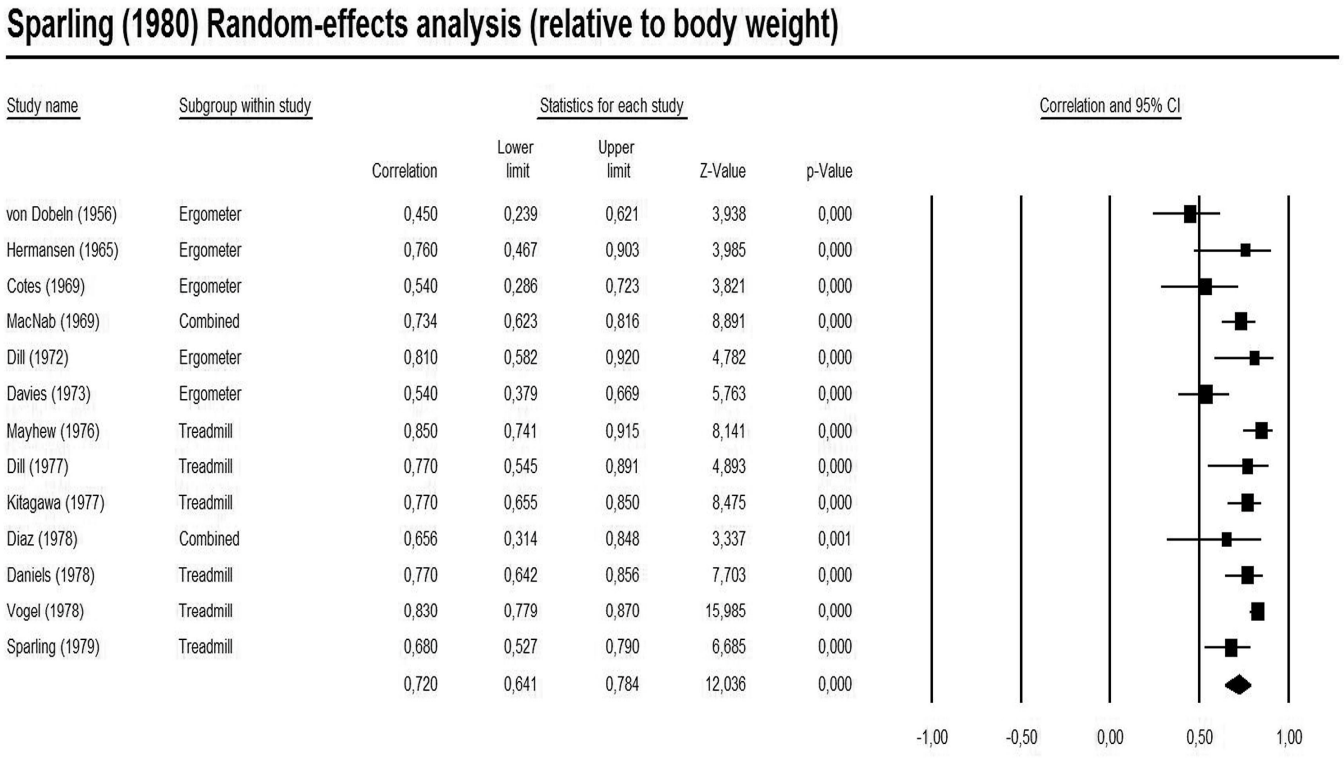


Figure D5

Sparling (1980) Fixed-effect analysis (relative to fat-free weight)

Study name	Subgroup within study	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Hermansen (1965)	Ergometer	0,630	0,246	0,843	2,966	0,003
MacNab (1969)	Combined	0,664	0,533	0,765	7,593	0,000
Dill (1972)	Ergometer	0,600	0,227	0,819	2,941	0,003
Mayhew (1976)	Treadmill	0,630	0,413	0,779	4,805	0,000
Dill (1977)	Treadmill	0,530	0,179	0,761	2,830	0,005
Kitagawa (1977)	Treadmill	0,650	0,492	0,766	6,440	0,000
Daniels (1978)	Treadmill	0,600	0,408	0,741	5,233	0,000
Vogel (1978)	Treadmill	0,650	0,558	0,726	10,431	0,000
Sparling (1979)	Treadmill	0,310	0,077	0,511	2,584	0,010
		0,607	0,552	0,657	16,688	0,000

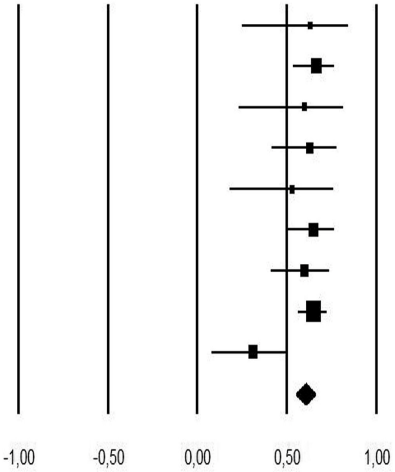


Figure D6

Sparling (1980) Random-effects analysis (relative to fat-free weight)

Study name	Subgroup within study	Correlation	Lower limit	Upper limit	Z-Value	p-Value
Hermansen (1965)	Ergometer	0,630	0,246	0,843	2,966	0,003
MacNab (1969)	Combined	0,664	0,533	0,765	7,593	0,000
Dill (1972)	Ergometer	0,600	0,227	0,819	2,941	0,003
Mayhew (1976)	Treadmill	0,630	0,413	0,779	4,805	0,000
Dill (1977)	Treadmill	0,530	0,179	0,761	2,830	0,005
Kitagawa (1977)	Treadmill	0,650	0,492	0,766	6,440	0,000
Daniels (1978)	Treadmill	0,600	0,408	0,741	5,233	0,000
Vogel (1978)	Treadmill	0,650	0,558	0,726	10,431	0,000
Sparling (1979)	Treadmill	0,310	0,077	0,511	2,584	0,010
		0,599	0,524	0,664	12,403	0,000

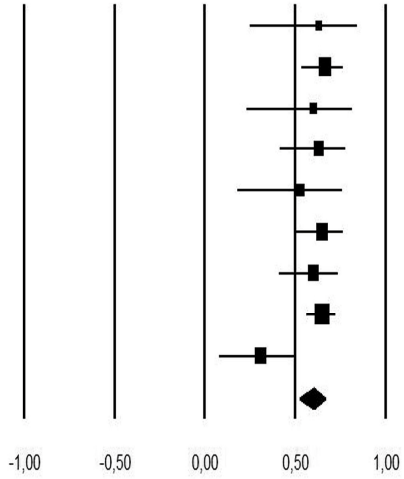


Figure D7

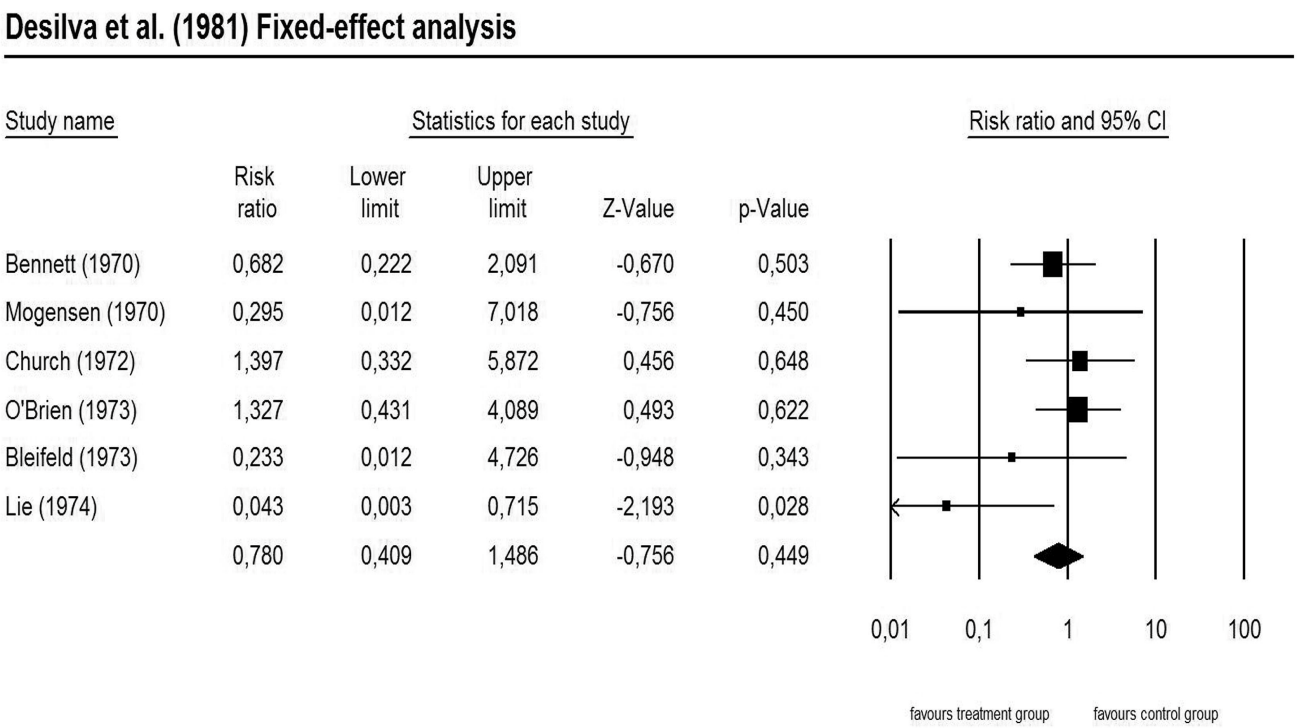


Figure D8

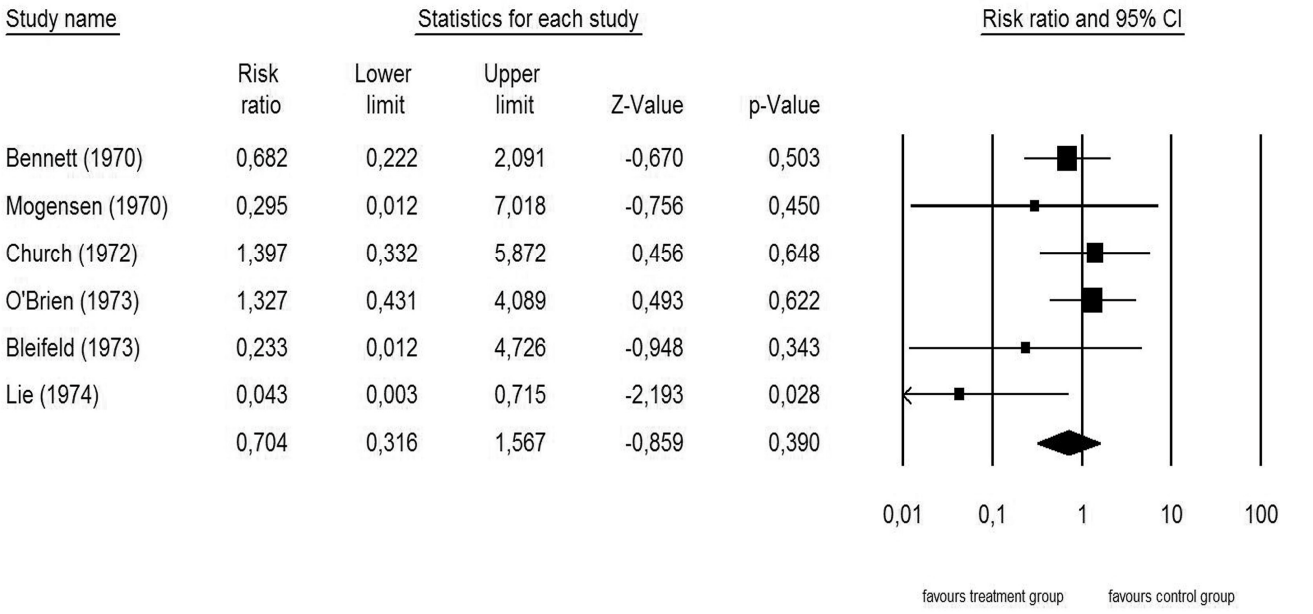


Figure D9

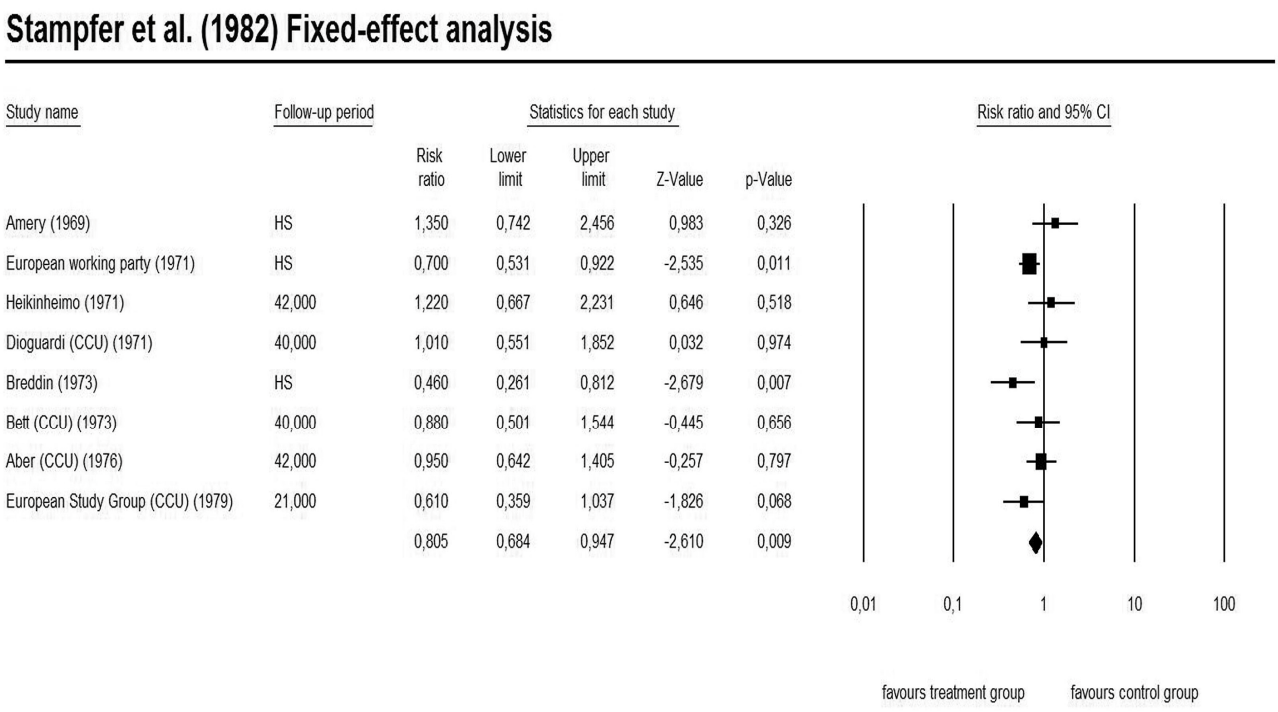
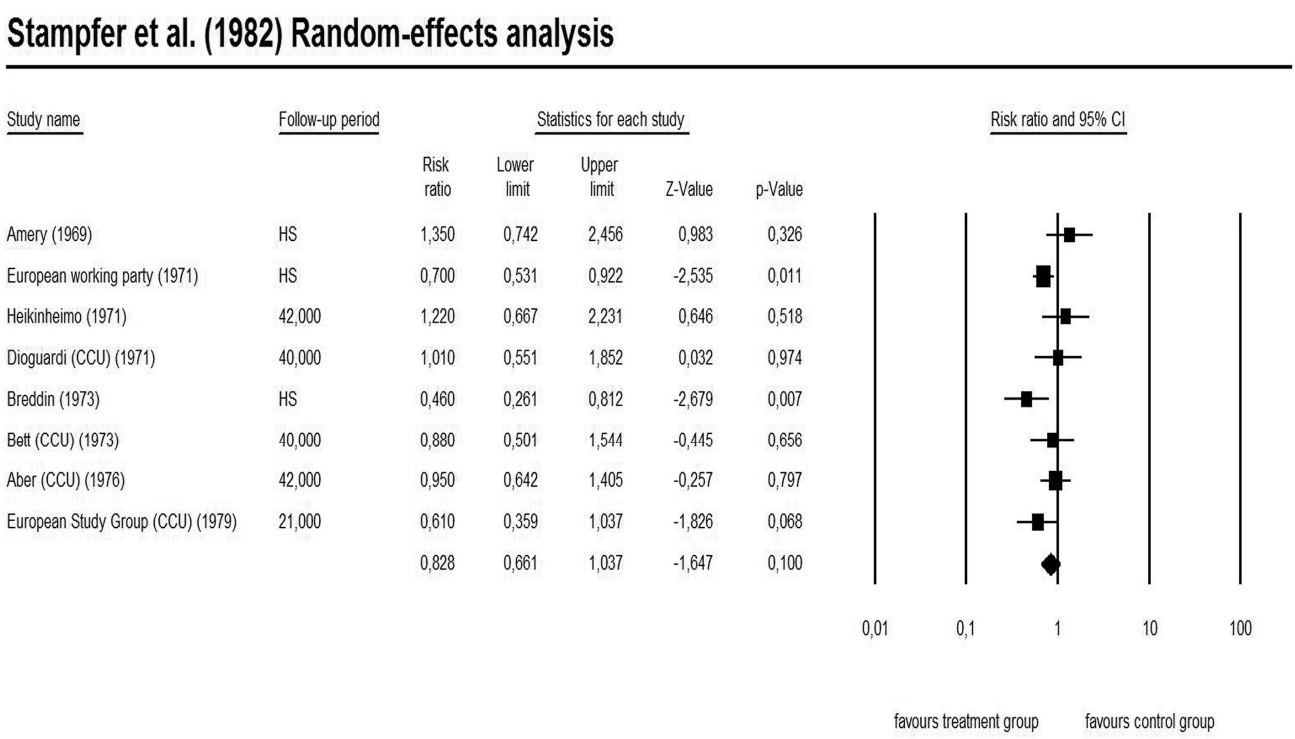


Figure D10



Appendix E: Funnel plots

Funnel plots for all reanalyses including observed studies as well as studies imputed by the Duval-Tweedie trim-and-fill method, which was once based on a fixed-effect model and the other time based on a random-effects model

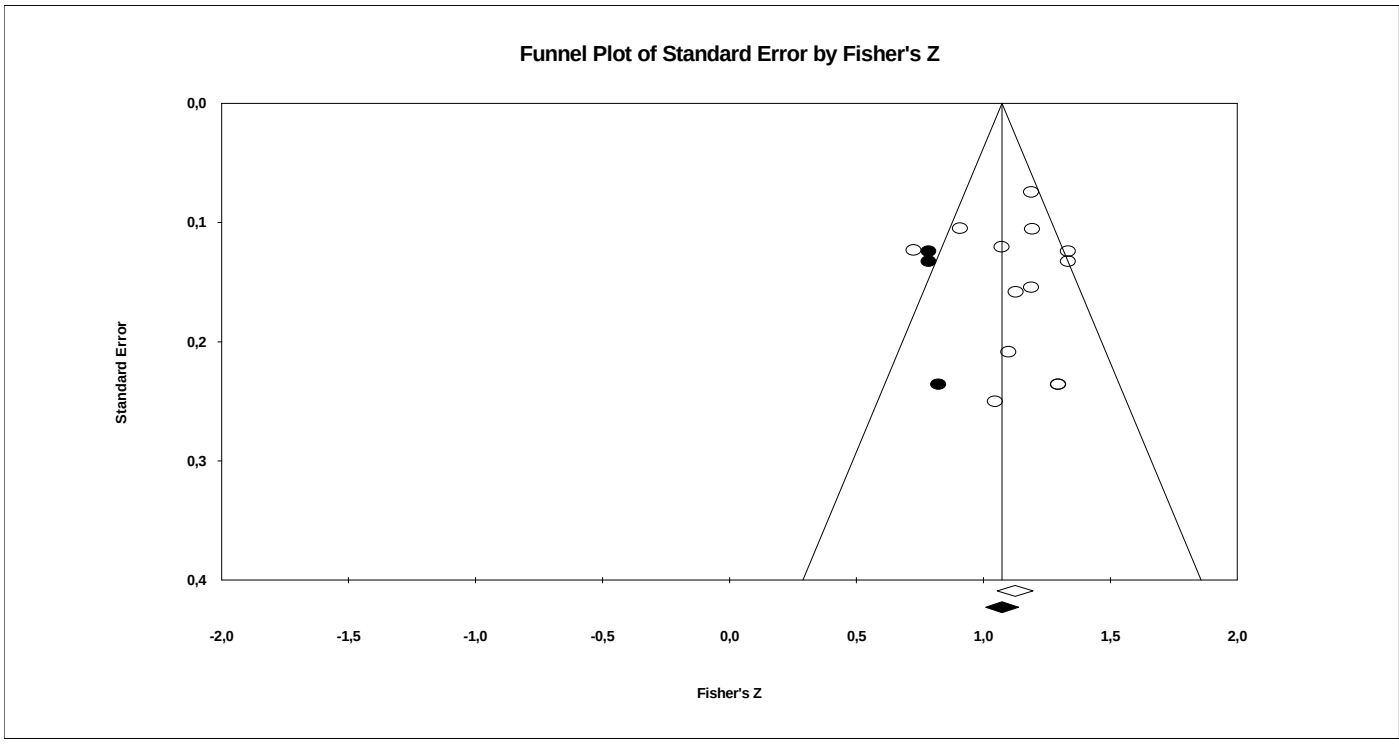


Figure E1. Sparling (1980); absolute terms; fixed-effect model.

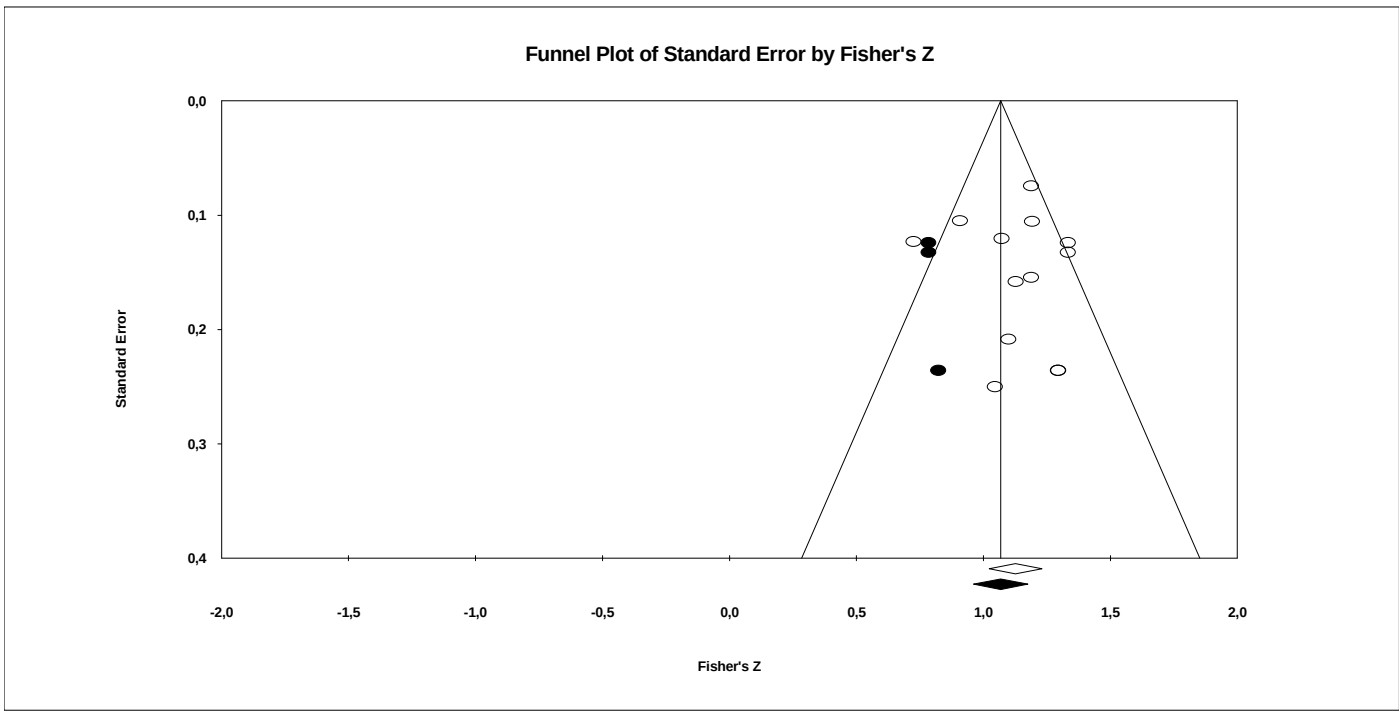


Figure E2. Sparling (1980); absolute terms; random-effects model.

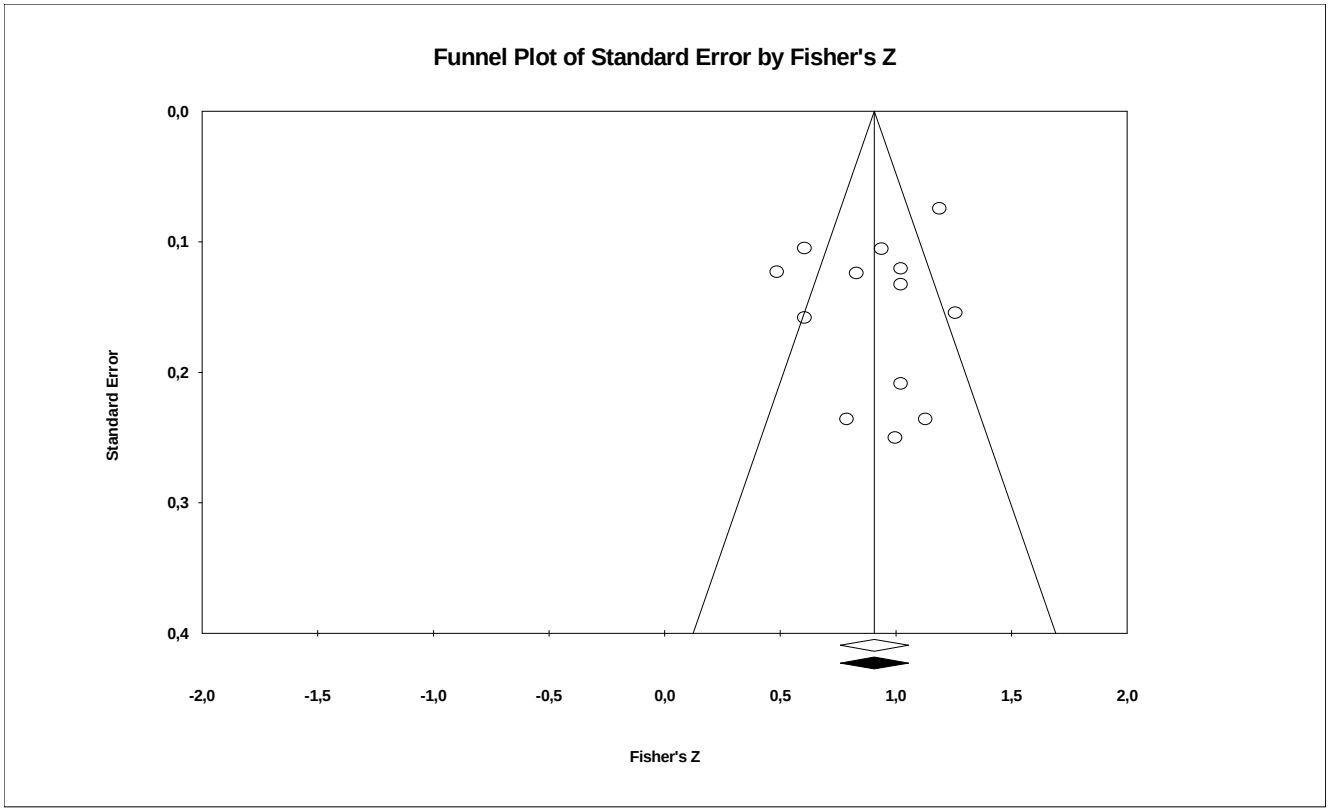


Figure E3. Sparling (1980); relative to body weight; fixed-effect analysis.

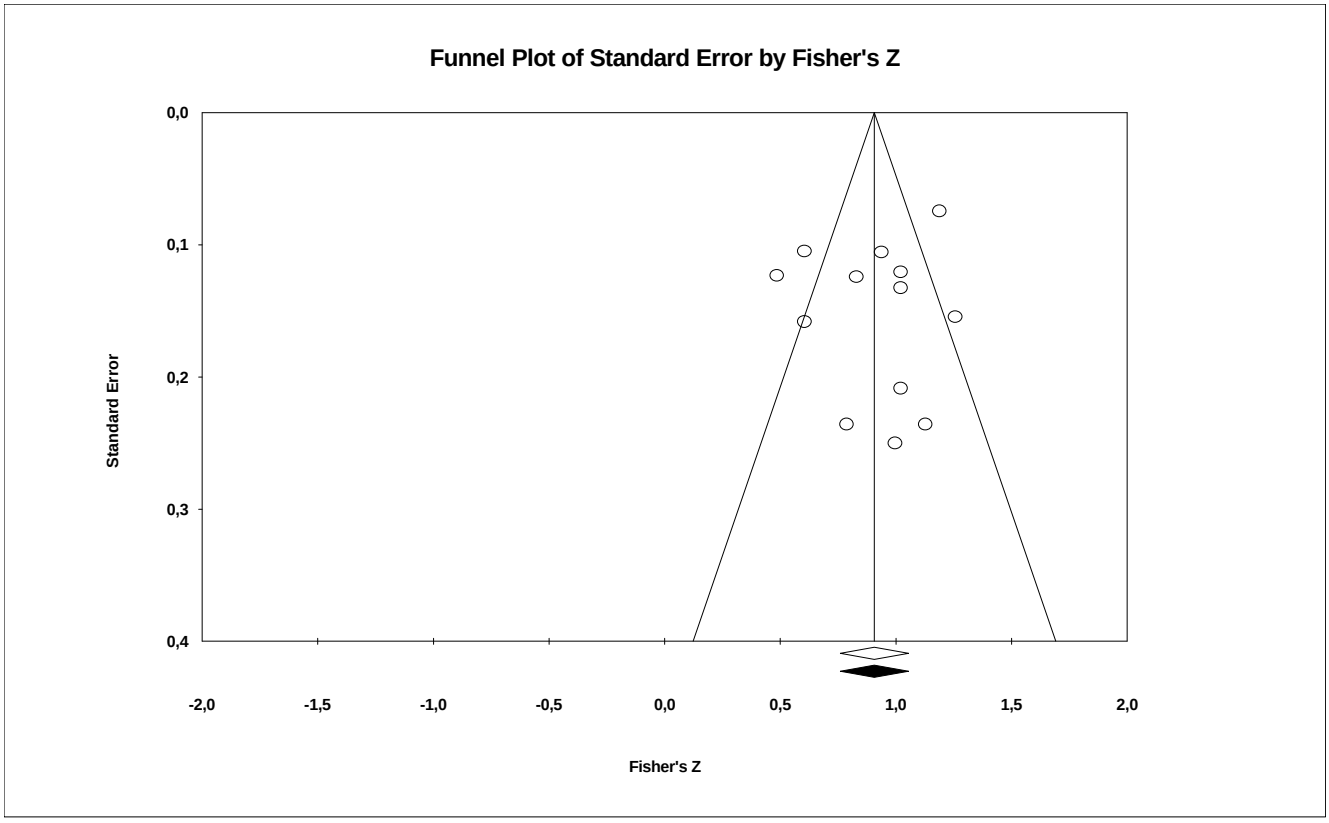


Figure E4. Sparling (1980); relative to body weight; random-effects analysis.

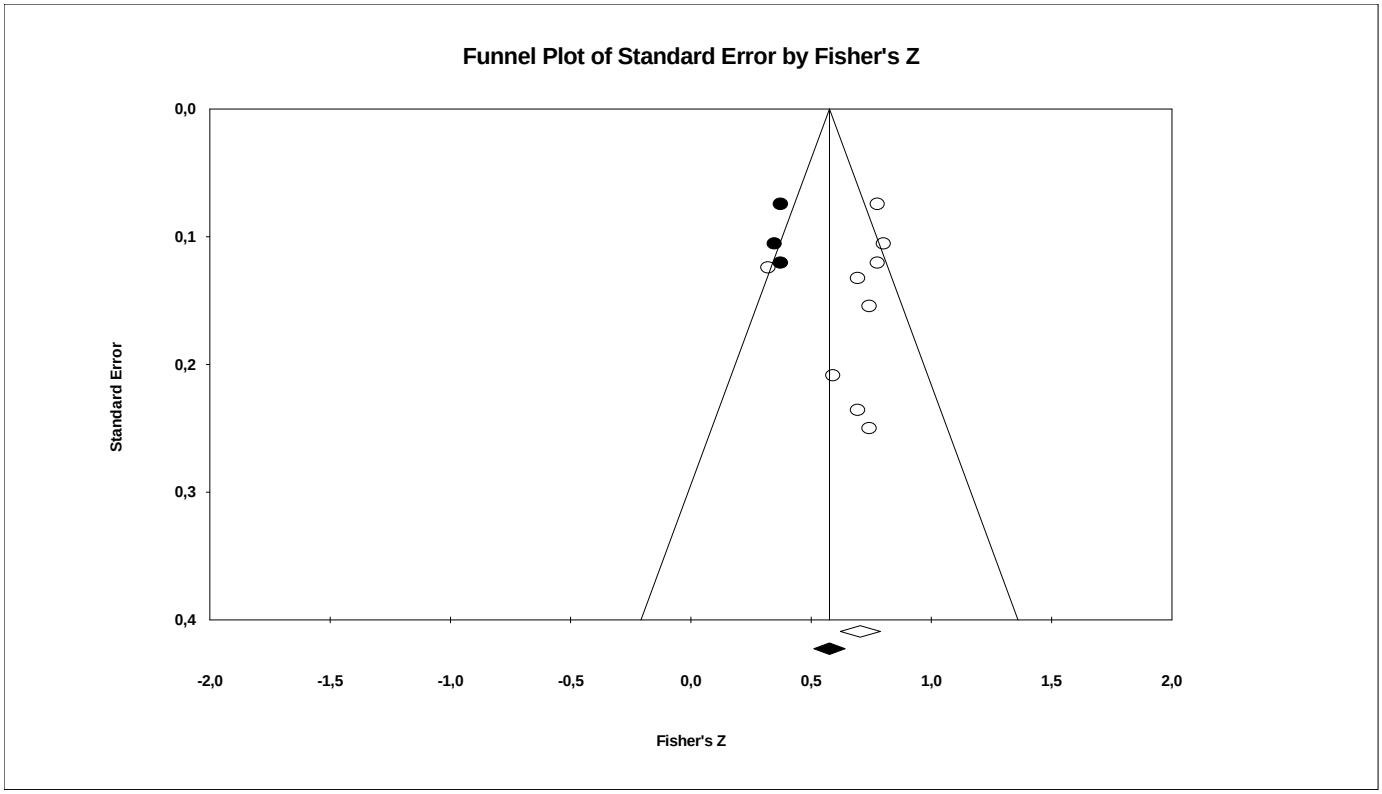


Figure E5. Sparling (1980); relative to fat-free weight; fixed-effect analysis.

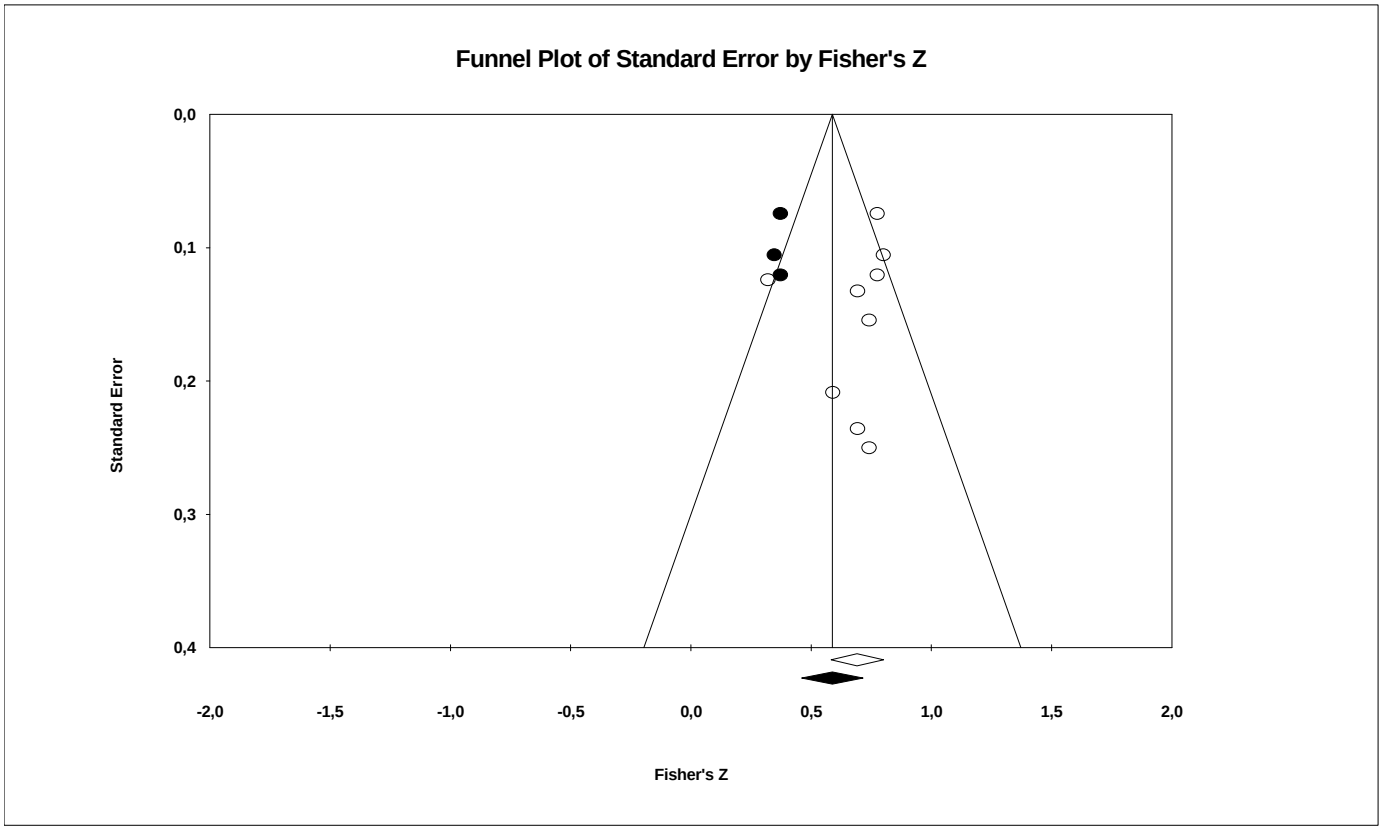


Figure E6. Sparling (1980); relative to fat-free weight; random-effects analysis.

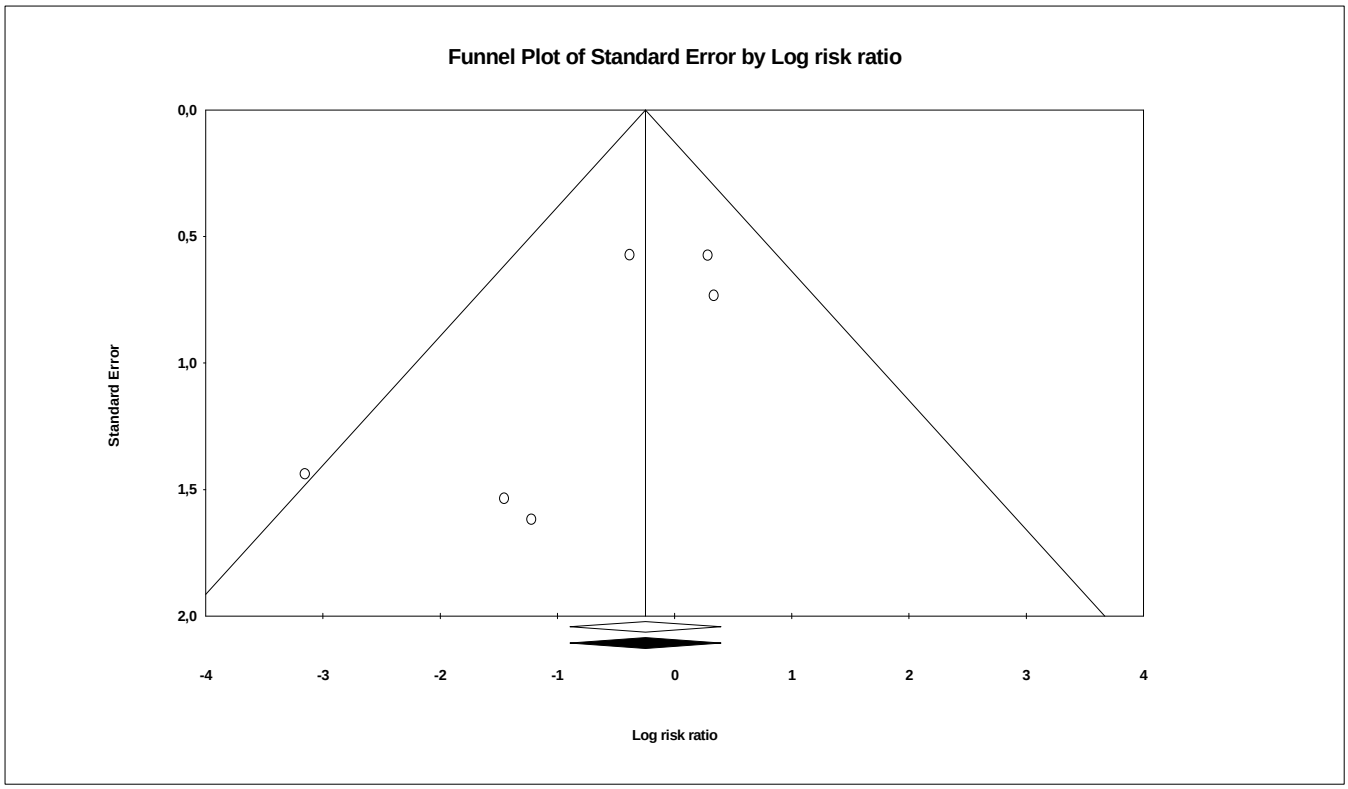


Figure E7. Desilva et al. (1981); fixed-effect analysis.

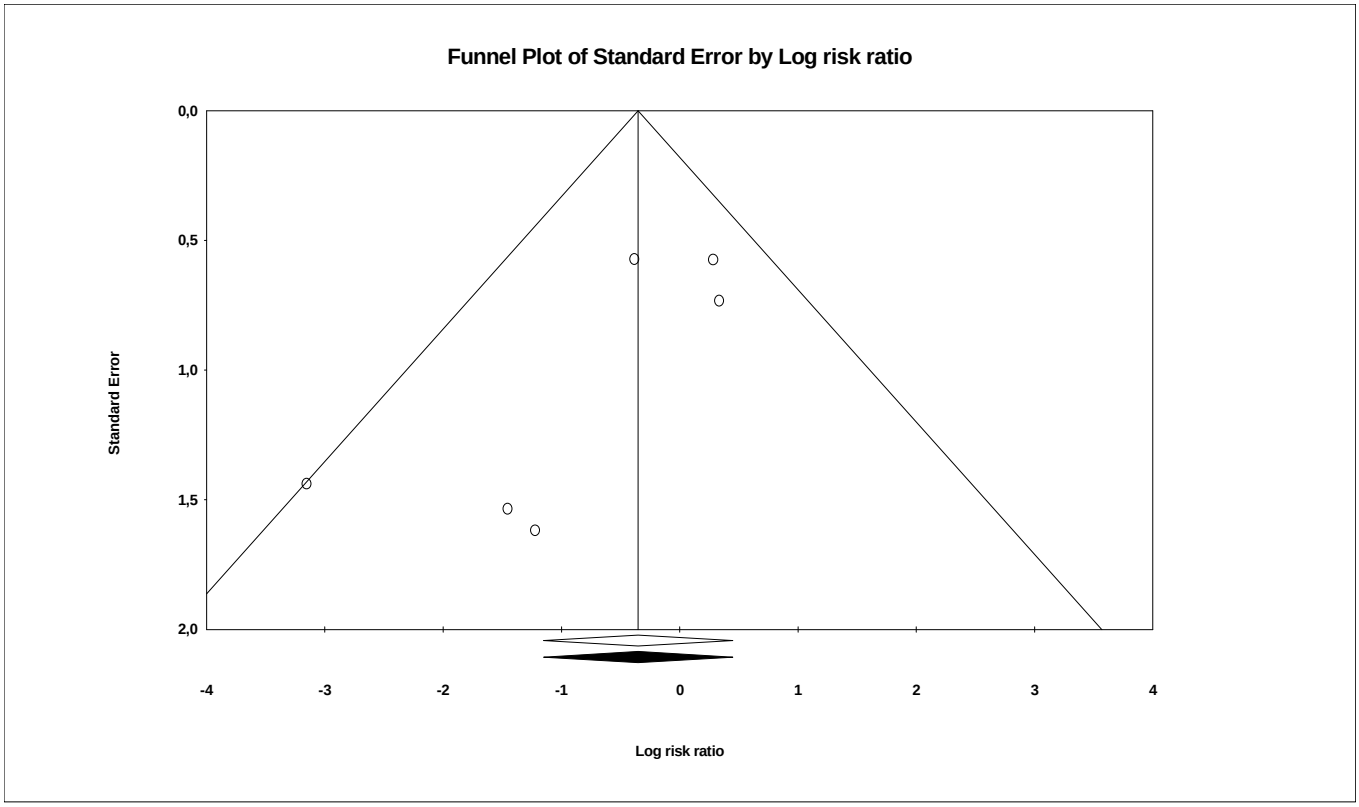


Figure E8. Desilva et al. (1981); random-effects analysis.

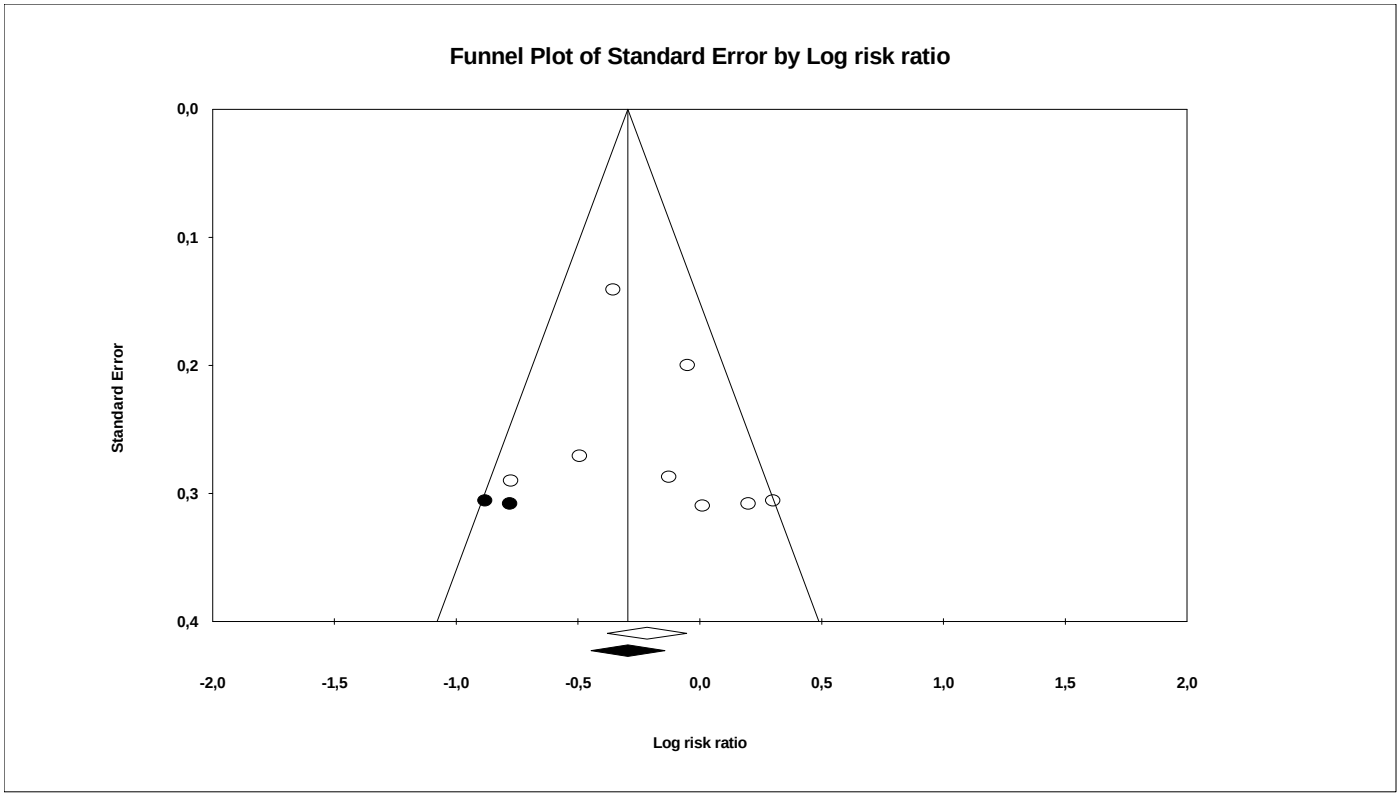


Figure E9. Stampfer et al. (1982); fixed-effect analysis.

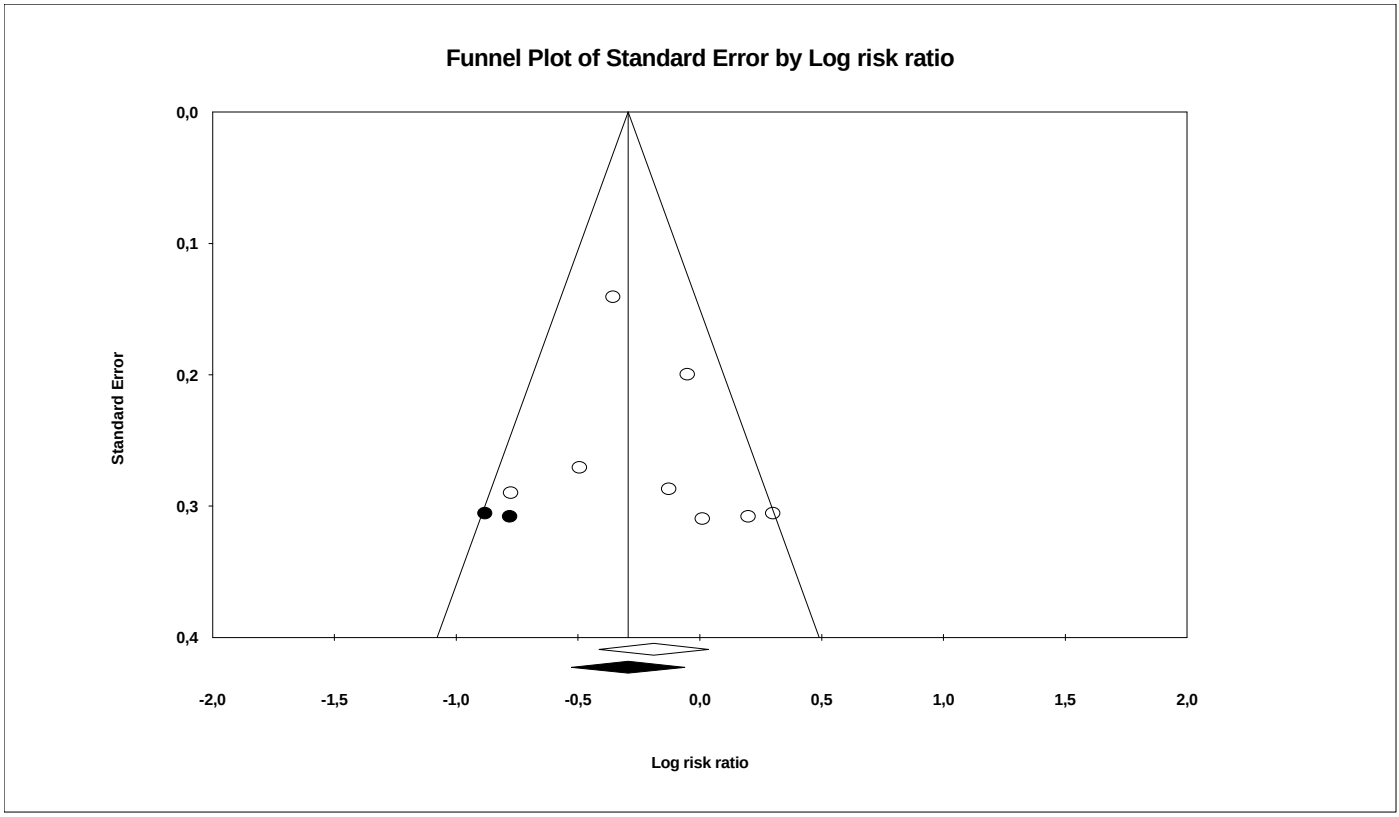


Figure E10. Stampfer et al. (1982); random-effects analysis.

References

Aber, C. P., Bass, N. M., Berry, C. L., Carson, P. H. M., Dobbs, R. J., Fox, K. M., … Stock,J. P. P. (1976). Streptokinase in acute myocardial infarction: A controlled multicenter study in the United Kingdom. *British Medical Journal, 2,* 1100-1104.

Allen, M., & Preiss, R. (1993). Replication and meta-analysis: A necessary connection.*Journal of Social Behavior and Personality, 8,* 9-20.

Amery, A., Roeber, G., Vermeulen, H. J., & Verstraete, M. (1969). Single-blind randomized multicenter trial comparing heparin and streptokinase treatment in recent myocardial infarction. *Acta Medica Scandinavica Supplementum, 505,* 5-35.

Andrews, G., Guitar, B., & Howie, P. (1980). Meta-analysis of the effects of stuttering treatment. *Journal of Speech & Hearing Disorders, 45,* 287-307.

Anonymous (1980). Aspirin after myocardial infarction. *Lancet, 1,* 1172-1173.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be?, *American Psychologist, 63,* 848-849.

Arkin, R. M., Cooper, H. M., & Kolditz, T. A. (1980). A statistical review of the literature concerning the self-serving attribution bias in interpersonal influence situations. *Journal of Personality, 48,* 435–448.

Berk, A. A., & Chalmers, T. C. (1981). Cost and efficacy of the substitution of ambulatory for inpatient care. *New England Journal of Medicine, 304,* 393-397.

Bett, J. H. N., Castaldi, P. A., Hale, G. S., Isbister, J. P., Mclean, K. H., O'Sullivan, E. F., … Rosenbaum, M. (1973). Australian multicenter trial of streptokinase in acute myocardial infarction. *Lancet, 1,* 57-60.

Blanchard, E. B., Andrasik, F., Ahles, T. A., Teders, S. J., & O'Keefe, D. (1980). Migraine and tension headache: A meta-analytic review. *Behavior Therapy, 11,* 613-631.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005). *Comprehensive meta-analysis, version 2.2.030.* Englewood, NJ: Biostat.

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

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

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4th ed.). Heidelberg, Germany: Springer.

Burger, J. M. (1981). Motivational biases in the attribution of responsibility for an accident: A meta-analysis of the defensive-attribution hypothesis. *Psychological Bulletin, 90,* 496-512.

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

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13,* 321-341.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (Eds.). (1992). *Meta-analysis for explanation: A casebook.* New York: Russell Sage.

Cooper, H. M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology, 37,* 131-146.

Cooper, H. M. (1984). *The integrative research review: A systematic approach*. Beverly Hills, CA: Sage.

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Cooper, H. M., Burger, J. M., & Good, T. L. (1981). Gender differences in the academic locus of control beliefs of young children. *Journal of Personality and Social Psychology, 40,* 562-572.

Cooper, H. M., & Rosenthal, R. (1980). Statistical versus traditional procedures for summarizing research findings. *Psychological Bulletin, 87,* 442-449.

Cordray, D. S. (1990). An assessment from the policy perspective. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 99-119). New York: Russell Sage.

Cotton, J. L., & Cook, M. S. (1982). Meta-analyses and the effects of various reward systems: Some different conclusions from Johnson et al. *Psychological Bulletin, 92,* 176-183.

Desilva, R. A., Hennekens, C.H., Lown, B., & Casscells, W. (1981). Lignocaine prophylaxis in acute myocardial infarction: An evaluation of randomised trials. *Lancet, 2,* 855-858.

Dioguardi, N., Lotto, A., Levi, G. F., Rota, M., Proto, C., Mannucci, P. M., … Agostoni, A. (1971). Controlled trial of streptokinase and heparin in acute myocardial infarction. *Lancet, 2,* 891-895.

Eagly, A. H., & Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: A meta-analysis of social influence studies. *Psychological Bulletin, 90,* 1-20.

European Cooperative Study Group For Streptokinase Treatment In Acute Myocardial Infarction. (1979). Streptokinase in acute myocardial infarction. *New England Journal of Medicine, 301,* 797-802.

Fricke, R. (1982). *Bibliographie zum Themenbereich "Metaanalyse"*. Bibliography (Arbeitsstelle für Unterrichtsforschung, Report No. 2), University Hannover, Germany.

Fricke, R., & Treinies, G. (1985). *Einführung in die Metaanalyse*. Bern, Switzerland: Huber.

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

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

Glass, G. V., & Smith, M. L. (1978). *Meta-analysis of research on the relationship of class-size and achievement.* Retrieved from ERIC database. (ED168129)

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis, 1,* 2-16.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin, 85,* 845-857.

Hattie, J. A., & Hansford, B. C. (1982). Self measures and achievement: Comparing a traditional review of literature with a meta-analysis. *Australian Journal of Education, 26,* 71-75.

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

Hedges, L. V., & Olkin, I. (1989). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Heikinheimo, R., Ahrenberg, P., Honkapohja, H., Iisalo, E., Kallio, V., Konttinen, Y., … Siitonen, L. (1971). Fibrinolytic treatment in acute myocardial infarction. *Acta Medica Scandinavica, 189,* 7-13.

Hereford, S. M. (1979). The Keller Plan within a conventional academic environment: An empirical "meta-analytic" study. *Engineering Education, 70,* 250-260.

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

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

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin, 86,* 721-735.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Hyde, J. S. (1981). How large are cognitive gender differences? *American Psychologist, 36,* 892-901.

Ide, J. K., Parkerson, J., Haertel, G. D., & Walberg, H. J. (1981). Peer group influence on educational outcomes: A quantitative synthesis. *Journal of Educational Psychology, 73,* 472-484.

Inglis, J., & Lawson, J. S. (1982). A meta-analysis of sex differences in the effects of unilateral brain damage on intelligence test results. *Canadian Journal of Psychology, 36,* 670-683.

Ingram, L. (1990). An overview of the desegregation meta-analyses. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 61-70). New York: Russell Sage.

Iverson, B. K., & Levy, S. R. (1982). Using meta analysis in health education research. *Journal of School Health, 52,* 234-239.

Iverson, B. K., & Walberg, H. J. (1982). Home environment and school learning: A quantitative synthesis. *Journal of Experimental Education, 50,* 144-151.

Jennions, M. D., & Möller, A. P. (2002). Relationships fade with time: A meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society of London B, Biological Sciences, 269,* 43-48.

Johnson, D. W., Maruyama, G., Johnson, R., Nelson, D., & Skon, L. (1981). Effects of cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis. *Psychological Bulletin, 89,* 47-62.

Kavale, K. (1981). Functions of the Illinois Test of Psycholinguistic Abilities (ITPA): Are they trainable? *Exceptional Children, 47,* 496-510.

Kazrin, A., Durac, J., & Agteros, T. (1979). Meta-meta analysis: A new method for evaluating therapy outcome. *Behaviour Research and Therapy, 17,* 397-399.

Kennedy, M. M. (1978). Findings from the follow through planned variation study. *Educational Researcher, 7,* 3-11.

Kremer, B. K., & Walberg, H. J. (1981). A synthesis of social and psychological influences on science learning. *Science Education, 65,* 11-23.

Kulik, C.-L. C., & Kulik, J. A. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. *American Educational Research Journal, 19,* 415-428.

Ladas, H. (1980). Summarizing research: A case study. *Review of Educational Research, 50,* 597-624.

Landman, J. T., & Dawes, R. M. (1982). Psychotherapy outcome: Smith and Glass' conclusions stand up under scrutiny. *American Psychologist, 37,* 504-516.

Levy, S. R., Iverson, B. K., & Walberg, H. J. (1980). Nutrition-education research: An interdisciplinary evaluation and review. *Health Education Quarterly, 7,* 107-126.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., … Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine, 6,* e1000100.

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

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

Mogensen, L. (1970). Ventricular tachyarrhytmias and lignocaine prophylaxis in acute myocardial infarction. *Acta Medica Scandinavica Supplementum, 513,* 1-80.

Mumford, E., Schlesinger, H. J., & Glass, G. V. (1982). The effects of psychological intervention on recovery from surgery and heart attacks: An analysis of the literature. *American Journal of Public Health, 72,* 141-151.

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

Muncer, S., Taylor, S., & Craigie, M. (2002). Power dressing and meta-analysis: Incorporating power analysis into meta-analysis. *Journal of Advanced Nursing, 38,* 1-7.

O'Brien, K. P., Taylor, P. M., Croxson, R. S. (1973). Prophylactic lignocaine in hospitalized patients with acute myocardial infarction. *Medical Journal of Australia Supplementum, 2,* 36-37.

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.

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

Press, S. J. (1990). Comments on the desegregation summary analysis. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 71-74). New York: Russell Sage.

R. Development Core Team (2011). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing Computer software.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Rosenthal. R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3,* 377-415.

Rosenthal, R., & Rubin, D. B. (1982). Further meta-analytic procedures for assessing cognitive gender differences. *Journal of Educational Psychology, 74,* 708-712.

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.

Schneider, J. M. (1990). Research, meta-analysis, and desegregation policy. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 55-60). New York: Russell Sage.

Schwab, D. P., Olian-Gottlieb, J. D., & Heneman, H. G. (1979). Between-subjects expectancy theory research: A statistical review of studies predicting effort and performance. *Psychological Bulletin, 86,* 139-147.

Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin, 92,* 581-604.

Smith, M. L. (1980). Sex bias in counseling and psychotherapy, *Psychological Bulletin, 87,* 392-407.

Smith, M. L., & Glass, G.V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752-760.

Smith, M. L., Glass, G.V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Baltimore, MD: Johns Hopkins University Press.

Smith, L. R., & Land, M. L. (1981). Low-inference verbal behaviors related to teacher clarity. *Journal of Classroom Interaction, 17,* 37-42.

Sparling, P. B. (1980). A meta-analysis of studies comparing maximal oxygen uptake in men and women. *Research Quarterly for Exercise and Sport, 51,* 542-552.

Stampfer, M. J., Goldhaber, S. Z., Yusuf, S., Peto, R., & Hennekens, C. H. (1982). Effect of intravenous streptokinase on acute myocardial infarction: Pooled results from randomized trials. *New England Journal of Medicine, 307,* 1180-1182.

Tornatzky, L. G., & Klein, K. J. (1982). Innovation characteristics and innovation adoption-implementation: A meta-analysis of findings. *IEEE Transactions on Engineering Management, 29,* 28-45.

Treinies, G., & Fricke, R. (1983). Deskriptive Methoden der Metaanalyse zur Integration experimenteller Studien, dargestellt an einem Beispiel aus der Instruktionspsychologie. Report (Arbeitsstelle für Unterrichtsforschung, Report No.5), University Hannover, Germany.

U.S. Department of Commerce Economics and Statistics Administration, U.S. Census Bureau. (n.d.). *Census Regions and Divisions of the United States*. Retrieved from http://www.census.gov/geo/www/us_regdiv.pdf

Wampler, K. S. (1982). Bringing the review of literature into the age of quantification: Meta-analysis as a strategy for integrating research findings in family studies. *Journal of Marriage and the Family, 44,* 1009-1023.

Williams, P. A., Haertel, E. H., Haertel, G. D., & Walberg, H. J. (1982). The impact of leisure-time television on school learning: A research synthesis. *American Educational Research Journal, 19,* 19-50.

Zusammenfassung

Ziel der vorliegenden Arbeit war einerseits die Überprüfung der Möglichkeit klassische Meta-Analysen (1977-1982) mittels State-of-the-Art Methoden zu reanalysieren.

Es wurden dafür a priori vier Kriterien aufgestellt, welche eine klassische Meta-Analyse erfüllen musste, um mittels modernen Methoden reanalysiert werden zu können. Die Studie musste Effektstärken metaanalytisch integrieren, Effektstärken mussten für jede inkludierte Primärstudie berichtet werden (oder berechenbar sein), berichtete Effektstärken mussten heute gebräuchlichen entsprechen (oder von Meta-Analyse Software verwertbar sein) und zuletzt musste die Varianz (oder Standardfehler) für jede inkludierte Primärstudie berichtet werden (oder berechenbar sein). Zusätzlich wurde nach Durchsicht aller in Frage kommenden Studien ein fünftes Kriterium aufgestellt, welches spezifische, über die vier genannten hinausgehende Voraussetzungen beinhaltete.

Andererseits sollten jene klassischen Meta-Analysen, welche alle notwendigen Voraussetzungen erfüllten, reanalysiert werden. Es wurden zu diesem Zweck die in der Originalstudie berichteten Daten herangezogen und moderne Methoden zur Beantwortung der ursprünglichen Forschungsfragen darauf angewendet. Ziel war, Schlussfolgerungen der Reanalyse mit jenen der Originalstudie zu vergleichen, um zu überprüfen, ob die Anwendung von State-of-the-Art Methoden neue oder andere Schlussfolgerungen zuließ.

Lediglich drei von 78 klassischen Meta-Analysen erfüllten alle notwendigen Voraussetzungen, um mittels State-of-the-Art Methoden reanalysiert werden zu können. Die Reanalysen erbrachten teils Ergebnisse, die mit den ursprünglich gezogenen Schlussfolgerungen übereinstimmten, teils diesen aber widersprachen oder sie erweiterten.

Die vorliegende Arbeit stellt einen ersten Schritt dar, herauszufinden, ob die Anwendung moderner Methoden auf Originaldaten früher Meta-analysen neue oder andere Einsichten erbringt. Darauf aufbauend könnten einerseits vor 1977 publizierte Studien, welche Vorläufer meta-analytischer Methoden anwendeten, und andererseits nach 1982 publizierte Meta-analysen bezüglich Reanalysierbarkeit überprüft und bei Vorliegen der notwendigen Voraussetzungen reanalysiert werden.

Eidesstattliche Erklärung

Ich bestätige, die vorliegende Diplomarbeit selbst und ohne Benutzung anderer als der angegebenen Quellen verfasst zu haben. Weiters ist sie die Erste ihrer Art und liegt nicht in ähnlicher oder gleicher Form bei anderen Prüfungsstellen auf. Alle Inhalte, die wörtlich oder sinngemäß übernommen wurden, sind mit der jeweiligen Quelle gekennzeichnet.

Wien, November 2011 _____

Barbara Zimmermann

Curriculum Vitae

Name:	Barbara Zimmermann
Geburtsdatum:	09.02.1988
Geburtsort:	Wien
Staatsbürgerschaft:	Österreich
Familienstand:	Ledig

SCHULISCHER WERDEGANG

1994-1998:	Volksschule Mondweg
1998-2006:	BG13 Fichtnergasse, neusprachlicher Zweig (Französisch, Latein)
2006-2007:	Studium der Rechtswissenschaften an der Universität Wien (Abschluss des 1. Abschnitts)
Seit 2007:	Studium der Psychologie an der Universität Wien Voraussichtlicher Abschluss Jänner 2012

BERUFSERFAHRUNG

07/2005	Praktikum bei Junkers/Robert Bosch AG - Thermotechnik im Bereich Marketing
07/2007	Rechtshörerschaft im Bezirksgericht Fünfhaus
10/2008-08/2010	Ehrenamtliche Mitarbeit im Nachbarschaftszentrum 6 des Wiener Hilfswerks (Lernhilfe für Kinder mit Migrationshintergrund)
05/2009 – 12/2009	Studentische Hilfskraft bei ISG Personalmanagement (Unterstützung des Bereichs Training & Personalentwicklung)
05/2010 - 08/2010	Pflichtpraktikum im Rahmen des Psychologiestudiums (240 Stunden) im Nachbarschaftszentrum 6 des Wiener Hilfswerks