

MetaWin

Statistical Software for Meta-Analysis

version 2.0

Michael S. Rosenberg¹
Dean C. Adams^{1,2}
Jessica Gurevitch¹

¹ *Department of Ecology and Evolution
State University of New York
Stony Brook, NY 11794-5245*

² *Department of Zoology and Genetics
Iowa State University
Ames, IA 50011*

Table of Contents

1. Getting Started	1
1.1 What is MetaWin?	1
1.2 What's new in version 2.0.....	2
1.3 Hardware Requirements and Installation.....	3
<i>Potential Problem with Windows System</i>	3
1.4 Overview of Manual	4
1.5 Basic Information	5
1.6 Technical Support and How to Cite MetaWin.....	6
2. Introduction to Meta-Analysis.....	8
2.1 Why Perform Meta-Analysis?	8
<i>Vote-Counting Methods</i>	9
<i>Combined Probability Methods</i>	10
<i>Modern Meta-Analysis</i>	12
<i>Summary</i>	13
2.2 Effect Size Calculations.....	14
<i>Means and Standard Deviation Data</i>	14
<i>Two x Two Contingency Data</i>	17
<i>Correlation Data</i>	20
<i>Z</i>	20
<i>Other Effect Sizes</i>	21
<i>Summary</i>	21
2.3 Meta-Analysis Summary Statistics	23
<i>Cumulative Effect Size and Total Heterogeneity</i>	24
<i>Incorporating Data Structure in Meta-Analysis</i>	25
<i>Fixed-Effects Models and Random-Effects Models</i>	30
2.4 A General Model for Meta-Analysis.....	33
<i>General Linear Models (GLM) for Meta-Analysis</i>	33
<i>Generalizing Odds Ratio Computations</i>	35
2.5 Summary	37
3. Additional Issues, Analyses and Calculations	38
3.1 Publication Bias and Data Exploration.....	38
<i>Does Publication Bias Exist?</i>	38
<i>Graphical Data Exploration</i>	40
<i>Rank Correlation Tests for Publication Bias</i>	42
<i>Fail-Safe Numbers</i>	43
3.2 Cumulative Meta-Analysis.....	45
3.3 Resampling Methods in Meta-Analysis	47
<i>Bootstrapping</i>	48
<i>Randomization Tests</i>	48
4. Running MetaWin	50
4.1 Data Spreadsheet.....	51

<i>Data Structure</i>	52
<i>Loading Data</i>	53
<i>Manipulating Data</i>	54
<i>Saving and Printing Data</i>	57
4.2 Output Window	58
4.3 Program Options	59
5. Conducting a Meta-Analysis.....	61
5.1 Calculating Effect Sizes.....	61
<i>Means and Standard Deviation Data</i>	62
<i>Two x Two Contingency Data</i>	64
<i>Correlation Data</i>	64
<i>Results from Effect Size Calculations</i>	65
5.2 Calculating Meta-Analysis Summary Analyses	66
<i>Basic Options Tab</i>	67
<i>Resampling Tests Tab</i>	68
<i>Refine Categories Tab</i>	69
<i>Refine Studies Tab</i>	70
<i>Graphical Output Tab</i>	71
<i>Miscellaneous Analyses Tab</i>	72
<i>Results from Summary Analyses Calculations</i>	73
6. Miscellaneous Analyses/Options	80
6.1 Data Exploration.....	80
<i>Graphical Methods Tab</i>	81
<i>Publication Bias Tab</i>	81
<i>Fail-Safe Tab</i>	82
<i>Results from Data Exploration</i>	83
6.2 Graphics	85
<i>Draw Graph Window</i>	85
<i>General Plot Options</i>	91
7. MetaWin Statistical Calculator	94
7.1 What is MetaCalc?.....	94
7.2 Statistical Functions.....	95
<i>Variance Functions</i>	95
<i>Probability Functions</i>	96
<i>Correlation Functions</i>	97
<i>Mean Difference Functions</i>	99
Appendix I: Example Data Sets	102
Gurevitch and Hedges (1993) Data Set	102
Berlin et al. (1989) Data Set.....	104
Appendix II: Warning and Error Messages.....	106
Glossary.....	112
References	123

ACKNOWLEDGMENTS

As with any revision, the second version of MetaWin has benefited from the suggestions and comments of many people. First and foremost we would like to thank Sinauer Associates, and Andy Sinauer in particular, for allowing us the opportunity to publish the second version of this software. Without his support we would never have undergone such a major revision. We also would like to thank the loyal users of version 1.0 (you know who you are) who have sent us many comments of support and provided suggestions for improvements. The many beta-testers (both individuals and meta-analysis workshop participants) that “crash-tested” early releases of version 2.0 deserve a special thanks; they endured a long list of “features” and strange errors that can only be found from careful plodding through the seemingly endless list of options that only beta-testers can find. Finally, we thank Larry Hedges for providing us with *in press* chapters from his forthcoming book.

Because MetaWin is a software package, we also would like to acknowledge those who have helped in software development. In particular, F. James Rohlf has been a consistent source of comments, hints and recommendations for the improvement of our programming. In addition, the routines and mathematical functions available in several published works greatly sped up our work, and allowed us to be biostatisticians rather than computer scientists. Among these are: Walden, 1986; Winer and Valley, 1988; Press et al., 1989; Born, 1995; and Microsoft Press, 1997.

1. GETTING STARTED

1.1 What is MetaWin?

This program will allow one to summarize the results of multiple independent studies using meta-analytic procedures. Version 2.0 is much more general than its predecessor, and allows for greater flexibility in both the effect sizes that can be used as well as the statistical models for summarizing meta-analytic data (see Section 1.2 below). This version can calculate both fixed effects models and random effects models, and can be used for a variety of meta-analytic data structures, including no data structure, categorical (grouped) data, and continuous (regression) data. MetaWin now comes with its own spreadsheet; data can be entered directly into the spreadsheet or can be read from a text file, a Microsoft Excel file, or a Lotus 123 file. A variety of commonly used meta-analysis effect sizes can be calculated, including Hedges' d , *response ratio*, *odds ratio*, *risk difference*, *relative risk*, and Fisher's z-transform (for details see Section 2.2, pp. xx). From these (or other) effect sizes, cumulative mean effects, their confidence intervals, and various heterogeneity statistics can be calculated. The total heterogeneity, Q_T , is calculated for each analysis, and, for categorical and continuous data models, the heterogeneity explained by the model, Q_M , and the residual error heterogeneity, Q_E , are calculated as well (for a discussion of these statistics, see Section 2.3, pp. xx).

MetaWin 2.0 also allows one to refine the analysis by removing certain studies or groups of studies from the analysis without having to alter the data file. A variety of exploratory data analyses can be performed, including various tests to evaluate potential publication bias. Cumulative meta-analyses can also be performed, in order to investigate changes in the cumulative mean effect size as new studies are added to the model. In addition, one can graphically explore the data through histograms, normal quantile plots and funnel plots. Scatter plots, regression plots, radial plots, and plots of cumulative mean effect sizes can also be generated.

Finally, because data may violate the underlying assumptions of meta-analysis, it may be useful to evaluate the significance of meta-analytic statistics using resampling methods (Adams et al., 1997). Therefore, this program will allow one to incorporate resampling tests into the meta-analysis. In particular, confidence intervals for cumulative mean effect sizes can be generated using two different bootstrap procedures (bootstrap confidence intervals and bias-corrected bootstrap confidence intervals). MetaWin also allows one to test the significance of the heterogeneity explained by the model, Q_M , using a randomization test (for a discussion of these statistics, see Section 3.3, pp. xx).

1.2 What's new in version 2.0

Many new analyses and features are found in MetaWin 2.0 that were not present in the previous version. In the first version of MetaWin (Rosenberg et al., 1997), one could load data from a specifically-structured text file, and calculate two different effect sizes (Hedges' d , $\ln \text{response ratio}$) from means, standard deviations, and sample size data (however, one could perform a meta-analysis on other effect sizes by using the "effect file format", in which the user provided MetaWin the effects and their weights). Effect sizes could then be statistically summarized in a categorical meta-analysis using either a fixed-effects or mixed/random effects-model. If the data were specially coded, a meta-analysis with no group structure could also be performed. The program allowed one to refine their data analysis, so that individual studies or groups could be excluded. The program was limited in the number of studies (500), the number of independent variables (10), the number of groups per variable (25), and the number of iterations for resampling tests (19999).

In contrast, Version 2.0 of MetaWin contains a spreadsheet editor, and one can load data from a text file, a Microsoft Excel file, or a Lotus 123 file. These files can now be in free matrix format, and data in the editor can be corrected and altered as needed. In addition, one can enter data directly into the spreadsheet for editing. The current version of MetaWin calculates effect sizes from three different types of primary data: (1) means, standard deviations, and sample sizes, (2) two x two contingency tables, and (3) correlation coefficients, and six different effect sizes can be calculated from these data. Like version 1.0 however, the user can provide data columns for effect sizes and their variances that are not directly calculated in MetaWin.

The meta-analytic summary analyses in version 2.0 are much more flexible than in the first version of MetaWin. Both fixed-effects and random-effects models can be calculated for three different types of data structure: data from one population (i.e., no structure), categorical data, or continuous data. Group mean effect sizes are calculated for categorical models, and regression coefficients (slope and intercept) are calculated for continuous models. Like version 1.0, the analysis can be refined, and groups or individual studies can be eliminated from the analysis. Additionally, this version presents graphical summaries of the meta-analysis, including plots of group mean effect sizes (for categorical models) and regression plots (for continuous models). Other graphical methods can be performed as well, including funnel plots, normal quantile plots, histograms, scatter plots, radial plots, and plots of individual effect sizes.

A set of exploratory data analysis tools is now included in MetaWin, including methods for exploring publication bias, as well as various fail-safe methods. A statistical calculator is included, so that one may transform test-statistics from various studies into a common metric (e.g., correlation coefficients) that may be used to generate effect sizes. Finally, the underlying data architecture of MetaWin has been substantially improved,

which has resulted in the elimination of many of the program limitations found in version 1.0. There is still a limit on the number of iterations that can be performed for resampling tests (now 64,999), but the number of independent variables, the number of studies, and the number of groups is limited only by the amount of available space on one's hard drive.

1.3 Hardware Requirements and Installation

To run MetaWin 2.0, you must have an IBM-compatible computer running Windows95 or better, with a minimum of 8MB of RAM. The program comes on a CD, with several subdirectories including a Manual subdirectory containing a PDF version of the manual and a Tools subdirectory containing a Windows system patch (see below).

To install MetaWin, choose Run from the Start Menu and enter "X:\Setup" where X is your CD-Rom drive, or double-click Setup.exe in the root directory of the CD using Windows Explorer. This will begin the installation process.

The installation of MetaWin is self-explanatory. At one stage, you will be asked to enter your name, company, and a serial number. The serial number can be found on the packing slip and on the CD. MetaWin will not run without a valid serial number. A successful installation will put the executable and help files in a directory of your choice and example data sets in a "Data" subdirectory. Program icons will be created on the Start Menu which can be used to run the program. See the ReadMe file (found in the same directory as the executable) for last minute information that could not be included in the manual.

To uninstall MetaWin, choose Add/Remove Programs from the Control Panel, find MetaWin on the program list, and press the Add/Remove button.

Potential Problem with Windows System

Problem: The toolbars on various MetaWin windows appear blank. The toolbar buttons are not displaying their icons properly.

Cause: There is a known incompatibility between MetaWin 2.0 and early versions of a Windows system file COMCTL32.DLL (usually found in the Windows/System directory of your hard drive). Older releases of this file contained bugs (including Y2K problems) which Microsoft has since fixed and updated. You can find out which version you have by clicking on the file in Windows Explorer with the right mouse button, choosing Properties, then clicking on the Version tab. MetaWin 2.0 requires version 4.72 or later. If you have Windows98 or Internet Explorer 4.01 (or later) on your system, you should already have an updated version of this file. If you have an older

version of COMCTL32.DLL and want to update it, you may do so by running 401COMUPD.EXE found in the Tools directory on the MetaWin CD, or by downloading the latest version of this file from the Microsoft website. Updating your system or installing newer versions of Microsoft Internet Explorer will also update the file. If you are running Windows NT 4.0, do not try to update this file until after installing the Service Pack 3 or later (see Microsoft's website for more details). People using Windows98 Second Edition, Windows2000, and Internet Explorer 5.x should not run 401COMUPD.EXE.

1.4 Overview of Manual

This manual is divided into two main sections. The first section (Chapters 2 – 3) describes the basic theory and statistical models of meta-analysis. In Chapter 2 we discuss why one wishes to perform a meta-analysis and present a brief summary of methods for synthesizing results from multiple studies, including vote-counting, combining probability values, and modern meta-analysis. Next is a comprehensive explanation of the basic calculations necessary for conducting a meta-analysis. This includes a discussion of effect sizes, and which effect sizes can be calculated from the various sorts of meta-analytic data. We then discuss summary meta-analysis statistics, cumulative mean effect sizes, and heterogeneity statistics. We describe the different methods for incorporating various data structures, and compare and discuss the use of fixed effects models and random effects models. Finally, we present a general model for meta-analysis, and show how methods from different disciplines can be considered part of this larger, more general framework.

In Chapter 3 we discuss other extensions of meta-analysis, including methods of data exploration and the detection of publication bias. We present the various fail-safe methods, as well as graphical means by which meta-analysts evaluate the robustness of the data. We also discuss cumulative meta-analysis, and why it may be useful. Finally, we describe the theory and procedures of resampling methods, including randomization and bootstrapping, and their place in meta-analysis.

The second section of this manual (Chapters 4 – 7) explains how MetaWin can be used to perform a meta-analysis. In Chapter 4 we describe how to start MetaWin, how to change the default options, explain the MetaWin spreadsheet, data entry and manipulation, and the MetaWin output window. In Chapter 5 we describe how to calculate effect sizes from various sorts of data, and how to perform the summary analyses of meta-analysis, including how to perform resampling tests and how to refine an analysis by removing individual studies or groups. Chapter 6 describes the miscellaneous analyses offered in MetaWin, including data exploration, fail-safe methods, methods for detecting publication bias, and all graphical output. Finally, in Chapter 7, we explain how the MetaWin statistical calculator can be used to convert the

statistical results of individual studies into common effect sizes. The manual also has appendices containing example data sets and warning and error messages, a glossary, an index, and references.

1.5 Basic Information

Though more comprehensive than the previous version, this manual does **not** provide sufficient information for someone unfamiliar with meta-analysis to learn everything necessary to carry out a good meta-analysis! Meta-analysis involves various complex and subtle considerations, some of which are quite controversial. While this program makes it easy to carry out the calculations involved, that also means that it becomes more easy to fall into some serious pitfalls without realizing what is happening. Unfortunately, there are few courses in meta-analysis, but there are several excellent texts to get you started. For general background information on meta-analysis and methods for research synthesis, we direct the reader to the following comprehensive texts on meta-analysis: Hedges and Olkin (1985; 2000), Hunter and Schmidt (1990), Rosenthal (1991), Cooper and Hedges (1994a), and Cooper (1998). We mention some of the things one needs to think about below.

- What are the steps of a meta-analysis?

There are many ways to approach a meta-analysis and the books listed above present a good framework for conducting an analysis. Here we summarize the basic steps:

1. Define Question(s) to be Answered
2. Conduct a Literature Search and Collect Data*
3. Choose an Appropriate Effect Size and Analytic Model(s)
4. Calculate Effect Sizes
5. Explore Statistical Properties/Characteristics of Effect Sizes
6. Conduct Summary Analyses
7. Examine and Interpret Results

* Meta-analysis can be used for quantitative syntheses of both published and unpublished data (e.g., Rosenthal and Rubin, 1978).

MetaWin can be used to perform steps 4 through 6.

- What is a “control?”

In carrying out a meta-analysis, one usually compares the performance of two groups: generally, a “control” group is compared with an “experimental” or “treatment” group. Sometimes, the identity of the groups is clearly consistent across studies. Often,

however, the person carrying out the meta-analysis must decide which group is comparable as a “control” across studies. For example, Gurevitch et al. (1992), in examining the effects of competition in field experiments, decided that the “controls” were those organisms which were kept at natural density, while the “experimental” group in each study was manipulated to be above or below natural density. It is the meta-analyst, not the author of each individual study, that must make the determination regardless of what a group might be called by the authors of each study. While the absolute value of d (and some other measures of effect size) will be the same regardless of which group is called the experimental and which one the control, there are other things that are easier to examine in a meta-analysis when the “controls” are consistent from one study to another (see Gurevitch et al., 1992 for several examples).

- What shall I do if there is more than one measure of outcome (e.g., growth *and* fecundity; reading scores *and* math scores, etc.) for each experiment?

One way to think about the purpose of meta-analysis is that, where there is a body of experimental data that measures the same effect in different ways, this approach offers the tools to examine the magnitude of that effect and its consistency among studies. It is almost never the purpose of meta-analysis to combine studies that are measuring things that are fundamentally different. In almost all cases, it does not make *scientific* sense to combine these sorts of fundamentally different measures in a single meta-analysis. Furthermore, these measurements are likely to be non-independent when they are measured on the same individuals in a study. Therefore, it is usually best *not* to combine different basic kinds of outcome in a single meta-analysis, unless they are similar enough that one believes that they are actually measuring essentially the same thing (e.g., biomass and % cover for plants, where one is essentially a surrogate for the other). One option when several measures are available is to carry out separate meta-analyses on each of the common measures of outcome.

There is a related issue, and that is what to do when there is more than one experimental outcome reported in a paper, but there is a single measure of outcome. For example, a paper might report the effects of predation on two different species of prey. Should one include the responses of each of these species? Many meta-analysts would choose to do so, despite the problem of lack of independence, because the more conservative approach of using only a single measure sacrifices too much information. This is an important and substantive issue for the person carrying out the analysis to think through with care.

1.6 Technical Support and How to Cite MetaWin

A World Wide Web site for MetaWin is located at:

<http://life.bio.sunysb.edu/ee/meta> **WE NEED AN ADDRESS FROM SINAUER!!**

This site contains information on the latest version of MetaWin, answers to frequently asked questions, and information on how to contact the authors. When requesting information, please provide information on the type of microprocessor you are using, the amount of RAM, and the operating system of your computer, a brief description of the problem, and, if possible, the data set you had problems running. This information will allow us to more quickly diagnose your problem.

If you have used MetaWin to perform your analysis, please cite the program in the following manner:

Rosenberg, M. S., D. C. Adams, and J. Gurevitch. 1999. *MetaWin: Statistical Software for Meta-Analysis*. Version 2.0. Sinauer Associates, Sunderland, Massachusetts.

2. INTRODUCTION TO META-ANALYSIS

2.1 Why Perform Meta-Analysis?

One of the most commonly asked questions of researchers unfamiliar with meta-analysis is “why should I perform a meta-analysis?” To answer this question we must put meta-analysis in the proper context. Although a complete historical account of meta-analysis (see Hunt, 1997) and its methodological predecessors is beyond the scope of this manual, a brief discussion of the roots of meta-analysis is necessary for gaining perspective on this important field.

Science is normally performed in two different ways. The most common method is through primary research studies, where original data addressing a particular hypothesis are presented and evaluated. Conversely, one may assess a hypothesis through a research synthesis, where the results from a set of published studies are evaluated in terms of the same hypothesis. In fact, because science is cumulative, the research synthesis is an integral part of all science, including primary research, for one must review what is currently known about a topic before expanding our knowledge to what is unknown. Until recently, however, most research syntheses were qualitative and narrative, where the results of primary studies were subjectively assessed with respect to the hypothesis of interest (e.g., the effect of smoking on cancer incidence), and the author decided whether the published evidence tended to support or refute the hypothesis.

Although narrative reviews are useful summaries of the knowledge in a discipline, a quantitative means of assessing the evidence for or against a particular hypothesis is often desirable. A quantitative research synthesis that analyzes the results of a set of analyses is called a *meta-analysis* (Glass, 1976). (Note: Although all quantitative research syntheses fall under the general heading of ‘meta-analysis’ we shall reserve the use of that term for the modern meta-analytic methods). The earliest known quantitative research synthesis was performed by Pearson (1904), who calculated the average correlation coefficient from several studies to assess the effectiveness of a vaccine against typhoid (Cooper, 1998). Other quantitative methods for combining the results of independent studies were soon developed, including methods for combining probability values (e.g., Tippet, 1931; Fisher, 1932). Unfortunately, these techniques were rarely used in research summaries.

Following these early methods of quantitative research synthesis, little progress was made until the 1970s, when several researchers independently developed methods for statistically combining the results of independent studies. Unlike previous methods, these meta-analytic techniques statistically combined measures of effect from each individual study (e.g., Glass, 1976; Smith and Glass, 1977; Rosenthal and Rubin, 1978).

Thus these methods could not only determine whether an effect was present in a set of studies, but also estimate the *magnitude* of that effect. Throughout the 1970s and 1980s methodological improvements to the basic meta-analysis were continually being developed, and the field of meta-analysis soon became its own sub-discipline within statistics. Several summary texts were written on the subject (see e.g., Light and Pillemer, 1984; Rosenthal, 1984; Hedges and Olkin, 1985), which helped galvanize the statistical and philosophical approaches for performing a meta-analysis (for a more complete history of meta-analysis, see Cooper and Hedges, 1994b; Hunt, 1997).

Today, methods for quantitative research syntheses and specifically meta-analytic methods are common statistical tools used in the medical sciences, the social sciences, and some areas of the biological sciences. Most quantitative research syntheses employ methods that fall into one of three categories: methods for vote-counting, methods for combining probabilities, and modern meta-analytic methods. These techniques vary greatly in their quantitative complexity, as well as in their statistical power. Although MetaWin only performs meta-analytic methods, we briefly discuss each of the three methods for research synthesis below, to provide perspective on the range of methods available for research synthesis, as well as to compare the relative strengths and weaknesses of meta-analysis and its alternatives.

Vote-Counting Methods

The simplest method of synthesizing primary research results is known as vote-counting (Light and Smith, 1971; Hedges and Olkin, 1980). With this method, one observes the statistical results of a set of studies in relation to a given hypothesis. The results are then placed in one of three categories: statistically significant in the expected direction (i.e., positive findings), statistically significant in the unexpected direction (i.e., negative findings), and not statistically significant. The proportion of findings in each of the three categories is then determined, and the hypothesis is evaluated based on these proportions. The category with the largest proportion is asserted to be the statistical trend summarizing the primary literature, and is used as evidence to support a given hypothesis, or is used as evidence refuting the hypothesis.

Although the method described above is by far the most-commonly used method of vote-counting (termed conventional vote-counting), several alternative procedures have been developed. For example, for studies whose results fall into one of two categories, a simple sign test can be used, where the proportion of a significant results is tested against a binomial distribution. Methods of obtaining confidence intervals for vote-counting procedures using both equal and unequal sample sizes have also been developed (e.g., Gibbons et al., 1977; Hedges and Olkin, 1980), which can then be used to statistically assess the significance of the overall effect. Finally, vote-counting

procedures have been developed to detect publication bias in a set of primary research studies (for a complete discussion of vote-counting methods see Bushman, 1994).

The main advantage of conventional vote-counting methods is that they are simple and straightforward, and thus have intuitive appeal. Vote-counting has been used in research summaries in the social sciences (e.g., Levin and Nalebuff, 1995; Schmidt et al., 1995), the medical sciences (e.g., Balas et al., 1996), and the biological sciences (e.g., Connell, 1983; Schoener, 1983; Møller, 1997; Clarke, 1998). However, there are several statistical difficulties with these procedures. First, vote-counting tends to be overly conservative (Hedges and Olkin, 1980). Furthermore, because the expected number of non-significant findings can be greater than the expected number of significant positive or negative findings, vote-counting tends to have low statistical power (Cooper, 1998) and thus may not be able to detect the real treatment effects (Hunter et al., 1982). A third limitation is that the number of significant outcomes has little necessary relation to the magnitude of an effect and cannot provide this critical information. For these reasons, an approach that is statistically more powerful, and not sensitive to sample size effects, is desirable.

Combined Probability Methods

An alternative to vote-counting procedures are methods for combining probability values. These methods have a long history (e.g., Tippet, 1931; Fisher, 1932; Pearson, 1933), and may be considered the precursors to modern meta-analysis. All of these methods combine statistical results from a set of studies based on exact probability values (or transformations of probability values) to provide an overall assessment of significance. They are often referred to as *omnibus* tests, because they only depend on the exact probabilities from each individual study, rather than the distribution of the underlying data (Hedges and Olkin, 1985).

One advantage of employing combined probability methods over vote-counting is that these methods use exact probabilities, so the different sample sizes of each study are taken into account in their statistical summary of primary research results. Over the years many methods have been developed, which are usually classified according to the statistical distribution employed (for reviews see Rosenthal, 1978, 1991; Becker, 1994). At least eighteen methods currently exist, which use the following statistical distributions: the uniform distribution, normal distribution, Student's *t*-distribution, χ^2 distribution, and logistic distribution. We discuss a few of the more commonly used methods below (for mathematical descriptions of all methods see Rosenthal, 1991 and Becker, 1994).

Minimum *p* method: This method, originally developed by Tippet (1931) utilizes the uniform distribution, and is among the simplest of the methods for combining probabilities. Using this method, one concludes that there is evidence of a significant

effect if *any* one study is significant at the critical α -level. The smallest probability value is found and compared to the critical value:

$$\mathbf{a} = 1 - (1 - \mathbf{a}^*)^{1/n} \quad (2.1)$$

where n is the number of studies and \mathbf{a}^* is the overall significance level (usually 0.05).

Sum of logs method: This method was developed by Fisher (1932) and uses the inverse of the c^2 distribution to determine significance. The test is performed by calculating the statistic P as:

$$P = -2\sum_1^n \log(p_i) \quad (2.2)$$

where n is the number of studies and p_i are the individual significance values for the n studies. P is then tested against a c^2 distribution with $2n$ degrees of freedom to assess whether there a significant effect is present in the data.

Sum of Z method: This method is the most commonly-used method in the social sciences (Becker, 1994) and makes use of a normal distribution. Proposed by Stouffer et al. (1949), the statistic Z is found as:

$$Z = \frac{\sum_1^n Z(p_i)}{\sqrt{n}} \quad (2.3)$$

where n is the number of studies and $Z(p_i)$ are the Z -scores for the individual significance values. Z is then tested against the critical value based on a normal distribution.

Sum of p method: Edgington (1972) described a method by which one may sum significance values from various studies to test whether an overall effect was present. This method, like the minimum p method, utilizes a uniform distribution. The statistic P is calculated as:

$$P = \frac{\left(\sum_1^n p_i\right)^n}{n!} \quad (2.4)$$

where n is the number of studies and p_i are the individual significance values. P is then tested against a critical α (usually 0.05).

The methods presented above represent the major classes of methods for combining probabilities. Further, only slight adjustments of these equations are needed to convert

these methods to other methods found in the literature. For example, a slight alteration of the minimum p method yield's Wilkinson's (1951) generalized method (Birnbaum, 1954; Becker, 1994); the sum of Z method has been generalized to the weighted sum of Z method (Mosteller and Bush, 1954); and the sum of logs method (Fisher, 1932) is a special case of the weighted sum of logs method, the inverse c^2 method, and the weighted inverse c^2 method (Becker, 1994). Therefore, although we've only presented four methods, they represent a large number of the methods for combining probability values that are commonly used in research synthesis.

As mentioned above, methods for combining probability values are often preferred over vote-counting techniques because they make use of sample size information from the individual studies, and are thus more statistically powerful. A disadvantage of these techniques however, is that while they can allow one to accept or reject the hypothesis based on a summary of the primary data, the methods cannot quantify the magnitude of the effect that is yielding significance, nor can they assess the overall agreement (homogeneity) or lack of it among studies or groups of studies. In order to calculate the average effect for a set of studies, one must use modern meta-analytic procedures, which are described below.

Modern Meta-Analysis

Modern meta-analytic techniques are computationally different from combined probability methods in that they combine the measures of effects from individual studies into an estimate of the overall strength of the effect, and use this to determine significance. As in combined probability methods, a significant result implies that there is statistical support for the hypothesis being tested. However, this determination is made by calculating an overall effect size for the combined results from the individual studies, and determining whether this combined effect size is greater than expected by chance. This is the real advantage of meta-analysis, because one can now interpret the strength of the statistical findings based on the magnitude of the overall effect size. Below we present a brief overview of meta-analytic techniques. Theory and computational details can be found in Sections 2.2 - 2.4, and Chapter 3.

A concept crucial to modern meta-analysis is that of the *effect size*. An effect is a statistical measure that portrays the degree to which a given event is present in a sample (Cohen, 1969). For example, one may wish to know the effect of smoking on cancer rates. This can be measured in a number of ways, including the percentage of smokers with cancer, the standardized mean difference of smokers versus non-smokers, or the ratio of the percentage of smokers with cancer versus the percentage of non-smokers with cancer (see methods below). These are all measures of the *effect of smoking on cancer rates*, and their magnitude is considered an *effect size*. Many measures of effect size exist, and which effect size should be used to represent the

results of a set of studies depends, in part, on the nature of the primary data (see below). All effect size measures are a means of representing the results of primary research in a common, standardized way, so that the results from individual studies can be compared and evaluated (Cooper, 1998).

There are two main stages to modern meta-analysis. First, individual effect sizes and their associated variances are calculated for each study in order to place the data from the primary studies on a common scale. The effect size used in this step depends on the type of primary data available and other considerations (see Section 2.2). The second stage of meta-analysis is to combine these effect sizes in a statistical summary based on a particular meta-analytic model. Unlike more traditional statistical approaches (e.g., ANOVA), effect sizes from individual studies in a meta-analysis should be combined using *weighted* statistical models. Several meta-analytic models exist, all of which are weighted linear models whose weights are based on the studies' sampling variances. Various statistical models can be implemented, which represent the variation of study effect sizes around the cumulative effect size as either fixed or random, and various types of data structure can also be included in the model (e.g., no structure or categorical structure). From these weighted models, one obtains estimates of the overall effect present in the data, as well as its variance. This can be used to determine whether there is a significant overall effect. An estimate of heterogeneity among studies can also be calculated; it is used to determine whether the individual studies likely come from one or more statistical populations. It should be noted that alternative approaches (e.g., Bayesian methods) for statistically summarizing effect sizes have been suggested (see Louis and Zelterman, 1994).

One interesting aspect of modern meta-analysis is that, because the methods were developed in several fields simultaneously, there appears to be a wealth of statistical models which are not related in any way to one another (e.g., compare the statistical models presented in Hedges and Olkin, 1985; DerSimonian and Laird, 1986; Berlin et al., 1989). Further, it also seems that there are certain statistical models that can only be used for certain effect sizes. In Section 2.4 we show that this is not the case. Rather, there is one, general model for all meta-analytic methods and models, and simple alterations of this model account for the variation of models among fields.

Summary

In the sections above, we briefly described the three main methods for conducting a comprehensive research synthesis. These methods ranged from computationally simple (vote-counting) to more complex (combined probability values and modern meta-analysis). Because MetaWin is a comprehensive package for performing modern meta-analysis, the next three sections deal exclusively with the theory and calculations necessary to perform such analyses. In Section 2.2 we present methods for calculating

effect sizes and their variances from the three main types of data usually available from primary studies. In Section 2.3 we describe the meta-analytic summary statistics, including calculation of overall effect sizes and heterogeneity statistics, show how data structure can be modeled in a meta-analysis, and discuss fixed effects models and random effects models. Finally, in Section 2.4 we discuss how all meta-analytic models are special cases of a more general statistical model, and show how seemingly disparate methodologies present in various fields are really part of the same, more general framework.

2.2 Effect Size Calculations

The calculation of effect sizes depends, in part, on the data available from the primary studies. Because the data from primary studies are presented in different forms, methods for calculating effect sizes have been developed for these different types of primary data. For most studies, the primary data fall into one of three categories: (1) data representing the means, sample sizes, and standard deviations for both the experimental (treatment) and control groups, (2) data presented as two x two contingency tables representing the four categorical outcomes from a controlled experiment (the number of individuals from the experimental and control groups that do and don't exhibit a given response), and (3) data whose summary statistics can be transformed and represented as a correlation coefficient. For these commonly used methods, it is assumed that the researcher is able to identify which values from the primary studies represent the experimental group, and which represent the control group. Because distinct effect size measures have been developed for each of these three types of data, we present our overview of effect sizes in a similar manner. First we discuss effect sizes that can be calculated from means and standard deviations, followed by effect sizes for two x two contingency tables, and finally, effect sizes from correlations.

Means and Standard Deviation Data

Often, the data found in primary studies represent the summary statistics for both the experimental and control groups. This type of data presentation is quite common in both the social sciences and the biological sciences. For such data, effect size measures have been developed that summarize the difference between experimental and control groups using the means, \bar{X}^C & \bar{X}^E , standard deviations, s^C & s^E , and sample sizes, N^C & N^E . The most common effect size that can be calculated from such data is the standardized mean difference, or Z-score. Several estimates of the standardized mean difference have been developed, and are commonly used in meta-analysis. One of the first effect sizes introduced explicitly for meta-analysis is Glass' (1976) estimate, Δ , which is calculated as:

$$\Delta = \frac{(\bar{X}^E - \bar{X}^C)}{s^C}. \quad (2.5)$$

The variance of Glass' Δ can be estimated as:

$$v_{\Delta} = \frac{N^C + N^E}{N^C N^E} + \frac{\Delta^2}{2(N^C - 1)}. \quad (2.6)$$

This estimate is a simple derivation of the standardized mean difference equation, where the difference in group means is standardized by the control group standard deviation. One difficulty with Glass' Δ is that only the sampling variance from the control group is used to standardize the difference in means. When the sampling variances of the experimental and control groups differ, Hedges (1981) proposed a modification of Glass' (1976) metric which uses an estimate of the pooled sampling variance. Hedges' g can be calculated as:

$$g = \frac{(\bar{X}^E - \bar{X}^C)}{S} \quad (2.7)$$

where the pooled standard deviation is found from:

$$S = \sqrt{\frac{(N^E - 1)(s^E)^2 + (N^C - 1)(s^C)^2}{N^E + N^C - 2}}. \quad (2.8)$$

The estimate of the variance of Hedges' g is similar to the estimate of the variance of Glass' Δ (Rosenthal, 1994), and is found as:

$$v_g = \frac{N^C + N^E}{N^C N^E} + \frac{g^2}{2(N^C + N^E - 2)}. \quad (2.9)$$

Cohen (1969) proposed a similar estimate of the standardized mean difference for a measure of effect size. Cohen's d is calculated as:

$$d_{Cohen} = \frac{(\bar{X}^E - \bar{X}^C)}{s} \quad (2.10)$$

where the pooled standard deviation is found from:

$$s = \sqrt{\frac{(N^E - 1)(s^E)^2 + (N^C - 1)(s^C)^2}{N^E + N^C}}. \quad (2.11)$$

The variance of Cohen's d is calculated as:

$$v_{d-Cohen} = \left(\frac{N^C + N^E}{N^C N^E} + \frac{d^2}{2(N^C + N^E - 2)} \right) \left(\frac{N^C + N^E}{N^C + N^E - 2} \right). \quad (2.12)$$

While Hedges' g is not affected by unequal sampling variances in the experimental and control groups, it does tend to be slightly biased when sample sizes are small (Hedges and Olkin, 1985). To correct for this, one may calculate a slight modification of Hedges' g , which accounts for the effects of small sample sizes. This effect, Hedges' d (which works well for $n > 10$ and even as small as $n = 5$), can be calculated as:

$$d = \frac{\left(\bar{X}^E - \bar{X}^C \right)}{S} J \quad (2.13)$$

where S is the pooled standard deviation (equation 2.8) and J is:

$$J = 1 - \frac{3}{4(N^C + N^E - 2) - 1}. \quad (2.14)$$

The variance of Hedges' d is found as:

$$v_d = \frac{N^C + N^E}{N^C N^E} + \frac{d^2}{2(N^C + N^E)}. \quad (2.15)$$

The four methods described above all estimate the standardized mean difference, and use this estimate as a measure of effect size. An alternative measure of effect size that may be very useful for some applications is the *response ratio* (Hedges et al., 1999). The *response ratio* is the ratio of some measure of outcome in an experimental group to that of the control group. The natural log of this measure has preferable statistical properties, and this effect size is calculated as:

$$\ln R = \ln \left(\frac{\bar{X}^E}{\bar{X}^C} \right) = \ln \left(\bar{X}^E \right) - \ln \left(\bar{X}^C \right) \quad (2.16)$$

with

$$v_{\ln R} = \frac{(s^E)^2}{N^E(\bar{X}^E)^2} + \frac{(s^C)^2}{N^C(\bar{X}^C)^2}. \quad (2.17)$$

Unlike the methods that estimate the standardized mean difference, this method estimates the effect as a proportionate change resulting in experimental manipulation. Such a measure is easily interpretable, and may prove to be as generally useful as the standardized mean difference. All of the methods for estimating effect sizes from means, standard deviations, and sample sizes range from $-\infty$ to $+\infty$, where 0 signifies no difference in effects between the experimental and treatment groups, negative values represent effects where the control group attains a greater value than the experimental group, and positive values represent effects where the experimental group attains a greater value than the control group.

Two x Two Contingency Data

When primary research studies represent data collected from controlled experiments, categorical results are often presented as two x two contingency tables. For such data, there are two groups, experimental and control, and two outcomes, response and no response (e.g., Table 2.1). For example, outcomes may be alive/dead, improved/not improved, no heart attack/heart attack, and we may want to compare the outcomes for members of the experimental and control group: do individuals who smoke have a higher rate of cancer than individuals who don't smoke? Does use of drug X reduce mortality rates? Although this type of primary data is most prevalent in the medical sciences (e.g., Brind et al., 1996), such data may also be found in biology: is mating success (success/failure) associated with the presence of bright plumage (presence/absence)? Thus these effect sizes may have broad applicability to many fields.

Table 2.1. Hypothetical two x two contingency table representing the data from a primary study. In this table, A, B, C, and D are the number of observations in each cell.

		Treatment	Control	Total
Response	A	B	A + B	
	C	D	C + D	
Total	n _t = A + C	n _c = B + D	A + B + C + D	

Before effect sizes can be calculated, one must determine the *rate* of response for both the treatment and control groups. Rate ranges from zero to one, and can be interpreted as the probability of a member of that group showing the response. For the treatment group, the *rate* is calculated as:

$$P_t = \frac{A}{n_t} \quad (2.18)$$

while the *rate* for the control group is:

$$P_c = \frac{B}{n_c}. \quad (2.19)$$

Using these rate values, one of several effect sizes can be calculated. The simplest of these is the *rate difference* (or *risk difference*), which is calculated as the difference in rate scores between the treatment and control groups (Normand, 1999):

$$RD = P_t - P_c. \quad (2.20)$$

An estimate of the sampling variance of *RD* is found as:

$$\nu_{RD} = \frac{P_t(1 - P_t)}{n_t} + \frac{P_c(1 - P_c)}{n_c} \quad (2.21)$$

(DerSimonian and Laird, 1986; L'Abbé et al., 1987; Berlin et al., 1989; Normand, 1999). The *rate difference* ranges from $-\infty$ to $+\infty$, where negative values represent a reduced rate for the treatment group, and positive values represent a greater rate for the treatment group. No difference in the rate between control and treatment groups leads to $RD = 0$. Although the *rate difference* is relatively straightforward to calculate and interpret, a shortcoming of this effect size is that the range of variation of *RD* is greatly limited by the magnitudes of P_t and P_c . That is, the possible values of *RD* are much greater when P_t and P_c are close to 0 and 1, and are constrained when P_t and P_c are both close to 0.5, and this mathematical constraint may cause apparent heterogeneity among studies that is not present in the sample (see Fleiss, 1994).

A second effect size that can be calculated from rate scores is the *relative rate* (*risk ratio*, or *rate ratio*). This effect size is calculated as the rate of the treatment group relative to that of the control group (Greenland, 1987; L'Abbé et al., 1987; Normand, 1999), or:

$$RR = \frac{P_t}{P_c}. \quad (2.22)$$

The variance of the natural log *RR* is found as (Normand, 1999):

$$\nu_{\ln RR} = \frac{(1 - P_t)}{n_t P_t} + \frac{(1 - P_c)}{n_c P_c}. \quad (2.23)$$

Because this effect size is a ratio, no difference in the rate of treatment and control is represented as one. Values of RR from zero to one represent studies where the rate for the treatment group is less than the control group, and values greater than one represent studies where the rate for the treatment group is greater than the control group. As with most ratio metrics, the summary meta-analytic statistics are typically calculated for $\ln RR$, which changes the expected value to zero, and transforms the range to $-\infty$ to $+\infty$.

A slightly more complicated effect size that can be calculated from data in this format is the *odds ratio*. The odds of an event is the probability of the event occurring divided by the probability that the event does not occur to one:

$$odds_t = \frac{P_t}{1 - P_t} : 1. \quad (2.24)$$

The *odds ratio* (or *relative odds*) estimates the odds of an event happening in the treatment group, relative to the odds of the same event happening in the control group (L'Abbé et al., 1987; Berlin et al., 1989; Sokal and Rohlf, 1995; Normand, 1999). OR is calculated as:

$$OR = \frac{\left(\frac{P_t}{1 - P_t} \right)}{\left(\frac{P_c}{1 - P_c} \right)} = \frac{P_t(1 - P_c)}{P_c(1 - P_t)} = \frac{AD}{BC}. \quad (2.25)$$

Many formulas for the variance of the natural log OR exist (see Robbins et al., 1986), but the earliest (and simplest) estimate of the variance (Hauck, 1979) is:

$$\nu_{\ln OR} = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}. \quad (2.26)$$

No difference in the odds of the treatment and control is represented as one. Values of OR from zero to one represent studies where the odds for the treatment group are less than the control group, and values greater than one represent studies where the odds for the treatment group are greater than the control group. Because OR is a ratio, summary statistics are typically performed on $\ln OR$.

An alternative estimate of the *odds ratio*, based on the Mantel-Haenszel (1959) procedure, has been proposed when the effects of multiple studies are to be combined in a meta-analysis (see Mantel and Haenszel, 1959; Yusuf et al., 1985; Berlin et al., 1989). Estimates of the sampling variance of these effects have also been developed, as have

special statistical methods for combining these effects in a summary meta-analysis. A full discussion of these statistics, as well as their derivation, is found in Section 2.4.

Correlation Data

One difficulty in standardizing the results from primary research studies is that the statistical summary information is often insufficient for calculating Hedges' d , $\ln OR$, or other related statistics (Rosenthal, 1991). Further, different studies may present their summary information in different forms: one study may use an F -statistic, while another uses c^2 . In these cases, it is often useful to transform the data to a product moment correlation, and calculate effect sizes from correlation coefficients (Rosenthal, 1991). For example, F -statistics, t -statistics, and c^2 -statistics can all be translated to correlation coefficients, and these can be used to calculate effect sizes (see Table 2.2). Alternatively, the study outcomes may commonly be published as correlations in some fields.

Table 2.2. Conversion of some commonly-found test statistics to r .¹

statistic	conversion equation
Z	$r = \frac{Z}{\sqrt{N}}$
t	$r = \sqrt{\frac{t^2}{t^2 + df}}$
F	$r = \sqrt{\frac{F}{F + df}}$
c^2	$r = \sqrt{\frac{c_{(1)}^2}{N}}$

¹ Note: Z in this table represents the standard normal deviate, not Fisher's z -transform. For conversion of the F -statistic, use 1 df in the numerator, and $df = n-1$ in the denominator (see Rosenthal, 1994). For conversion of the t -statistic, $df=n_c+n_e-2$.

To calculate effect sizes, one first transforms the statistical results from each study to correlation coefficients, using the equations found in Table 2.1 (see also Cohen, 1965; Friedman, 1968). If data are presented only as significance values, these can be converted to standard normal deviates, Z , which can then be transformed to correlation coefficients (Rosenthal, 1991). Using these correlation coefficients, effect size estimates are then calculated from Fisher's (1928) z -transformation (Rosenthal, 1991; Cooper, 1998). Fisher's z -transform is found as:

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) \quad (2.27)$$

and the asymptotic variance of z is:

$$\nu_z = \frac{1}{n-3} \quad (2.28)$$

where n is the sample size (Sokal and Rohlf, 1995). Fisher's z ranges from $-\infty$ to $+\infty$, where negative values of z represent a negative effect, positive values of z represent a positive effect, and $z = 0$ represents no effect. As mentioned above, this method has the advantage that the data from primary research studies may be in almost any form, and need not contain all summary data (e.g., means, standard deviations, and sample sizes). Therefore, one may include a larger number of primary studies in their meta-analysis, since these studies all need not present their results in similar forms. A disadvantage of this method however, is that it is slightly biased when sample sizes are low, but for most practical purposes this bias can be ignored (see Snedecor and Cochran, 1989; Rosenthal, 1994). Another potential disadvantage is that studies reporting outcomes in such a wide range of formats may not be estimating the same thing, and it may not always be reasonable to combine them (but see Rosenthal, 1991).

Other Effect Sizes

In the section above, we have presented many different effect size measures that can be calculated from the data most commonly found in primary research studies. However, it is certainly possible to calculate other measures of effect size, and use these in a summary meta-analysis. Some authors have objected to the use of certain measures of effect size, and thus alternative approaches have been described (e.g., see Greenland et al., 1986; Greenland, 1994; Osenberg et al., 1997; Osenberg and St. Mary, 1998). For typical weighted meta-analyses, this requires knowledge of the sampling distribution of the measure one wishes to use. Although MetaWin calculates six of the effect sizes mentioned above directly from primary data, one may still perform summary meta-analyses using other effect size measures (for details on how to use other effect sizes in MetaWin, see Section 5.2).

Summary

A crucial step in any modern meta-analysis is to record the information from the individual research studies and represent them such that the data are all presented as a common measure and are properly standardized. This is done through the use of effect

sizes which measure the magnitude of effect present in each study. Many effect size measures exist, and the choice of effect size depends, in part, on the type of primary data available. We outlined three major classes of data commonly found in primary research studies and described effect sizes that can be calculated from these types of data (Table 2.3). We presented formulas for calculating these effects, their variances, and the general properties of the effect sizes were also described.

While the choice of effect size is partially based on the type of primary data available, there are other considerations. For example, if all studies contain the necessary information (e.g., means, standard deviations, sample sizes), it may be desirable to use an effect specially-designed for such data (e.g., Hedges' d). On the other hand, the advantage of the correlation methods is that summary test-statistics are almost always available from the primary literature (Rosenthal, 1991), so choosing Fisher's z-transform as the effect size may allow one to include a larger number of primary research studies in the meta-analysis. Fortunately, regardless of which effect size is chosen, the calculation of meta-analytic summary statistics is identical and comparable. We discuss these methods in Section 2.3 below.

Table 2.3. Summary of effect sizes that can be calculated from various sorts of primary data. Equations for sampling variances of these effect sizes are found above.

Name/s	Symbol	Data Type ¹	Equation
Glass' Δ	Δ	\bar{X}, s, N	$\Delta = \frac{(\bar{X}^E - \bar{X}^C)}{s^C}$
Hedges' g	g	\bar{X}, s, N	$g = \frac{(\bar{X}^E - \bar{X}^C)}{S}$
Cohen's d	d_{Cohen}	\bar{X}, s, N	$d_{Cohen} = \frac{(\bar{X}^E - \bar{X}^C)}{s}$
Hedges' d	d	\bar{X}, s, N	$d = \frac{(\bar{X}^E - \bar{X}^C)}{S} J$
<i>response ratio</i>	$\ln R$	\bar{X}, s, N	$\ln R = \ln \left(\frac{\bar{X}^E}{\bar{X}^C} \right)$
<i>rate difference</i> <i>risk difference</i>	RD	P_t, P_c	$RD = P_t - P_c$
<i>relative rate</i> <i>risk ratio</i> <i>rate ratio</i>	$\ln RR$	P_t, P_c	$RR = \frac{P_t}{P_c}$
<i>odds ratio</i> <i>relative odds</i>	$\ln OR$	P_t, P_c	$OR = \frac{P_t(1 - P_c)}{P_t(1 - P_t)}$
Fisher's z -transform	z	r, N	$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$

¹ Note: P_t and P_c are calculated from two x two contingency tables; r can be calculated from any test-statistic.

2.3 Meta-Analysis Summary Statistics

One of main the goals of a quantitative research synthesis is to determine whether there is evidence in a set of primary studies supporting a particular hypothesis. For meta-analysis, the first step in this process is to place the results of the primary studies in a common framework, which is done through the calculation of effect sizes. The second step is to statistically summarize these effect sizes to determine whether there is evidence for the hypothesis. This is typically done through a set of summary analyses

which are based on a particular statistical model. In this section, we describe the commonly-used summary analyses of a typical meta-analysis. First we discuss cumulative effect sizes, and how they can be used to assess the overall evidence supporting a particular hypothesis. We then describe statistics for assessing the degree of heterogeneity between studies, followed by a description of methods for incorporating data structure into meta-analytic models. This includes a discussion of cumulative effect sizes for groups in a categorical analysis, the relation of effect sizes to other independent variables, and how the total heterogeneity in a sample can be partitioned into various statistical components. Finally, we end this section with a discussion of the differences between fixed-effects meta-analytic models and random-effects meta-analytic models.

Cumulative Effect Size and Total Heterogeneity

The most common statistic for summarizing a sample of data points is the average value, or the *mean*. The mean is a measure of central tendency and estimates the true average effect for a population. One difficulty with using the mean to summarize effect sizes in meta-analysis, however, is that individual studies frequently do not have common sample sizes, and thus some form of weighting is necessary. Further, because the variance of effect sizes are a function of sample size, large-sample theory states that studies with larger sample sizes will have lower variances, and thus will provide more precise estimates of the true population effect size (see Hedges and Olkin, 1985; Shadish and Haddock, 1994; Cooper, 1998). To take this into consideration, a weighted average is used in meta-analysis to estimate the *cumulative effect size* for a sample of studies, where the weight for the i^{th} study is the reciprocal of its sampling variance, $w_i = 1/v_i$. The cumulative effect size is calculated as:

$$\bar{\bar{E}} = \frac{\sum_{i=1}^n w_i E_i}{\sum_{i=1}^n w_i} \quad (2.29)$$

where n is the number of studies and E_i is the effect size for the i^{th} study. The variance of $\bar{\bar{E}}$ is a function of the individual weights, and is found as:

$$s_{\bar{\bar{E}}}^2 = \frac{1}{\sum_{i=1}^n w_i}. \quad (2.30)$$

Using $s_{\bar{\bar{E}}}^2$, the confidence interval around $\bar{\bar{E}}$ is:

$$CI = \bar{\bar{E}} \pm t_{\alpha/2[n-1]} * \frac{s}{E} \quad (2.31)$$

where t is the two-tailed critical value found from the Student's t -distribution at the critical level, α . Using this confidence interval, the overall effect present in a set of studies can now be evaluated. The cumulative effect size represents the overall magnitude of the effect present in the studies; this value is considered to be significantly different from zero if its confidence limits do not bracket zero (i.e., $\bar{\bar{E}}$ is significant at $P = \alpha$). Thus, with the cumulative effect size and its confidence interval, we can determine whether there is significant evidence supporting a particular hypothesis, and also estimate the *magnitude* of that support.

In addition to the cumulative effect size, it is also of interest to determine whether a set of effect sizes are homogeneous. The *total heterogeneity* of a sample, Q_T , (Hedges and Olkin, 1985) is calculated as:

$$Q_T = \sum_{i=1}^n w_i E_i^2 - \frac{\left(\sum_{i=1}^n w_i E_i \right)^2}{\sum_{i=1}^n w_i} \quad (2.32)$$

where n , E_i , and w_i are the same quantities as used above. Through simple algebra, the equation for Q_T can be rearranged and expressed as:

$$Q_T = \sum_{i=1}^n w_i \left(E_i - \bar{\bar{E}} \right)^2. \quad (2.33)$$

To assess whether there is significant heterogeneity in a sample, Q_T is tested against a χ^2 -distribution with $n - 1$ degrees of freedom. The null hypothesis for this test is that all effect sizes are equal (see Gurevitch and Hedges, 1993). A significant Q_T indicates that the variance among effect sizes is greater than expected by sampling error (Cooper, 1998) and implies that other explanatory variables should be investigated (see below). As is obvious from equation 2.33, Q_T is a weighted sum of squares, and is thus comparable to the total sum of squares in an analysis of variance (ANOVA).

Incorporating Data Structure in Meta-Analysis

As mentioned above, a significant Q_T indicates that the variance among effect sizes is greater than expected by sampling error. One possible explanation for this is that there may be some underlying structure to the data. Thus far, we have presented the summary statistics of meta-analysis under the assumption that all studies came from

the same population. This is the simplest model of data structure, and actually models the distribution of effect sizes as if there was *no structure* in the data. However, this may not always be the case. Examples of meta-analyses with data structure include Rosenthal and Rubin's (1978) study of interpersonal expectancy effects for dissertation and nondissertation studies and Gurevitch et al.'s (1992) study of the effects of competition in ecological field experiments, in which trophic levels were compared. For such *categorical* factors, one would like to address not only if the overall cumulative effect size is significant, but also whether there are differences among the effect sizes for particular groups of studies (in the meta-analytic literature, these grouping variables are often referred to as *predictors* or *moderators*).

A second model that may be used describes variation in effect sizes by their covariation with another independent variable. In these *continuous* models, one describes the relationship between effect sizes and the independent variable using a linear model, and determines whether this model explains a significant portion of the variation in effect sizes across studies. Such models have been used to investigate the effects of television on school learning (Williams et al., 1982), as well as the effect of productivity on competition in plant communities (Goldberg et al., 1999). Though both categorical and continuous models are generally considered to be subsets of *general linear models* (GLM), we first describe the computational details of each separately, and then discuss generalizations of these models.

Categorical Data: When studies can be segregated into more than one group, a *categorical meta-analysis* is appropriate. For such data, one can calculate the overall cumulative effect size, \bar{E} , and its confidence interval as above (equations 2.29 – 2.31). In addition, the cumulative effect size for each group, \bar{E}_j , can be calculated using only the studies for that group. \bar{E}_j is calculated as:

$$\bar{E}_j = \frac{\sum_{i=1}^{k_j} w_{ij} E_{ij}}{\sum_{i=1}^{k_j} w_{ij}} \quad (2.34)$$

where k_j is the number of studies in the j^{th} group, and w_{ij} and E_{ij} are the weight and effect size for the i^{th} study in the j^{th} group. The variance of \bar{E}_j is found as:

$$s_{\bar{E}_j}^2 = \frac{1}{\sum_{i=1}^{k_j} w_{ij}} \quad (2.35)$$

and the confidence interval around \bar{E}_j is:

$$CI = \bar{E}_j \pm t_{\alpha/2[k_j-1]} * s_{\bar{E}_j}. \quad (2.36)$$

For each group, \bar{E}_j represents an estimate of the true cumulative effect size for that set of studies, and like $\bar{\bar{E}}$, the cumulative effect size for each group is considered to be significant if its confidence interval does not bracket zero (i.e., \bar{E}_j is significant at $P = \alpha$).

The heterogeneity within the j^{th} group, Q_{Wj} , is found as:

$$Q_{Wj} = \sum_{i=1}^{k_j} w_{ij} (E_{ij} - \bar{E}_j)^2 \quad (2.37)$$

where k_j is the number of studies in the j^{th} group, w_{ij} and E_{ij} are the weight and effect size for the i^{th} study in the j^{th} group, and \bar{E}_j is the cumulative effect size for that group. Q_{Wj} may be tested against a χ^2 -distribution with $k_j - 1$ degrees of freedom.

In addition to testing if the cumulative effect sizes for each group are significantly different from zero, one would also like to know whether the groups differ from each other. This is analogous to the procedure of analysis of variance (ANOVA), where one is interested in determining whether or not there is a difference among group means. In ANOVA, this is determined by partitioning the total variation in a sample into several components: the variance explained by the model (groups), and the residual error variance. Thus,

$$SS_T = SS_M + SS_E \quad (2.38)$$

and the difference among groups is statistically tested using an F -ratio of the model variance versus the error variance, where each is divided by its degrees of freedom (see Sokal and Rohlf, 1995). In meta-analysis, we can partition the total heterogeneity, Q_T , in a similar manner (Hedges and Olkin, 1985) such that:

$$Q_T = Q_M + Q_E \quad (2.39)$$

where Q_M is the variation in effect sizes that is explained by the model, and Q_E is the residual error variance in effect sizes not explained by the model. For categorical data, Q_M is thus a description of the difference among group cumulative effect sizes, and is calculated as:

$$Q_M = \sum_{j=1}^m \sum_{i=1}^{k_j} w_{ij} \left(\bar{E}_j - \bar{\bar{E}} \right)^2 \quad (2.40)$$

where m is the number of groups, k_j is the number of studies in the j^{th} group, w_{ij} is the weight for the i^{th} study in the j^{th} group, \bar{E}_j is the cumulative effect size for the j^{th} group, and $\bar{\bar{E}}$ is the overall cumulative effect size. The residual error heterogeneity, Q_E , is found through the summation of the individual within-group heterogeneity values, Q_{Wj} , or through the calculation of:

$$Q_E = \sum_{j=1}^m Q_{Wj} = \sum_{j=1}^m \sum_{i=1}^{k_j} w_{ij} (E_{ij} - \bar{E}_j)^2 \quad (2.41)$$

where m is the number of groups, k_j is the number of studies in the j^{th} group, w_{ij} and E_{ij} are the weight and effect size for the i^{th} study in the j^{th} group, and \bar{E}_j is the cumulative effect size for the j^{th} group. Both Q_M and Q_E can be tested against a χ^2 -distribution with $m - 1$ degrees of freedom for Q_M and $n - m$ degrees of freedom for Q_E . A significant Q_M implies that there are differences among cumulative effect sizes for the groups, while a significant Q_E implies that there is still heterogeneity among effect sizes not explained by the model. An important distinction between the notation used above and that used in most meta-analysis texts is that we have chosen the terms Q_M and Q_E to represent the quantities usually referred to as Q_B and Q_W (see e.g., Hedges and Olkin, 1985; Cooper, 1998). We have done this because partitioning the total heterogeneity can be performed for models of data structure *other* than categorical data models. Therefore, ‘model’ heterogeneity and ‘error’ heterogeneity are more general terms.

Continuous Data: When the effect sizes from individual studies can be explained by another independent variable, a *continuous model meta-analysis* is appropriate (Hedges and Olkin, 1985; Greenland, 1987). Continuous model meta-analysis requires three variables from each study, the effect size, E_i , the weight, w_i , and the value of the independent variable, X_i . From these variables, the relationship between effect size and the independent variable is determined using weighted least squares regression (for a description of weighted least squares, see Neter et al., 1989). The general equation for least squares regression is found as:

$$E_i = b_0 + b_1 X_i + \epsilon \quad (2.42)$$

where E_i and X_i are the effect size and independent variable of the i^{th} specimen, b_0 and b_1 are the intercept and slope of the regression, and ϵ is the error (Sokal and Rohlf, 1995). Using a weighted least squares regression model, the relationship between effect sizes and the independent variable, X , is represented by the slope, b_1 , which is found as:

$$b_1 = \frac{\sum_{i=1}^n w_i X_i \sum_{i=1}^n w_i E_i - \sum_{i=1}^n w_i X_i \sum_{i=1}^n w_i E_i}{\sum_{i=1}^n w_i \left(\sum_{i=1}^n w_i X_i \right)^2} \quad (2.43)$$

where n is the number of studies, and w_i , E_i , and X_i are the weight, effect size, and value for the independent variable of the i^{th} study. The intercept from this regression is found as:

$$b_0 = \frac{\sum_{i=1}^n w_i E_i - b_1 \sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i} \quad (2.44)$$

Unfortunately, the standard errors for the slope and intercept cannot be calculated through standard least squares procedures (see Hedges and Olkin, 1985). Rather, the standard error of the slope is calculated as:

$$s_{b_1} = \sqrt{\frac{1}{\sum_{i=1}^n w_i X_i^2 - \left(\sum_{i=1}^n w_i X_i \right)^2}} \quad (2.45)$$

and the standard error of the intercept is calculated as:

$$s_{b_0} = \sqrt{\frac{1}{\sum_{i=1}^n w_i - \left(\sum_{i=1}^n w_i X_i \right)^2}} \quad (2.46)$$

Dividing the slope and intercept by the corresponding standard error yields their Z -score, which is compared to a normal distribution to determine whether they are statistically significant. A significant regression coefficient (i.e., slope) implies that the independent variable explains a significant portion of the variation in effect sizes. Like the categorical data model, one can partition the total heterogeneity into several components, Q_M , which describes the amount of heterogeneity explained by the

regression model, and Q_E , the amount of residual error heterogeneity. For a continuous model, Q_M is calculated as:

$$Q_M = \frac{b_1^2}{s_{b_1}^2} \quad (2.47)$$

and Q_E can be found as:

$$Q_E = Q_T - Q_M \quad (2.48)$$

where Q_T is found in the usual manner. The significance level of Q_M (which is tested against a χ^2 -distribution with 1 degree of freedom) is exactly identical to the significance level of b_1 , which is tested against a normal distribution. Again, a significant Q_M indicates that the independent variable explains a significant amount of the variability in effect sizes, and a significant Q_E (tested against a χ^2 -distribution with $n - 2$ degrees of freedom) implies that there is still heterogeneity among effect sizes not explained by the model.

Fixed-Effects Models and Random-Effects Models

To this point, we have described three different statistical models for performing meta-analysis, which are distinguished from one another based on data structure. The simplest model is one where all studies come from the same population, which we have called the *no structure model*. Using this model, one may calculate the overall cumulative effect size, and the total heterogeneity. Alternatively, if the studies can be separated into distinct groups, one may use a *categorical model*, where, in addition to the overall cumulative effect size and the total heterogeneity, one may calculate and test the cumulative effect sizes for each group. Like ANOVA, the total heterogeneity can be partitioned into several components, one that represents the variance in effect sizes explained by the categorical variable, and one that represents the residual error variance. Finally, if one has an independent variable, a weighted least squares regression analysis can be performed, where the slope and intercept of the *continuous model* can be calculated, as well as the heterogeneity in effect sizes explained by the regression model.

Although all of these models can be written in a more general format as special cases of the *General Linear Model* (see Section 2.4), there is an additional commonality to these methods, namely, that all are *fixed-effects* meta-analysis models. Fixed effects models are those models that assume that there is one true effect size shared by all studies, or one true effect size for each group of studies (see Hedges 1994; Hedges and Vevea, 1998; Gurevitch and Hedges, 1999). Fixed-effects models calculate summary statistics based

on the assumption that the only variation in effect sizes is due to sampling error. A second class of models accounts for the fact that, in addition to sampling error, there is true random component of variation in effect sizes between studies. Such models are called *random-effects* models (see Raudenbush, 1994). In random-effects models the random component of effect size variation is calculated and incorporated into the summary statistics.

Performing a random-effects model is a three-step process. First, one performs a fixed-effects model analysis to determine the values of the summary statistics. Using these summary statistics, one then calculates an estimate of the *pooled study variance* (or between-study variance), s_{pooled}^2 . This is used to calculate the weights for the random-effects model. In a fixed-effects model, one calculates the individual weights as: $w_i = 1/v_i$ where v_i is the individual study variance. In a random effects model, s_{pooled}^2 is incorporated into this weight formulation such that:

$$w_{i(rand)} = \frac{1}{v_i + s_{pooled}^2}. \quad (2.49)$$

Finally, these new weights are used in the random-effects model, which is calculated using the same equations as described above (equations 2.29 – 2.48).

The calculation of s_{pooled}^2 depends on the underlying structure of the data. For a model with no structure, s_{pooled}^2 is calculated as:

$$s_{pooled}^2 = \frac{Q_T - (n - 1)}{\sum_{i=1}^n w_i^2} \cdot \frac{\sum_{i=1}^n w_i}{\sum_{i=1}^n w_i} \quad (2.50)$$

where Q_T is the total heterogeneity for the sample of studies, n is the number of studies, and w_i is the fixed-effects weight for the i^{th} study. The numerator of this estimate is thus the unexplained variance in effect sizes minus the degrees of freedom, and the denominator is a summed function of the weights. For a categorical model, a slightly different formulation of s_{pooled}^2 is used, which is:

$$\mathbf{s}_{pooled}^2 = \frac{Q_E - (n - m)}{\sum_{j=1}^m \left(\sum_{i=1}^{k_j} w_{ij} - \frac{\sum_{i=1}^{k_j} w_{ij}^2}{\sum_{i=1}^{k_j} w_{ij}} \right)}. \quad (2.51)$$

In this formulation, Q_E is the residual error heterogeneity from the fixed-effects model, n is the total number of studies, m is the number of groups, k_j is the number of studies in the j^{th} group, and w_{ij} is the fixed-effects weight for the i^{th} study in the j^{th} group. This equation is very similar to equation 2.49, in that the numerator is the unexplained variance in effect sizes minus the degrees of freedom, and the denominator is a function of the weights, summed over groups. Note that this random-effects model for categorical data is sometimes referred to as the *mixed-effects model* (e.g., Gurevitch and Hedges, 1993; Hedges and Vevea, 1998), analogous to mixed-effects models in ANOVA, because it includes random variation among studies within a group and fixed differences between groups.

The equation of \mathbf{s}_{pooled}^2 for a continuous model is slightly more complicated than the two equations above, but the same general format applies. For a continuous model, \mathbf{s}_{pooled}^2 is calculated as:

$$\mathbf{s}_{pooled}^2 = \frac{Q_E - (n - 2)}{\sum_{i=1}^n w_i - \sum_{i=1}^n w_i^2 \left(\frac{\sum_{i=1}^n w_i X_i^2 - 2X_i \sum_{i=1}^n w_i X_i + X_i^2 \sum_{i=1}^n w_i}{\sum_{i=1}^n w_i \sum_{i=1}^n w_i X_i^2 - \left(\sum_{i=1}^n w_i X_i \right)^2} \right)} \quad (2.52)$$

where Q_E is the residual error heterogeneity from the fixed-effects continuous model, n is the number of studies, and w_i and X_i are the fixed-effects weight and independent variable for the i^{th} study.

While equations 2.50 and 2.51 have been published elsewhere (see e.g., Hedges and Olkin, 1985), the between-study variance component for the continuous random model (equation 2.52) has not previously been represented exactly in this format. We derived this equation from a more general formula (equation 2.62) for estimating between-study variance components for random-effects models (Overton, 1998; Hedges and Olkin, 2000), of which the three equations above are special cases. This formula is part of a general format for all meta-analytic summary analyses, which we present in the next section.

2.4 A General Model for Meta-Analysis

In the previous section, we presented six different statistical models for meta-analysis. We first presented three different types of data structure (no structure, categorical data structure, continuous data structure), and showed how each of these models for data structure could be incorporated into either a fixed-effects meta-analysis or a random-effects meta-analysis. While the variance of effect sizes is typically modeled in one of two ways (i.e. fixed and random), there are many possible models for data structure. For example, one may have two categorical variables of interest, or one categorical *and* one continuous variable. These statistical models can be calculated using weighted *general linear models*, or GLM (Hedges and Olkin, 1985; 2000). Although MetaWin does not currently include a GLM module, all of the statistical models calculated in MetaWin are special cases of GLM, so a general understanding of this procedure helps to place the equations of the various meta-analytic models in a common framework. We therefore present a brief introduction to GLM below.

General Linear Models (GLM) for Meta-Analysis

General linear models use matrix algebra to determine the relationship between a set of dependent and independent variables. We first define \mathbf{E} as the vector of dependent variables, which, for meta-analysis, are effect sizes. Thus,

$$\mathbf{E} = \begin{bmatrix} E_1 \\ \vdots \\ E_n \end{bmatrix} \quad (2.53)$$

is a $1 \times n$ vector of effect sizes. The *design matrix* for a GLM is the matrix of independent variables. This matrix, \mathbf{X} , is defined as:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p1} \\ \vdots & \vdots & & \\ 1 & X_{1n} & \cdots & X_{pn} \end{bmatrix}. \quad (2.54)$$

The first column of \mathbf{X} is always a vector of ones, and is the only column included for a meta-analysis with no data structure. Additional columns of values are added to the \mathbf{X} matrix if there is structure in the data. This includes columns representing a continuous variable, or columns representing a categorical grouping variable. Incorporation of independent variables is a general procedure, so if multiple independent variables exist, they can be incorporated in the \mathbf{X} matrix simultaneously as multiple vector columns.

Interactions between variables can also be represented by the inclusion of additional columns in the \mathbf{X} matrix (see Johnson and Wichern, 1988). Using GLM, the relationship between the independent variables and effect sizes is found through the equation:

$$\mathbf{E} = \mathbf{X}\mathbf{b} + \mathbf{e} \quad (2.55)$$

where \mathbf{E} and \mathbf{X} are defined as above, \mathbf{b} is a vector of model coefficients, and \mathbf{e} represents the model error. Using this equation, one can find estimates of the model coefficients, \mathbf{b} , as well as estimates of the variance explained by the model. However, in a meta-analysis one must take into consideration the fact that the effect sizes vary in their precision around the true cumulative effect size, so a weighted estimate of \mathbf{b} is found as:

$$\mathbf{b} = (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{X}' \mathbf{W} \mathbf{E} \quad (2.56)$$

where \mathbf{X} , \mathbf{E} , and \mathbf{b} are defined as above, and \mathbf{W} is a diagonal matrix with the individual study weights, w_i , on the diagonal (Hedges and Olkin, 1985). \mathbf{b} yields estimates of the coefficients for the GLM model:

$$\mathbf{E}_i = \mathbf{b}_o + \mathbf{b}_1 X_{1i} + \mathbf{b}_2 X_{2i} + \dots + \mathbf{b}_p X_{pi} + \mathbf{e}_i. \quad (2.57)$$

Estimates of the variance for the \mathbf{b} coefficients are found from the diagonal of the matrix:

$$\sum_b = (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1}. \quad (2.58)$$

Note that equations 2.45 and 2.46 are derived from the diagonal elements of Σ_β . Using the GLM model, one can also find estimates of heterogeneity among effect sizes. The heterogeneity explained by the model, Q_M , is found as:

$$Q_M = \mathbf{b}^t \sum_{b^*}^{-1} \mathbf{b}^* \quad (2.59)$$

where \mathbf{b}^* and $\sum_{b^*}^{-1}$ are equivalent to \mathbf{b} and Σ_β , except that the first entry of \mathbf{b} , and the first row and column of Σ_β have been eliminated. The residual error heterogeneity of the model, Q_E , can be computed as:

$$Q_E = (\mathbf{E} - \mathbf{X}\mathbf{b})^t \mathbf{W} (\mathbf{E} - \mathbf{X}\mathbf{b}). \quad (2.60)$$

The total heterogeneity, Q_T , may be found as the summation of Q_M and Q_E , or:

$$Q_T = Q_M + Q_E. \quad (2.61)$$

Finally, between-study variance components for random-effects models can be found using the generalized equation:

$$\mathbf{s}_{pooled}^2 = \frac{Q_E - (n - k)}{tr\mathbf{W} - tr[\mathbf{WX}(\mathbf{X}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{W}]} \quad (2.62)$$

where Q_E is the residual error heterogeneity for the model, k is the number of columns of \mathbf{X} ($n - k$ is the degrees of freedom of Q_E), and \mathbf{W} and \mathbf{X} are as defined above (Note: “ tr ” denotes the trace of a matrix). Because this equation is general, the between-study variance components for any random-effects model can be found (see Hedges and Olkin, 2000).

Presenting meta-analysis in the GLM notation, it is now possible to represent all meta-analytic models in a similar manner. Using weighted GLM, one can perform both fixed-effects and random-effects models, and any type of data structure can be incorporated, including simple designs with no date structure, as well as complicated designs where both categorical and continuous independent variables exist. Thus, the GLM notation is completely general, and thinking of the different ‘models’ of meta-analysis as special cases of weighted GLM allows one to see the common attributes of these seemingly different sets of computations.

Generalizing Odds Ratio Computations

From the section above, it is clear that all statistical models for meta-analysis are special cases of weighted GLM. However, in the medical literature, one often encounters a set of special summary statistics that do not seem to conform to this common meta-analytic framework. On the surface, these equations seem dissimilar to those found in Section 2.3, but they are, in fact, completely compatible. Recall that in the medical sciences, the most prevalent form of primary data are categorical results presented as two x two contingency tables (e.g., Table 2.4). For such data, one typically calculates the *odds ratio* for each study, and combines these to estimate cumulative effects (Mann, 1990; 1994). However, rather than directly using the procedures outlined above, the Mantel-Haenszel (1959) procedure is commonly used for combining *odds ratios* (Yusuf et al., 1985; L’Abbe et al., 1987; Berlin et al., 1989; Haddock et al., 1998).

Table 2.4. Hypothetical two x two contingency table representing the data from a primary study. In this table, A, B, C, and D are the number of observations in each cell.

		Treatment	Control	Total
Response	A	B	A + B	
	C	D	C + D	
Total	$n_t = A + C$	$n_c = B + D$	$N = A + B + C + D$	

The Mantel-Haenszel procedure for combining *odds ratios* is as follows. First, for each of m independent studies, the total number of events in the i^{th} study is defined as N_i , and the observed responses from the treatment group $O_i = A_i$. Thus, the expected number of responses, assuming no treatment effect is:

$$\hat{O}_i = \frac{(A_i + B_i)(A_i + C_i)}{N_i}. \quad (2.63)$$

Assuming no treatment effect, $O_i - \hat{O}_i$ will vary randomly around zero, with variance:

$$V_i = \hat{O}_i \left(\frac{A_i + C_i}{N_i} \right) \left(\frac{C_i + D_i}{N_i - 1} \right). \quad (2.64)$$

Using the expected response and its variance for each study, the common log odds ratio (i.e. the mean *odds ratio* or cumulative effect size) from the m studies is estimated as:

$$\ln \overline{\overline{OR}} = \frac{\sum O_i - \hat{O}_i}{\sum V_i} \quad (2.65)$$

and the total heterogeneity, Q_T , is calculated as:

$$Q_T = \sum \frac{(O_i - \hat{O}_i)^2}{V_i} - \frac{(\sum O_i - \hat{O}_i)^2}{\sum V_i}. \quad (2.66)$$

On the surface, these calculations seem very different than those presented in Section 2.3. However, a slight alteration of the equations above allows us to estimate the natural log *odds ratio* for each study as:

$$E_i = \ln OR_i = \frac{O_i - \hat{O}_i}{V_i} \quad (2.67)$$

with variance $v_i = 1/V_i$. Using this estimate of the effect size and its variance, we can thus calculate the cumulative effect size and the total heterogeneity using the general framework outlined in Section 2.3. It can be shown that the results using our estimate of the effect size and its variance are identical to those found from the Mantel-Haenszel procedure. For example, using our estimates of $\ln OR$ in equation 2.29 yields the same value for the cumulative effect size as calculated using from the Mantel-Haenszel procedure (equation 2.65). In fact, the equivalence of equations estimating cumulative effect sizes (equations 2.29 and 2.65), and those estimating the total heterogeneity (equations 2.32 and 2.66), can be determined through simple algebraic manipulation.

The relationship between the two estimates of $\ln OR$ (equations 2.25 and 2.67) is somewhat more complicated. Equation 2.67 is not equal to the natural log of equation 2.25. Equation 2.25 is the calculation of the odds ratio for an individual study from a single two x two contingency table. Equation 2.67 is the estimate of the *common* $\ln odds ratio$ based on a single study. Although one could use equations 2.25 and 2.26 to estimate the odds ratio for use in a meta-analysis, it has been shown that the Mantel-Haenszel (1959) estimator gives better results, particularly for data with small sample sizes (see Hauck, 1989). However, because statistical results using our estimate of the common $\ln OR$ are identical to those found from the Mantel-Haenszel procedure, thinking of meta-analysis from our general model framework is valid.

2.5 Summary

In this chapter, we described methods for combining results from independent studies in a quantitative research synthesis. We presented a brief history of meta-analysis and compared the methods of vote counting, combining probability values, and modern meta-analysis. We described methods for calculating effect sizes from the three most common sources of primary data: means, standard deviations, and sample sizes; two x two contingency tables; and correlation coefficients, and showed how various test-statistics can be converted to a common framework. We then discussed methods for combining these effect sizes using meta-analytic summary analyses. We presented an introduction to fixed-effects and random-effects models, and showed how data structure can be incorporated in the calculation of cumulative effect sizes and effect size heterogeneity. Finally, we described a general model for all meta-analyses (GLM), and showed how the summary statistics for *odds ratios* can fit into this general framework.

In the next chapter, we describe additional calculations and analyses that may be performed on meta-analytic data. First we discuss various methods of data exploration that can be used to determine whether biases exist in the data. Next, we discuss cumulative meta-analysis, and show its uses. Finally, we present a brief introduction to resampling tests and describe why one may wish to use these in meta-analysis for the purposes of statistical evaluation.

3. ADDITIONAL ISSUES, ANALYSES AND CALCULATIONS

In the previous chapter, we presented methods for quantitative research synthesis, and discussed the procedures and steps necessary to conduct a modern meta-analysis. These steps include transforming the data from primary research studies into a common *effect size*, and combining these effect sizes according to some meta-analytic model. In this chapter, we discuss additional procedures and calculations that have been developed to complement meta-analysis. The techniques discussed here include methods for detecting publication bias, fail-safe measures and other methods of data exploration, cumulative meta-analysis, and resampling procedures that can be incorporated into a modern meta-analysis.

3.1 Publication Bias and Data Exploration

A common concern when conducting a literature review, especially quantitative reviews such as meta-analysis, is the potential for publication bias. Publication bias is the selective publication of articles showing certain types of results over those showing other types of results (Begg, 1994). The most commonly suspected publication bias is the tendency for journals to publish only those studies with statistically significant results. This deficit of non-significant published studies has been termed the “file-drawer problem” (Rosenthal, 1979) and will lead to an overestimate of the number of significant results on a given topic, sometimes called “anticonservative inferences” (Begg and Berlin, 1988). Such publication bias can have a number of causes; the most obvious are self-selective publication by researchers who do not try to publish studies whose results are statistically non-significant, and an editorial policy against publication of statistically non-significant studies (Hedges, 1992). Publication bias can also be caused by selective publication based on other factors such as sample size or effect size. A thorough treatment of publication bias can be found in Begg and Berlin (1988).

Does Publication Bias Exist?

A number of studies have actively looked for evidence of publication bias. In an early study, Sterling (1959) reported indirect evidence for publication bias, finding that 286 out of 294 articles (97%) in four psychological journals reported statistically significant results. Smart (1964) found that only 9% of published papers showed non-significant results, while 30% of unpublished theses showed non-significant results. Smith (1980) showed that effect sizes were smaller in unpublished studies than in published studies. Similarly, Coursol and Wagner (1986) showed that decisions to submit or publish research were correlated with positive results. Easterbrook et al. (1991) and Dickersin et al. (1992) both showed that statistical significance increased the chance of publication in

medical literature; Csada et al. (1996) showed a similar result in biology (although see Bauchau, 1997). Kleijnen and Knipschild (1992) have reviewed the methodologies used in studies of publication bias.

Of course, it is possible that the lack of non-significant studies signifies just that, non-significance. Nevertheless, most studies tend to produce significant results. This could be due to the lack of randomness in researchers' null hypotheses: researchers do not choose null hypotheses at random, but rather choose to study a phenomenon which they already believe exists (see Bauchau, 1997). Such selection of hypotheses will increase the proportion of significant results found in the literature. The truth is probably somewhere in between the extremes of publication bias on one end and pure corroborative research on the other.

An alternative approach to studying publication bias is modeling. A number of attempts have been made to model publication bias and its potential impact on meta-analytic summaries (e.g., Lane and Dunlap, 1978; Begg, 1985; Hedges, 1984, 1992; Iyengar and Greenhouse, 1988; Begg and Berlin, 1988; Dear and Begg, 1992; Vevea and Hedges, 1995; Hedges and Vevea, 1996; Bradley and Gupta, 1997). Many of these approaches are based on the concept of the *Weighted Distribution Theory* (Begg, 1994), where studies are included in an analysis based on their outcome. The approach is to model the probability of a study's inclusion based on its observed significance. Specification of this inclusion function gives an estimate of the degree to which publication bias may exist in a given data set; this bias may then be accounted for by complex factoring of the modeled missing publications. This type of research is still in its infancy, but offers an exciting potential solution to the problem (see Begg, 1994, for a review of these models).

In recent years, there have been suggestions that publication bias may not be as serious a problem as in the past. In medical research, there has been a large push to register research trials before data collection (Begg and Berlin, 1988), which could lead to a more accurate assessment of the true number of studies (Kleijnen and Knipschild, 1992). With the advent of the internet, the availability of unpublished data is becoming more and more common. Groups such as the Cochrane Collaboration have set up formats and standards that allow for the full exchange and storage of data. It has been suggested that funding agencies should take an active role in publishing all of the studies which they fund (Begg and Berlin, 1988). These and other types of policy adjustments have helped to eliminate some sources of publication bias because meta-analysts will not be dependent solely on journal publications for their data.

In the remainder of this section we shall discuss in more detail a number of simple methods that can be used to identify and examine the impact of publication bias.

Graphical Data Exploration

A number of graphical methods have been suggested to explore meta-analytic data. They can be used as methods for the potential identification of publication bias, as well as for a greater understanding of the data in general.

A simple way to examine the basic distribution of data is through a *Weighted Histogram* (Greenland, 1987). A weighted histogram (Figure 3.1) differs from a normal histogram (Sokal and Rohlf, 1995) in that the height of the bar in each class is made up of the combined weight of the studies which fall within that class, rather than their frequency. Because meta-analytic methods assume normality, a histogram may help one identify outliers. The shape of the weighted histogram can tell one about the structure and nature of the data. For example, if the histogram is bimodal or multimodal, it may indicate multiple groups necessitating a categorical model structure. Furthermore, if there is a publication bias against reporting studies with little or no effect, the histogram may be depressed around the region of no effect. Because of this type of publication bias, collections of studies with zero cumulative effect may be bimodal with a cluster of positive and negative effects on either side of zero; collections of studies with a real effect may be skewed in the direction of the effect (Greenland, 1987).

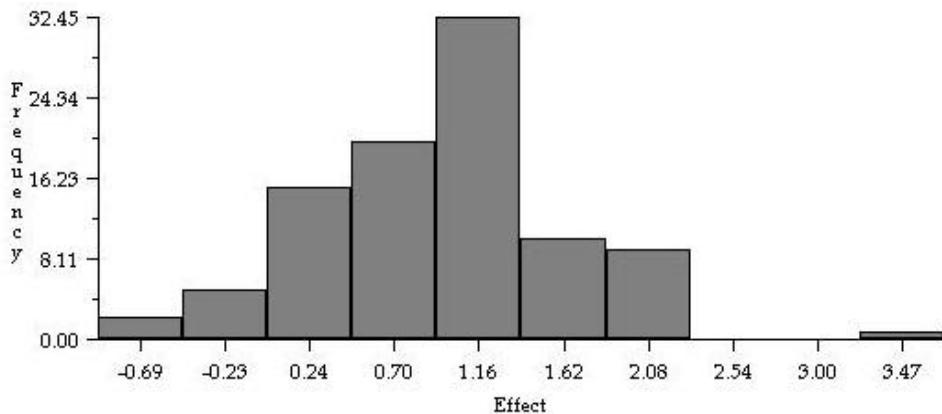


Figure 3.1. Example of a weighted histogram.

The *Funnel Plot* is a useful, informal method for exploring the distribution of one's data and searching for publication bias. A funnel plot is a scatter plot of effect size vs. sample size (Light and Pillemer, 1984). If there are no biases in publication, three basic predictions can be made about the data (Palmer, 1999). First, studies with small sample sizes should have increased sampling error relative to those with larger sample sizes; the variation around the cumulative effect size should decrease as sample size increases. Second, the cumulative effect size should be independent of sample size. Third, at a given sample size, individual studies should be normally distributed around the cumulative effect size due to random sampling error. If these predictions hold, the

resulting plot is shaped like a funnel with the large opening at the smallest sample sizes (Figure 3.2).

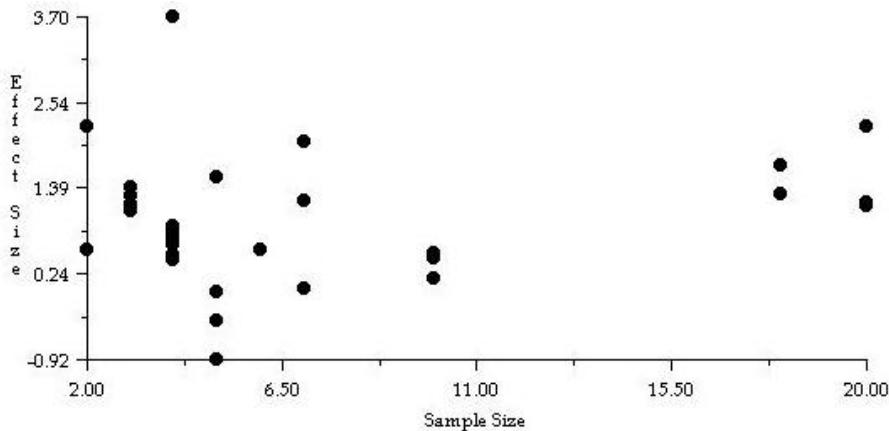


Figure 3.2. Example of a funnel plot.

Publication bias will tend to skew the shape of the funnel or the distribution of points within the funnel (Begg and Berlin, 1988; Palmer, 1999). If studies with small effect or nonsignificance are not included, there will tend to be gaps in the mouth of the funnel, especially near the null effect. Funnel plots can be enhanced by plotting points of different groups with different symbols. This plot may help one visually determine if there is structuring to the data or if certain types of data are driving the summary results (Palmer, 1999). Finally, it should be noted that funnel plots are occasionally represented as effect size versus variance, rather than versus sample size.

The *Normal Quantile Plot* has recently been suggested as an alternative to the funnel plot (Wang and Bushman, 1998). The normal quantile plot is a special case of a quantile quantile plot; this is a plot where two distributions are compared by plotting their quantiles against each other. If the two distributions are similar the quantiles will also be similar and the points will fall close to the line $X = Y$. Deviations from this line reveal how the distributions are different. In a normal quantile plot, one of the distributions is the standard normal distribution while the other is the distribution one wishes to compare it against (Figure 3.3).

Wang and Bushman (1998) suggest plotting the standardized effect size, $E_i / \sqrt{v_i}$, against the standard normal distribution. They list three advantages of the normal quantile plot to the funnel plot. First, it can be very difficult to determine if a funnel plot is actually funnel shaped, especially when the number of points is small; the normal quantile plot should be linear, which is much easier to interpret. This is further enhanced if confidence intervals are constructed around a normal quantile plot (Chambers et al., 1983). Second, funnel plots do not test the assumption of normality which most meta-

analytic procedures are based upon. Third, data can appear funnel shaped even when the studies come from multiple populations.

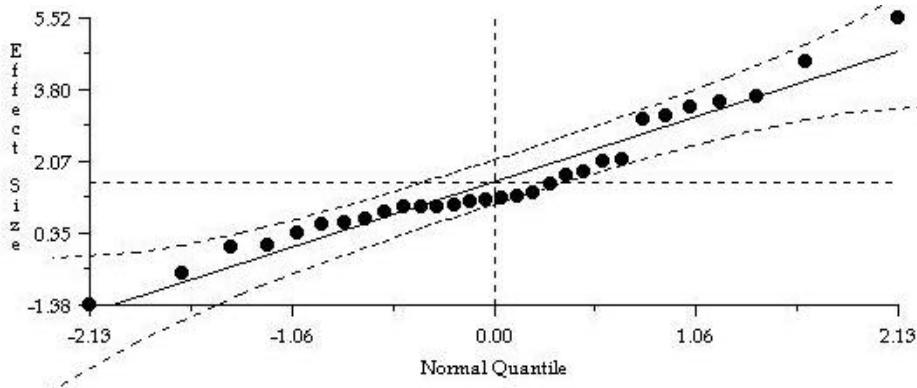


Figure 3.3. Example of a normal quantile plot.

The normal quantile plot has a number of interesting features. Deviations from linearity indicate deviations from normality; if the points do not fall within the confidence bands the data are not normally distributed. The slope of the linear regression line indicates the standard deviation of the data; this should be 1 for the standardized effect size if the studies come from a single population and have large sample sizes. The y -intercept of the regression indicates the mean. Publication bias will tend to leave strange gaps in the plot or lead to very non-linear curves. Multiple populations in the same data set can also lead to strange curves. See Wang and Bushman (1998) for more details on the interpretation of normal quantile plots in meta-analysis.

Rank Correlation Tests for Publication Bias

While the graphical methods are useful for exploring publication bias, they do not allow one to perform formal statistical tests. Begg (1994) suggested using a rank correlation test to search for a relationship between the standardized effect size, E^* , and sample size, n , across studies. This test is a statistical analogue of the funnel plot. A rank correlation test is one in which a correlation coefficient is calculated based on the relative ranks of the two variables, rather than on the actual values.

The standardized effect size of each study, E_i^* , is calculated as

$$E_i^* = \frac{(E_i - \bar{E})}{\sqrt{v_i}} \quad (3.1)$$

where E_i is the effect size for the i^{th} study, $\bar{\bar{E}}$ is the mean effect size, and v_i^* is the standardized variance of the i^{th} study, calculated as

$$v_i^* = v_i - \left(\sum \frac{1}{v_j} \right)^{-1} \quad (3.2)$$

where v_i is the variance of the i^{th} study and the summation is over all studies. The correlation between E^* and n is calculated using either of the common rank correlation tests, Kendall's tau or Spearman's rho (see Sokal and Rohlf, 1995, for more information about these tests). A significant correlation between E^* and n may indicate a publication bias where larger effect sizes in one direction (e.g. positive effects) are more likely to be published than smaller effect sizes. This test cannot detect publication bias where larger effect sizes are more likely to be published than smaller effect sizes, regardless of direction. The test has low power with only about 25 studies, but can be quite powerful as the number of studies reaches 75 or more (Begg and Mazumdar, 1994). Because there should be a strong correlation between sample size and study variance, the rank correlation could also be calculated between the standardized effect size and the standardized variance. More details and information about the characteristics of these tests can be found in Begg and Mazumdar (1994).

Fail-Safe Numbers

One of the simplest methods for estimating the magnitude of the file-drawer problem is the calculation of a fail-safe number. A fail-safe number is the number of non-significant, unpublished, or missing studies that would need to be added to a meta-analysis in order to change the results from the meta-analysis from significance to non-significance. If this number is large relative to the number of observed studies, one can feel fairly confident that the observed results, even with some publication bias, can be treated as a reliable estimate of the true effect.

The original (and most commonly used) fail-safe calculation was suggested by Rosenthal (1979). This method is based on the sum of Z method for combining probabilities (Stouffer et al., 1949; see Section 2.1 and Equation 2.3: **pp. xx**). Stouffer's method calculates a new Z -score by summing the Z 's of the individual studies and dividing by the square-root of the number of studies. The significance of this new Z is the significance of the combined studies. Rosenthal's method calculates the number of additional studies, N_R , with a mean effect size of zero, needed to reduce the combined significance to a desired α level (usually 0.05). N_R is calculated as:

$$N_R = \frac{\left(\sum_i^n Z(p_i)\right)^2}{Z_a^2} - n \quad (3.3)$$

where n is the number of studies, $Z(p_i)$ are the Z -scores for the individual significance values, and Z_a is the one-tailed Z -score associated with the desired α .

An alternative method, proposed by Orwin (1983), is based on the standardized mean difference effect size (see Section 2.2: **pp. xx**). Rather than test a specific α , this method calculates the number of additional studies, N_O , needed to reduce an observed mean effect size to a desired “minimal” effect size.

$$N_O = \frac{n(\bar{E}_0 - \bar{E}_m)}{\bar{E}_m - \bar{E}_n} \quad (3.4)$$

where n is the number of studies, \bar{E}_0 is the mean of the original n studies, \bar{E}_n is the mean of the additional N_O studies, and \bar{E}_m is the desired minimal mean effect size. Cohen (1969) defines a standardized mean difference effect size of 0.2 as “small,” 0.5 as “medium,” and 0.8 as “large.” Generally the minimal effect size chosen using Orwin’s calculation is 0.2 (e.g. VanderWerf, 1992).

Neither of these methods accounts for the sample size or variance of the studies. In order to include this in the calculation of a fail-safe number, L’Abbé et al. (1987) suggested simulating a single study of large negative effect and determining the sample size of that study necessary to raise an observed significance above 0.05. Similarly, they suggest simulating several relatively small studies of no effect and calculating how many of these studies it would take to raise an observed significance above 0.05.

Rosenberg (2000) suggested an alternative approach based on the modern meta-analysis methods of calculating summary statistics (see Section 2.3: **pp. xx**). This approach is independent of the effect size measure being used and is based on the calculation of the cumulative effect size (Equation 2.29). This method calculates how much total additional weight, N (associated with an unknown number of studies whose mean effect is zero) would be required to reduce the significance of an observed cumulative mean effect (as calculated from a t -test) down to a desired α level.

$$N = \frac{\left(\sum_i w_i E_i\right)^2}{t_{a[v]}^2} - \sum_i w_i \quad (3.5)$$

where w_i and E_i are the weight and effect size of the i^{th} study and t is the value of Student's t -distribution with v degrees of freedom (see discussion of v below). Dividing N by the mean weight, $\frac{\sum w_i}{n}$, yields

$$N^* = \frac{Nn}{\sum w_i}. \quad (3.6)$$

where n is the original number of studies. N^* is equivalent to the number of studies of null effect and mean weight necessary to reduce the observed significance level to α . N^* could also be interpreted as the relative size of a single study of no effect needed to reduce the significance level to α . Relative size means the study would have to have a weight N^* times the mean weight. Because we expect the weights to be roughly proportional to the sample size of each study (for some metrics this is exactly the case), N^* could also be thought of as an estimate of how many times larger the sample size of a single study (compared to the mean sample size) would need to be in order to reduce the significance of the mean effect to α .

One complication of this approach is that the degrees of freedom of the t -test, v , is based on the number of studies used to construct the mean. If N^* is interpreted as the relative size of a single study, $v = n - 1 + 1 = n$. However, if N^* is interpreted as multiple studies of mean weight, $v = n + N^* - 1$. In this case, N^* has to be solved for iteratively; because variation in t is quite small with even moderate degrees of freedom, only a few iterations are required for convergence (Rosenberg, 2000).

The advantage of this final method is that the fail-safe number is derived directly out of the methodology of modern meta-analysis; it should be a more accurate reflection of the number of additional studies needed to overturn observed results.

How large does a fail-safe number need to be in order for one to feel confident about ones results? This may be dependent on the field and how likely it is for unpublished studies to exist, but there are no hard rules. Rosenthal (1979) suggests the value $5n + 10$ (where n is the original number of studies) is a reasonable, conservative critical value to test a fail-safe calculation against.

3.2 Cumulative Meta-Analysis

If the primary research studies used in a modern meta-analysis can be arranged in a particular order (e.g., chronological), one may perform a *cumulative meta-analysis*. Cumulative meta-analysis is actually a series of meta-analyses, where studies are successively added to the analysis based on a predetermined order (Chalmers, 1991;

Yusuf et al., 1991; Antman et al., 1992; Lau et al., 1992; Chalmers and Lau, 1993). For each set of included studies, the summary meta-analytic statistics (usually the cumulative effect size and its variance) are calculated, another study is added, and the statistics are recalculated. The summary statistics are then placed in sequential order, and compared to one another (Figure 3.4). If the overall cumulative effect size from all studies is significant, one may ask when a statistically significant conclusion was first able to be detected. Although each cumulative meta-analysis is typically tested against the critical value of $\alpha = 0.05$, one may wish to adjust the critical value using a Bonferroni procedure, to reflect the fact that multiple statistical tests are performed on the same data (Chalmers, 1991; Yusuf et al., 1991; Lau et al., 1992; for a discussion of Bonferroni procedures, see Sokal and Rohlf, 1995). Typically, as studies are added to the analysis, the cumulative effect size first changes greatly from analysis to analysis, then it stabilizes and converges on the value found from the last cumulative analysis.

The idea of cumulative meta-analysis was first mentioned by Chalmers (1991), who later co-authored the first two cumulative meta-analyses (Antman et al., 1992; Lau et al., 1992). In one of these now classic papers (Lau et al., 1992), the authors analyzed the effectiveness of streptokinase drug therapy for patients with myocardial infarction. They included data from 33 published studies from 1959 and 1988 in their analysis. Using cumulative meta-analysis they discovered that the evidence in favor of drug therapies was significant, and that this overall result could actually have been determined thirteen years before the experts recommended the widespread use of such therapies!

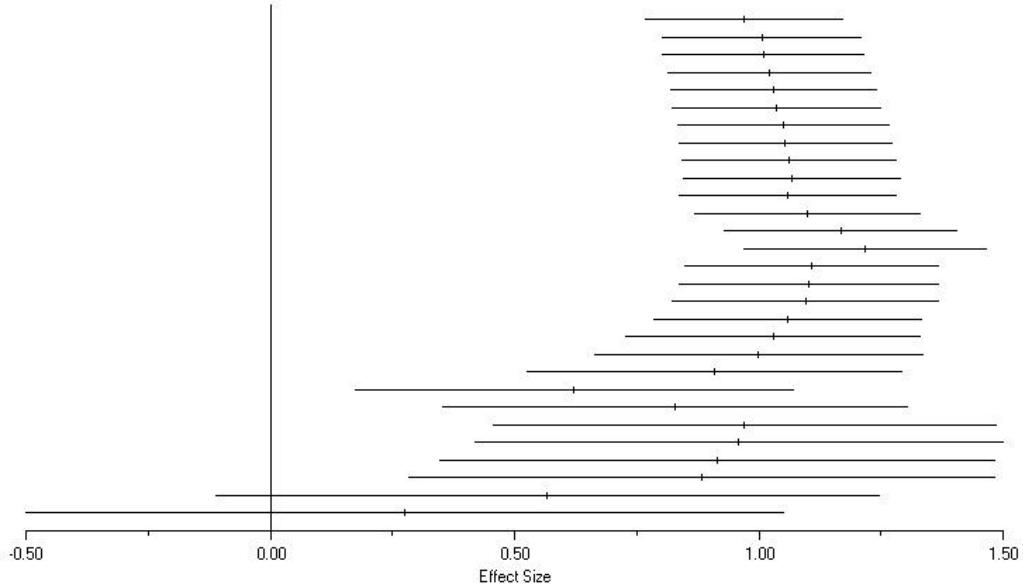


Figure 3.4. Example of a cumulative meta-analysis (Note: Results are presented such that the analysis at top of the graph includes all studies, while the analysis at the bottom includes only the first study. This is opposite to the presentation in Lau et al., 1992).

Interestingly, in the medical sciences, advances in cumulative meta-analysis has rekindled an old debate between those that advocate the use of large clinical trials, versus those that feel many smaller studies could yield the same results. Traditionally, recommendations for widespread use of a new drug or therapy were given only after one or more large, randomized controlled studies had been performed. However, several authors have demonstrated that the data from a series of smaller studies, when accumulated, is as strong (or stronger) than the evidence from a larger, controlled study. Thus, several authors advocate the use of small studies and cumulative meta-analysis as a means of identifying the benefit of a medical procedure as quickly as possible (for a discussion see Pogue and Yusuf, 1997). Indeed the results from Lau et al. (1992) implied that relying on traditional large-scale studies delayed the widespread use of drug therapies for myocardial infarction by nearly thirteen years. While a full discussion of this debate is beyond the scope of this manual, interested readers are referred to the following: Chalmers and Lau (1993), Flather et al. (1997), Pogue and Yusuf (1997), and Yusuf (1997).

3.3 Resampling Methods in Meta-Analysis

Resampling statistics are computer-intensive techniques that allow one to evaluate the significance of a given test value (Crowley, 1992; Manly, 1997). They are an alternative to conventional parametric statistics, which generally have more restrictive assumptions. Resampling tests are often useful when the original data do not conform to the distributional assumptions of parametric tests (Manly, 1997). Moreover, because they use the actual data rather than ranks, they are more powerful than traditional nonparametric approaches as well (Edgington, 1987; Adams and Anthony, 1996; Manly, 1997). Finally, resampling tests have been found to be particularly useful in taking into account the peculiarities of a given data set (Manly, 1997), such as a non-standard test statistic, or a data set with both normal and non-normal distributions (Adams and Anthony, 1996).

Resampling statistics are performed by calculating a statistic from the original data, and evaluating it by permuting the original data in some way, recalculating the test value of interest, and then repeating this procedure many times. Each permutation of the original data is called an *iteration*. The test values from all of the iterations are then used to generate a distribution of test values, and the original test value is compared to this generated distribution to determine the statistical significance of the original data (Edgington, 1987; Crowley, 1992; Manly, 1997).

Because meta-analytic data often have small sample sizes and may violate basic distributional assumptions (e.g., normality), resampling techniques can be important for accurately determining the significance of meta-analytic metrics.

Bootstrapping

Bootstrapping can be used to generate confidence intervals around a given statistic. Bootstrapping works by randomly choosing (with replacement) n studies from a sample size of n , and then calculating the desired statistic. For example, if there were twenty studies in total, twenty studies would be chosen for each bootstrap iteration. However, because bootstrapping is sampling with replacement, some of the studies from the original sample would be chosen more than once, while others would not be chosen at all. This procedure is repeated many times to generate a distribution of possible values. The lowest and highest 2.5% values are then chosen to represent the lower and upper 95% bootstrap confidence limits (or whatever percentiles are appropriate for the desired α value).

Confidence intervals generated in this way are called *percentile bootstrap confidence intervals*, because they are calculated by merely choosing certain percentile values (Dixon, 1993). These confidence intervals assume that the distribution of bootstrap values is centered around the original value. When this is the case, the percentile bootstrap is known to produce the correct confidence intervals (Efron, 1982; Dixon, 1993). However, when more than 50% of the bootstrap replicates are above or below the original value, the bootstrap confidence interval must be corrected for this bias. This is done by first finding the fraction, F , of bootstrap replicates smaller than the observed value. The upper and lower percentiles for the bias-corrected bootstrap are found as

$$\Phi[2\Phi^{-1}(F) \pm Z_{\alpha/2}] \quad (3.7)$$

where Φ is the normal cumulative distribution function, Φ^{-1} is the inverse cumulative normal distribution function, $Z_{\alpha/2}$ is the Z -score, and α is the Type I error associated with the confidence interval (i.e. for a 95% confidence interval, $\alpha = 0.05$). If $F = 0.5$, there is no bias and the upper and lower percentiles equal 2.5% and 97.5%. For complete details see Dixon (1993). In meta-analysis, we can use bootstrapping to generate a confidence interval about the cumulative mean effect size (Adams et al., 1997). In a categorical model, bootstrapping can be used to generate confidence intervals around cumulative effect sizes for groups, as well as the overall cumulative mean effect size.

Randomization Tests

Randomization tests are most frequently used to determine the significance level of a given test statistic. In meta-analysis, randomization tests can be used to test the

significance of the model structure, Q_M (see Section 2.3: **pp. xx**). For each iteration, the original effect size data are randomly reassigned to the structural design. For example, in a categorical model, studies are randomly assigned to groups such that the number of studies in each group is unchanged. A test statistic is then calculated using the randomly shuffled data. This represents one possible outcome based on the data. By performing many iterations, a frequency distribution of possible outcomes (i.e. test statistics) is generated. The actual test statistic is then compared to this frequency distribution of randomly generated statistics, and the proportion of randomly generated statistics more extreme than the actual statistic is taken to be the significance level for that data set. For example, if 301 of 4999 randomly generated values are equal to or larger than the actual test value, the probability of the original data value being more extreme than is expected by chance is 302/5000, or $P = 0.0604$. (the original data structure is considered to be an additional permutation, thus 302 and 5000, rather than 301 and 4999 are used: see Edgington, 1987).

With randomization tests, the value of the probability level is partially a function of the number of iterations used. Increasing the number of iterations allows one to decrease the value of the smallest detectable probability. For example, a randomization test using 49 iterations can detect a probability as low as 0.02, whereas a randomization test using 4999 iterations can detect a probability as low as 0.0002. Usually, between 999 and 4999 iterations are sufficient to get a reliable probability level (Potvin and Roff, 1993; Adams and Anthony, 1996; Manly, 1997). As the number of iterations increases, the standard error of replicated randomization tests decreases such that a further increase of iterations does not provide much more accuracy.

4. RUNNING METAWIN

You can start MetaWin the same way you start any Windows-based application: by double-clicking the metawin.exe program icon in Explorer, or by selecting it from the Start-Menu tree. The Main Dialog window of MetaWin is visible upon start-up (Figure 4.1). This window is a shell which contains all subsequent program windows. When first opened, the MetaWin data spreadsheet is visible in the Main Dialog window. The output window is layered behind the spreadsheet editor, and can be viewed by choosing it from the Windows menu, or by rearranging the windows with the Tile or Cascade commands found in the Windows menu.

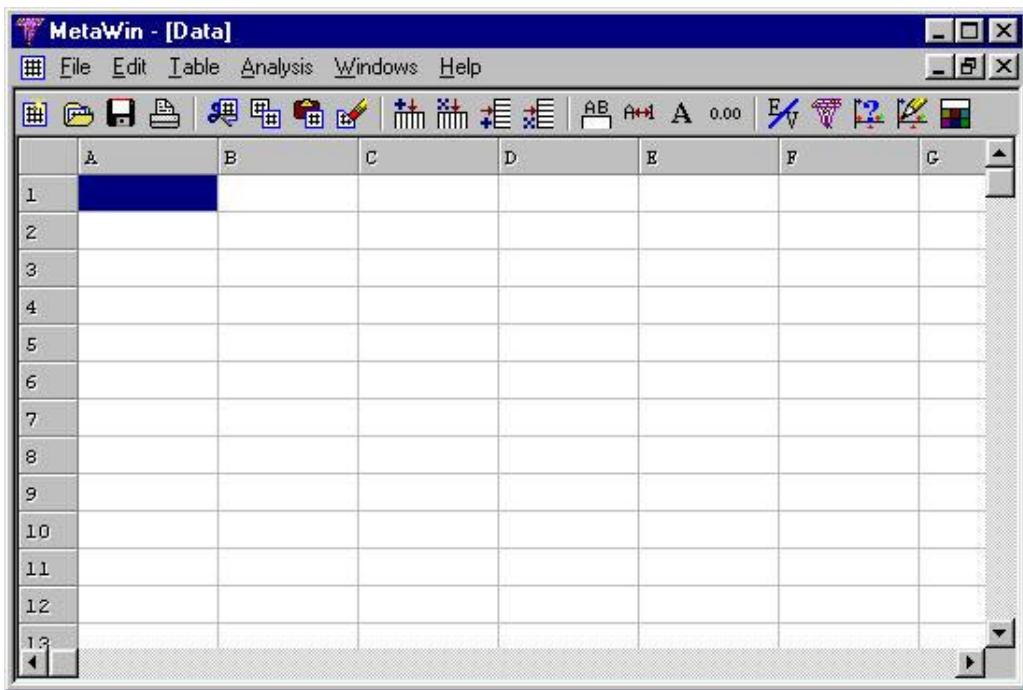


Figure 4.1. Main Dialog window of MetaWin with the data spreadsheet displayed as the active window.

The general commands for operating MetaWin can be found in the pull-down menus. Additionally, on each sub-window of MetaWin (e.g., data spreadsheet) there are special program buttons for those features unique to that window. There are seven menus: File, Edit, Table, Graph, Analysis, Windows, and Help. Some of these menus (e.g., Table and Graph) are visible only when the appropriate window is active. The File (Figure 4.2) and Edit menu choices change based on which sub-window of MetaWin is active. An explanation of the commands found in the pull-down menus will be found in the appropriate sections in the following chapters. A list of menu commands is found in Table 4.1.

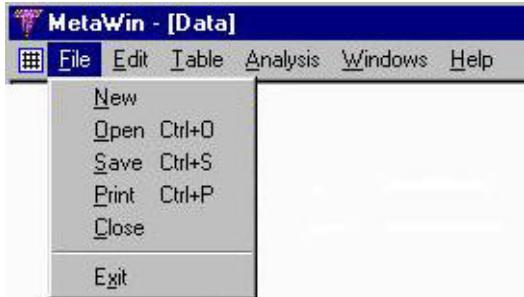


Figure 4.2. Commands found under the File menu. Note that some commands may be accessed through “Control” key strokes.

Table 4.1. Commands found under the menus from the Main Window of MetaWin.

Menu	Commands
File	New, Open, Save, Print, Close, Exit
Edit	Cut, Copy, Paste, Clear Cells, Select All, Show Toolbar
Table	Insert Row, Delete Row, Insert Column, Delete Column, Column Format, Rename Column, Mark As Label Column, Unmark Label Column, Font, Decimal Places, Text Column Color, Numeric Column Color, Label Column Color
Graph	Point/Line Style, Format Graph, Show Legend, Show Labels, Sort Effects By, # Histogram Classes
Analysis	Effect Sizes, Summary Analyses, Data Exploration, Draw Graph, S-Calculator, Options
Windows	Tile, Cascade, Arrange Icons
Help	Contents, Index, About

4.1 Data Spreadsheet

When first opened, the MetaWin data spreadsheet is visible and active in the Main Dialog window. This spreadsheet allows for general data input, and the size of the file is limited only by the amount of RAM and free space on your hard drive (i.e. there is no longer a limit on the number of studies or grouping variables). There are several program buttons on the MetaWin spreadsheet toolbar that allow you to manipulate data and data files (Table 4.2). These include buttons for file manipulation, data manipulation, and spreadsheet manipulation, and are identical to commands found under the various menus. There are also buttons for the general meta-analytic procedures performed in MetaWin. In addition to program buttons, when the spreadsheet editor is the active window, the Table menu is available.

Table 4.2. Program buttons from the MetaWin spreadsheet toolbar.

Button	Name	Menu	Command
	New	File	Create blank spreadsheet
	Open	File	Open data file
	Save	File	Save data to file
	Print	File	Print data
	Cut	Edit	Cut selected data cells to clipboard
	Copy	Edit	Copy selected data cells to clipboard
	Paste	Edit	Paste from clipboard
	Clear Cells	Edit	Clear selected data cells
	Insert Column	Table	Insert column in spreadsheet
	Delete Column	Table	Delete column from spreadsheet
	Insert Row	Table	Insert row in spreadsheet
	Delete Row	Table	Delete row from spreadsheet
	Rename Column	Table	Rename a column
	Column Format	Table	Change column format
	Change Font	Table	Change spreadsheet font
	Decimal Places	Table	Change decimal places
	Effect Sizes	Analysis	Calculate effect sizes
	Summary Analyses	Analysis	Calculate summary statistics
	Data Exploration	Analysis	Perform data exploration
	Draw Graph	Analysis	Draw graphics
	S-Calculator	Analysis	Start statistical calculator

Data Structure

Unlike MetaWin version 1.0 (Rosenberg et al., 1997), the data format is now completely general: no header line is necessary and data columns may be in any order. The only restriction on data format is that rows are data for individual studies, and columns are variables. Thus, with a more general file format and no limits on the size of the data file, the current version of MetaWin is much more flexible than the previous version.

MetaWin recognizes two types of data: *numeric data* and *text data*. Each data column is defined as a numeric column or a text column depending on the type of data that is stored in it. Numeric data are stored as real, double-precision floating-point numbers. For a column to contain numeric data, every row in the column *must* either contain a valid number or be blank. All other data are stored as text. By distinguishing types of data in the spreadsheet, MetaWin is more memory efficient and is better able to prevent potential calculation errors before they occur (e.g., trying to use study names as an effect size!). While MetaWin allows you to choose which column to use for various types of analyses, the program will only display those columns which are appropriate for the type of analysis being performed. Note that columns containing grouping variables *must* be designated as text columns, even if the groups are defined as: 1, 2, 3, etc. (see Section 5.2).

MetaWin automatically detects what type of data a column contains when it is loaded (see next Section, “Loading Data”). Any column with a non-numeric entry in it is automatically considered to contain text data. It is possible to change the format of a column from numeric to text or text to numeric (if the text entries are all valid numbers). See the Section “Manipulating Data” for details.

Loading Data

Data can be input into the MetaWin spreadsheet in one of three ways. First, an empty spreadsheet may be obtained by selecting File | New from the menu, or by clicking on the  button. Data may then be entered directly into the new spreadsheet by typing the data values into the spreadsheet cells. Alternatively, you can load data from a data file. To open a file, select File | Open from the menu (or click on the  button), then choose the file name in the Open Dialog box (Figure 4.3). You must choose whether the data file is a text file, a Lotus file, or an Excel file. Currently, MetaWin reads all versions of Excel through Excel2000 (9.0), and all versions of Lotus through version 3.x. For any type of data file, you may select whether or not the data file contains a header line. A header line contains the names of the data columns, but is not required. For text files, you may select whether the data are separated by spaces, tabs, commas, or another delimiter (or combinations thereof). Multiple consecutive delimiters are treated as just one.

There is an additional way to load data into MetaWin. If the data are in a spreadsheet (e.g., Excel), you may copy and paste directly into the MetaWin spreadsheet. You can also copy *from* MetaWin and paste into Excel (this may be useful after you have calculated effect sizes). Note that you cannot paste column headers into the MetaWin spreadsheet.

There may be limitations on MetaWin's ability to import Excel and Lotus files. Newer versions of Excel and Lotus contain workbooks with multiple worksheets. MetaWin can only handle a single worksheet and will therefore only read the *first* worksheet in a file. Additionally, MetaWin does not directly evaluate formulas, but rather reads the value calculated by the spreadsheet program at the time the file was saved. MetaWin also does not read cell formats; this may cause types of data to load in unexpected ways. For example, spreadsheets usually store dates internally as integers. In Excel, the date "5/21/97" will import as "35571" and "4/23/96" as "35178." If you are having problems importing a spreadsheet file, try copying the data from the spreadsheet and pasting it into MetaWin, or save it as a text file before importing into MetaWin.

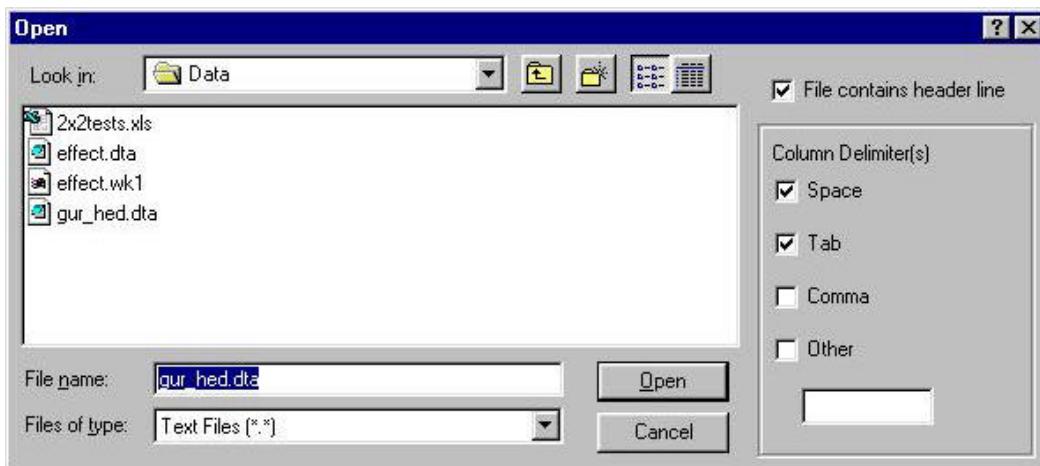


Figure 4.3. Open Dialog box of MetaWin.

Manipulating Data

With the MetaWin spreadsheet we have *not* attempted to reproduce all of the features found in Excel or Quattro Pro. Therefore, if the preliminary data manipulations involve more complicated operations (such as the use of equations) we recommend that you perform these in a more comprehensive spreadsheet package prior to importing the data into MetaWin. Nevertheless, you will find many of the commonly-used features of a spreadsheet editor in the MetaWin spreadsheet, which we have included to make data entry and editing easier. We discuss these operations below.

Using the MetaWin spreadsheet is much like using any other spreadsheet package. Of the commands included several are aesthetic in nature, while others are designed to make data entry and analysis easier. These commands may be found under the Table menu (Figure 4.4), and many have special buttons on the MetaWin spreadsheet toolbar (see Table 4.2). For example, you may change the font of the data in the spreadsheet by selecting Table|Font, or by clicking on the button. You can also change the

number of decimal places displayed for numeric data in a similar manner (all numeric columns are displayed to the same number of decimal places). The widths of rows and columns may be altered by clicking and dragging on the separator lines between column headers or row labels.

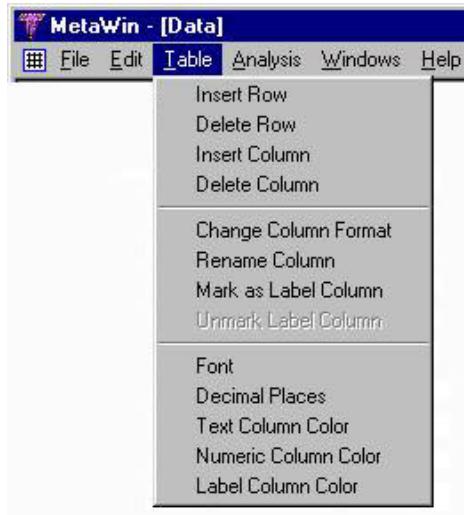


Figure 4.4. Commands found under the Table menu.

Other common spreadsheet commands can also be found in the MetaWin spreadsheet. These commands include cut, copy, paste, clear cells, and select all. They may be accessed from the Edit menu (Figure 4.5), or by using the program buttons on the MetaWin spreadsheet toolbar (see Table 4.2). Additionally, clicking the right mouse button in the main body of the spreadsheet accesses a sub-menu of data editing commands. Finally, all of the editing commands can be accessed by using “Control” key strokes, which are found in Figure 4.5 below.

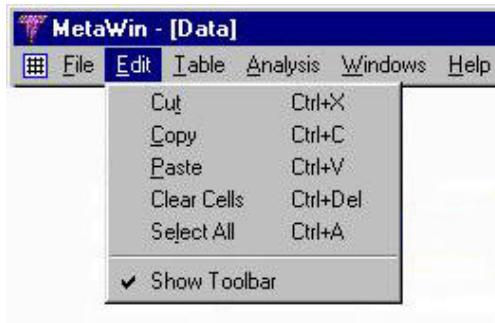


Figure 4.5. Commands found under the Edit menu and their “Control” key strokes.

Using the editing commands in MetaWin is similar to using the edit commands in any spreadsheet program. First, you highlight the desired data cells in the spreadsheet and then select the appropriate operation to perform on those cells. This applies for cutting

data, copying data, pasting data, and clearing data cells. When data cells are selected, they appear as light blue, while the active data cell is dark blue (Figure 4.6). It should be noted that the paste operation *begins* from the active cell, so if you have selected a section of the spreadsheet editor to paste data into, **be sure to select the section from the bottom-right cell to the top-left cell, so that the top-left cell is the active cell.** Also, there is no undo command in the MetaWin spreadsheet, so save your data frequently!

	Sex	Tree	State	+/-
1	M	Oak	NY	+
2	M	Maple	NY	-
3	M	Maple	NY	-
4	M	Oak	NJ	-
5	M	Maple	NJ	+
6	M	Maple	NJ	+

Figure 4.6. Example of highlighted cells selected in the data spreadsheet, with bottom-right cell as the active cell.

There are several other spreadsheet manipulation commands available from the Table menu (Figure 4.4). These allow you to insert and delete rows, insert and delete columns, rename a column, and change a column's format.

Inserting Rows and Columns: You may insert blank rows or columns by selecting Table | Insert Row or Table | Insert Column from the menu, clicking the or button, or by right clicking on the row or column header and choosing insert. New rows are inserted above the active row; new columns are inserted to the left of the active column.

Deleting Rows and Columns: To delete the active column or row, choose Table | Delete Row or Table | Delete Column from the menu, click the or button, or right click the row or column header to be deleted. Note that you cannot select multiple rows or columns. To delete multiple rows or columns you must delete each individually.

Naming Columns: Columns may be given names by clicking on any cell in that column and selecting Rename column from the Table menu, or by clicking the button. Additionally, you may select a column and click the right mouse button on the header for that column, and choose Rename Column from the sub-menu that appears. If you loaded a data file with column headers, they will be used as column names.

Changing Header Colors: MetaWin displays column headers in different colors to distinguish between the different data types (see “Data Structure” above). The default colors are gray for text columns and yellow for numeric columns. These colors can be changed by choosing the appropriate menu option from the Table menu (Figure 4.4). These color choices will be remembered by MetaWin the next time you run the program.

Label Column: You can also designate a particular column as the label column. This column typically has the names of the individual studies, and, when designated, these names will be used throughout all subsequent analyses. To designate a label column click on any data cell in the desired column and select Table|Mark As Label Column from the Table menu, or click the right mouse button on the column header and select Mark As Label Column from the sub-menu. Only text columns can be chosen as labels. The default color of the label column is green. You may also unmark a label column by choosing Unmark Label Column from the Table menu. The default label column color may be changed by selecting the appropriate command from the Table menu.

Changing Column Format: The format of a column (numeric or text) may be changed by using the Change Column Format option on the table menu, by clicking the  button, or by right clicking on the column header and choosing Change Column Format. Numeric columns can always be changed to text columns, but a text column can only be changed to a numeric column if all of the rows contain valid numbers, or are blank (see “Data Structure” for more information).

Note: Changing a numeric column to text and then back to a numeric column may cause a loss of information due to rounding error. When the conversion from a number to text is initially made, the new string is written using as many decimal places as are currently displayed in the spreadsheet. If the number is accurate to more decimal places, these are lost in the subsequent change back to numeric format. For example, suppose the number 4.234763 is stored in a numeric column which is displaying three decimal places. The number appears in the data cell as “4.235” but if you were to change the number of decimal places to six, it would appear as “4.234763.” When you change the format of that column to text, the resulting cell contains “4.235.” Changing the column format back to a numeric column stores the number as 4.235. If the number of decimal places is now changed to six, the value would be “4.235000.”

Saving and Printing Data

You may save the data by selecting File|Save or clicking the  button. This brings up the Save Dialog box (Figure 4.7), which is shown below. Data from the MetaWin

spreadsheet can only be saved as text files; MetaWin will not save data as Excel or Lotus files. If one of these file formats is desired, one may click on Edit | Select All to highlight the data, then copy and paste it into the appropriate spreadsheet package. While MetaWin can only save data from the spreadsheet as text, you may choose between four different text file formats. MetaWin saves text as tab delimited, space delimited, and comma delimited, as well as formatted text. Formatted text files are space delimited files where additional spaces are added so that the columns of data line up vertically in the text file, thus making it easier to read.

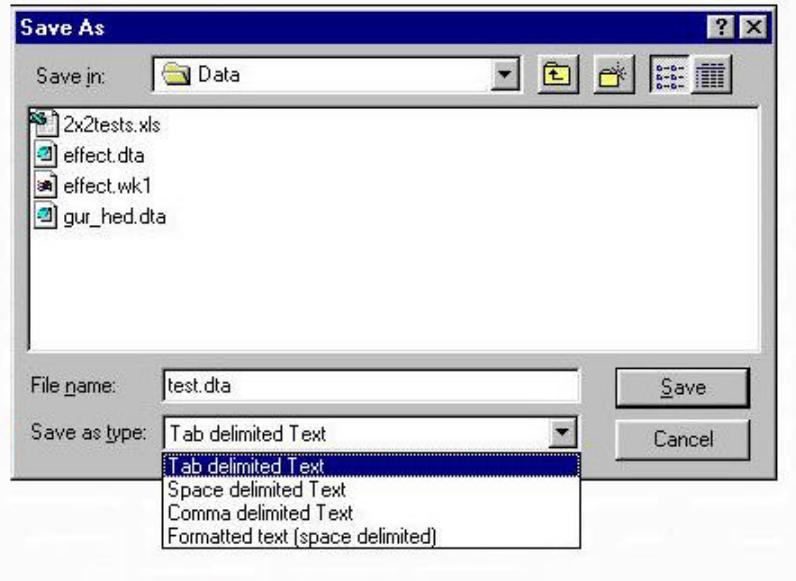


Figure 4.7. Save Dialog box of MetaWin.

Note: When saving the data in a text file, all numeric columns are written to the number of decimal places displayed in the spreadsheet. Any numbers which were accurate to more decimal places will be truncated. See the note on changing column formats (pp. xx) for more details.

Data in the spreadsheet may be printed by selecting File | Print or by clicking the button. Although the gridlines are not printed, the data is printed such that it resembles the spreadsheet as closely as possible, using the currently selected font, number of decimal places, and column widths. The printed text closely resembles the "Formatted Text" option of saving data files.

4.2 Output Window

MetaWin records the results from all of its operations in the MetaWin output window (Figure 4.8). For example, loading and saving data files, calculating effect sizes, and

performing summary analyses are all recorded in this window (specific descriptions of the output for each analysis is found with the description of that analysis). The MetaWin output window is always present, but is not always the active window. It can be viewed by choosing Windows|Output from the menu, or by rearranging the windows with Tile or Cascade.

Several operations may be performed while the MetaWin output window is active, the most important of which are saving and printing. Because the output window is the only location where summary meta-analytic statistical results are found, you may wish to save the contents of the output window after each analysis. Any information saved from the output window is saved as a text file. Additionally there are several editing options that are available in the output window which can be accessed through the Edit menu, or by clicking the buttons on the output window toolbar (see Table 4.2). These function identically to the edit options in the MetaWin spreadsheet and include cut , copy , paste , and clear . The option for changing the output font is also found under the Edit menu, or one may use the Font button on the output window toolbar. You may also type information directly into the MetaWin output window, so additional notes and comments from analyses can be included for future reference. Finally, all information from the output window may be eliminated by selecting Edit|Clear All output.

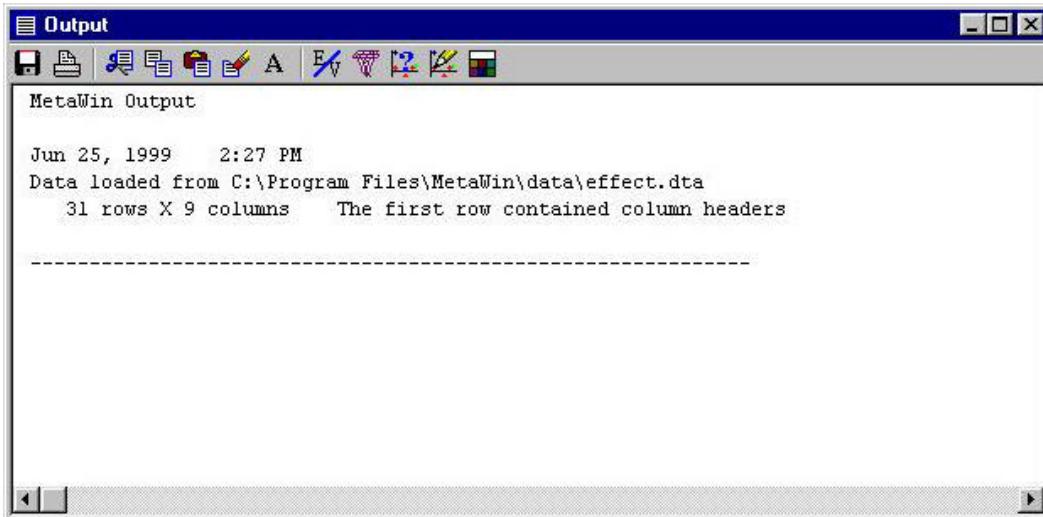


Figure 4.8. MetaWin Output window.

4.3 Program Options

MetaWin has several general program options that may be altered by the user. All of these options are saved when the program is closed and will be reloaded automatically

the next time you run MetaWin. To open the Options Dialog box (figure 4.9), choose Analysis | Options from the menu. The options include:

Decimal Places: You can choose how many decimal places to display for different types of output. These include effect size estimates, heterogeneity, Q , statistics and variances, and probability values. Note that probability values from resampling tests (see Section 5.2) are automatically displayed to the correct number of decimal places based on the number of iterations.

Confidence Intervals: A number of routines calculate confidence intervals around the cumulative mean effect size estimates. You may select the width of the confidence interval to be calculated. This applies to parametric confidence intervals and confidence intervals based on resampling tests (bootstrapping). For most applications you will most likely want to use 95% confidence intervals, but MetaWin allows you to choose any interval between 75% and 99%.

Case-Sensitivity: This option tells MetaWin whether or not to treat group names as case-sensitive (see Section 5.2 for more details on group names). Case-sensitive names distinguish between capitalized and lowercase letters; non-case-sensitive names do not. For example, if case-sensitivity is turned on “New York” and “new york” will be treated as separate groups.

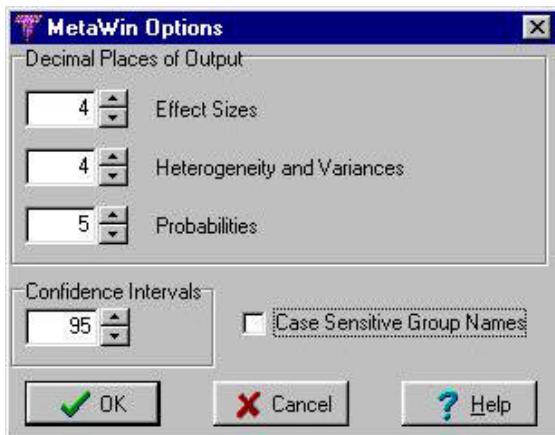


Figure 4.9. MetaWin Options box.

5. CONDUCTING A META-ANALYSIS

In this chapter we discuss how to use MetaWin to perform the basic calculations of a meta-analysis. This includes the calculation of effect sizes and the calculation of meta-analytic summary analyses. The sub-windows for performing these operations are accessed by selecting the appropriate option from the Analysis menu (Figure 5.1), or by clicking on one of the program buttons located on the MetaWin toolbars. In the following sections we use several example data sets to illustrate how to perform the basic meta-analytic calculations. Full descriptions of these data sets are found in Appendix I. In addition, the sample data files distributed with MetaWin are the same as the example data files used in this manual.

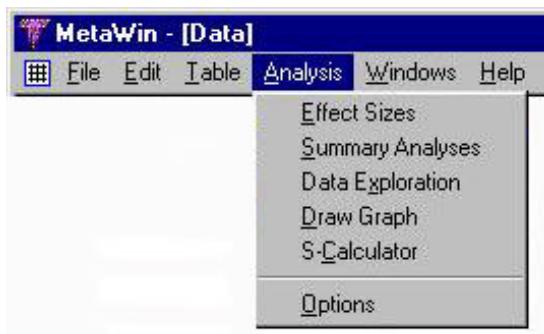


Figure 5.1. Commands found under the Analysis menu.

5.1 Calculating Effect Sizes

Unlike version 1.0, effect sizes are calculated in MetaWin 2.0 as a separate effect size routine, rather than being calculated as the first part of the meta-analysis summary statistics routine. This is done because the mathematical machinery for meta-analytic summary statistics is the same regardless of the choice of effect size (see Section 2.4). If the data file already includes effect sizes and their variances, you may skip the procedures in this section and proceed directly to the calculation of summary statistics (see Section 5.2).

The effect size window is accessed by selecting Analysis | Effect Sizes from the MetaWin menu, or by clicking the button. The effect size window has two tabs: one is for selecting the type of data to be used and one is for selecting the location of the data columns (Figure 5.2). MetaWin can be used to calculate effect sizes and their sampling variances from three types of primary data: 1) means, standard deviations, and sample sizes, 2) two x two contingency table data, and 3) correlation coefficients and sample sizes. Once a data type has been selected, a list of effect sizes that can be calculated from this data appear in the Effect Size box on the right side of the Data Type tab. Results from effect size calculations appear in both the MetaWin output window and

the data spreadsheet (see “Results from Effect Size Calculations” below). Note that the OK button will not be active until **ALL** necessary data columns have been selected! Because the options for data column selection vary with data type, we discuss each separately.

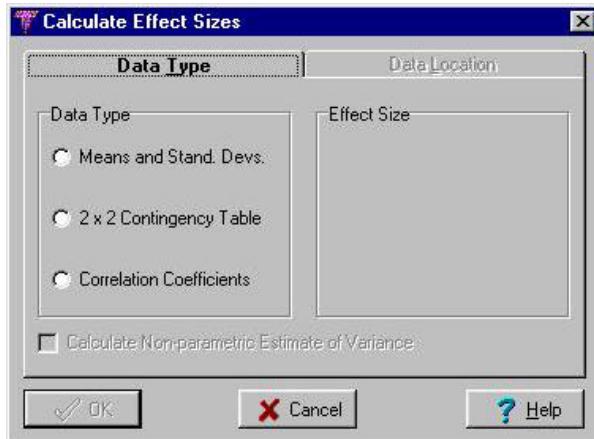


Figure 5.2. Effect size calculation window.

Means and Standard Deviation Data

If you select means and standard deviation data as your data type you can use MetaWin to calculate two different effect sizes: Hedges’ d , and the natural log of the *response ratio*, $\ln R$ (for a discussion of these statistics, see Section 2.2: pp. xx). Choose the type of effect size you wish to calculate from the Effect Size box on the Data Type tab (Figure 5.3). Next, click on the Data Location tab. On the Data Location tab you will find list boxes for the experimental and control group means, sample sizes, and standard deviations. Use these to select the appropriate data columns from the data spreadsheet. Notice that only *numeric* columns are visible in these list boxes! If one of your data columns is not listed, cancel out of the effect size window, return to the MetaWin Spreadsheet, and change the format of that column from text format to numeric format.

There are several other options that may be selected when using means and standard deviations as data. For example, a checkbox for a reversal column is present on the data location tab (Figure 5.3). If selected, choose the appropriate column from the data spreadsheet corresponding to the column of reversal markers. This column may consist of ‘+’ and ‘-’ or ± 1 , and may be either a text or numeric data column. Reversal markers are used to keep the sign of the expected effect comparable among studies. This is necessary when some studies have been measured in such a way that a “positive” response for those studies is in the opposite direction from other studies. For example, Gurevitch et al. (1992) were interested in examining the effect of competition on biomass, which could be experimentally determined by either removal or addition of

the competing species. For their data, a positive effect size would indicate the existence of competition in experiments where the competitor was removed, while a negative effect would indicate the same in experiments where the density of competitors was increased. By using reversal markers, responses from both removal and addition experiments could be included in the same analysis while keeping “natural density” as the control for all studies.

When using Hedges' d , you may additionally choose to calculate nonparametric estimates of the sampling variance for each study by clicking on the checkbox found on the data type tab. Nonparametric variance estimates use only the sample sizes from the experimental and control groups rather than incorporating the effect size into the calculation (Adams et al., 1997). Such estimates may be less constrained by the assumptions of large sample theory. If chosen, the nonparametric variances are calculated as:

$$\nu_{np} = \frac{N_i^C + N_i^E}{N_i^C N_i^E} \quad (5.1)$$

where N_i^C and N_i^E are the sample sizes from the experimental and control group of the i^{th} study (see Adams et al., 1997). Finally, it is nonsensical to use $\ln R$ when the experimental and control group means are of different signs (i.e., one is negative and one is positive), because the use of a ratio measures the *relative* magnitude of two effects, and thus assumes that their responses are in the same direction.

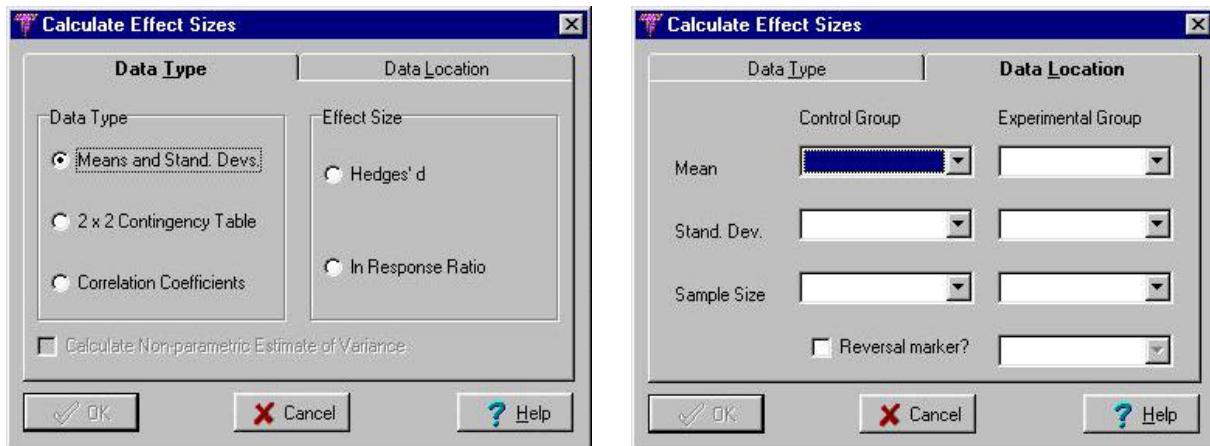


Figure 5.3. Options found on the Data Type tab and the Data Location tab of the effect size window when means and standard deviation data are selected.

Two x Two Contingency Data

If you select two x two contingency data as your data type you can calculate three different effect sizes using MetaWin: *odds ratio*, *rate difference*, and *relative rate* (for a discussion of these statistics, see Section 2.2: **pp. xx**). Choose the effect size metric you wish to calculate from the Effect Size box on the Data Type tab (Figure 5.4). Next, click on the Data Location tab. On the Data Location tab you will find list boxes for the four data cells of a two x two contingency table (e.g., Table 2.1): the number of responses in the control and treatment groups and the number of non-responses in the control and treatment groups (these are equivalent to A, B, C, and D in Table 2.1). Use these to select the appropriate *numeric* data columns from the data Spreadsheet. If there are empty cells in the two x two table for a particular study (i.e., zeros), MetaWin will not calculate the effect size for that study. When this is the case, 0.5 is often added to the values in each of the four data cells (for a discussion see Haddock et al., 1998). This correction is not performed by MetaWin.

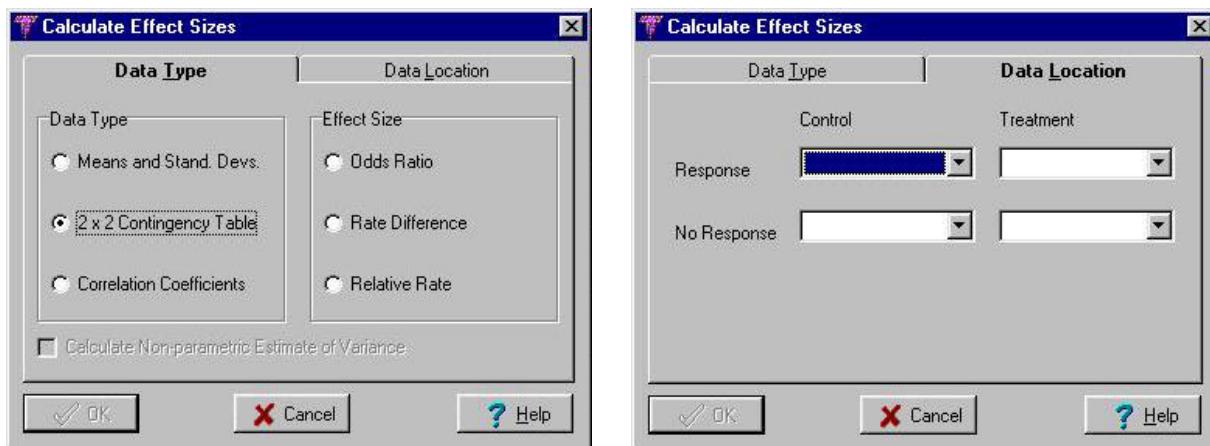


Figure 5.4. Options found on the Data Type tab and the Data Location tab of the effect size window when two x two contingency table data are selected.

Correlation Data

If you select correlation coefficient data as your data type, MetaWin will calculate Fisher's z-transformation as the appropriate effect size (see Section 2.2: **pp. xx**). Choosing correlation coefficients as your data type selects Fisher's z-transformation in the effect size box on the Data Type tab (Figure 5.5). On the Data Location tab are list boxes for correlation coefficients and their sample sizes. Use these to select the appropriate *numeric* data columns from the data Spreadsheet. Note that MetaWin only calculates Fisher's z-transform from correlation coefficients and their sample sizes. Other test-statistics may be converted to correlation coefficients using the MetaWin Statistical Calculator (see Chapter 7).

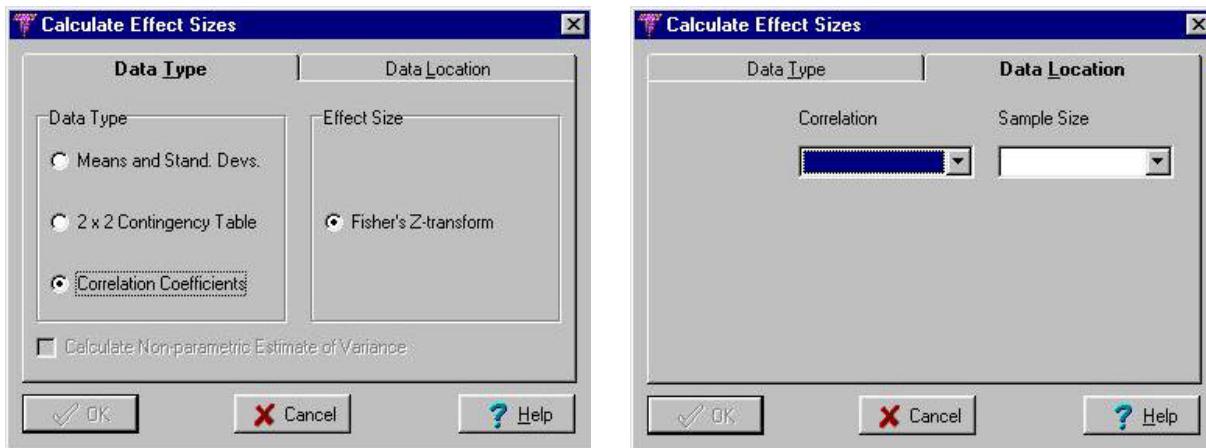


Figure 5.5. Options found on the Data Type tab and the Data Location tab of the effect size window when correlation coefficient data are selected.

Results from Effect Size Calculations

Once you have selected the appropriate data type and effect size for your data and have filled in the necessary list boxes, the OK button is activated and you can calculate effect sizes for your data. Results from effect size calculations appear in both the output window and the data spreadsheet. In the output window, MetaWin records both the time and the date of each analysis, states which type of effect size was calculated, and provides a table of effect sizes and variances for each study. If effect sizes for any individual studies are not able to be calculated (see below), a message will appear in the output window. In the data spreadsheet, several additional data columns are appended to the original data matrix. These columns are labeled as the effect size calculated (e.g., Hedges' d) and its variance. If you calculate more than one effect size for the same data, additional columns are added to the spreadsheet for each effect size. Note that if Hedges' d was chosen as the effect size and nonparametric estimates of study variances was selected, three columns will be added to the data spreadsheet (Hedges' d and its parametric and nonparametric variance: see above).

An example of the output from the calculation of effect sizes is found in Figure 5.6. In this example, no label column was specified, so by default, the studies are labeled "Study 1, Study 2," etc. If a label column had been selected, each study would be identified in the output window by the name found in the label column. In addition, the effect size for one study (Study 7) could not be calculated. The output displayed in Figure 5.6 has been cut to save space.

MetaWin - [Output]		
File	Edit	Analysis
Jul 7, 1999	3:16 PM	
Calculation of Effect Sizes - Hedges' d		
Study	Effect	Variance

Study 1	0.0362	0.2858
Study 2	0.7289	0.3047
Study 3	0.5651	0.3466
Study 4	1.5329	0.5175
Study 5	2.0139	0.4306
Study 6	1.8799	0.4806
Study 7	- No effect size could be calculated	
Study 8	1.3996	0.8299
Study 9	1.0889	0.7655
... Data Cut to Save Space ...		
Study 41	0.8199	0.5420
Study 42	-1.9561	0.7392
Study 43	-0.3989	0.4080

Figure 5.6. Sample output from effect size calculations, showing the date, the time, the type of effect size calculated, and the table of results (data are a subset of the Gurevitch and Hedges, 1993 data set: see Appendix I).

In circumstances where particular studies contain missing or invalid data for the effect size chosen (e.g. $N \leq 3$ for Fisher's z -transform, or negative study variances for Hedges' d) MetaWin will not calculate effects for those studies. Those studies will contain blanks in their effect size and variance columns of the spreadsheet, and a list of studies whose effect sizes could not be calculated will appear in the output window. In addition, a Warning message will be displayed informing the user that there was a problem calculating effects for some studies (for a list of Warning and Error messages see Appendix II). MetaWin will calculate effect sizes and variances for all other studies whose data are valid for the effect size chosen.

5.2 Calculating Meta-Analysis Summary Analyses

Once effect sizes have been calculated for each study, you may perform a summary analyses of a meta-analysis. Although MetaWin calculates six different effect sizes from three types of primary data, you may wish to use other effect sizes in your summary analysis (for a discussion of other effect sizes see Section 2.2: pp. xx). To do this, simply load a data file with a column of effect sizes and a column of variances that you have

generated. The procedures for performing the summary analysis are then identical to those described below.

The summary analyses window is accessed by selecting Analysis | Summary Analyses from the MetaWin menu, or by clicking the  button. There are six tabs on the summary analyses window, which contain all of the options for performing the various summary analyses. These tabs are: Basic Options, Resampling Tests, Refine Categories, Refine Studies, Graphical Output, and Miscellaneous Analyses. Results from summary analyses appear in the MetaWin output window (see “Results from Summary Analyses Calculations” below). Note that MetaWin will not activate the RUN button until **ALL** necessary data options have been selected! Below we discuss the options found on each of the tabs individually, and discuss how you can use MetaWin to perform summary analyses for different meta-analytic models (see Section 2.3: **pp. xx**).

Basic Options Tab

The basic options tab is where you select the columns containing the effect size and variance for each study, as well as where you choose the type of statistical model you wish to use in your summary analyses (Figure 5.7). The column in the data spreadsheet containing effect sizes is selected in the Effect Size list box, and the column containing the variance of that effect is selected in the Variance list box. Note that MetaWin only displays *numeric* columns as possible effect sizes and variances. If the effect size you have chosen is a log-transformed effect size (e.g., $\ln R$ or $\ln RR$), you may check the unlog effect size box and MetaWin will display the cumulative effect size results in their unlogged form in the output window. Any studies with missing values or invalid data (e.g., negative variances) will automatically be excluded from the analysis; excluded studies will be listed in the output window (see below). **Note:** To perform an “unweighted” summary analysis create a column of 1’s in the data spreadsheet and select this column in the Variance list box.

A set of six radio buttons arranged in a two x three table is also found on the MetaWin Basic Options tab. These buttons are for selecting the statistical model you wish to use in your summary analysis. The rows in this table correspond to selecting the type of data structure (no structure, categorical structure, or continuous structure: see Section 2.3: **pp. xx**), and the columns of this table allow you to choose either a fixed-effects model or a random-effects model (Section 2.3: **pp. xx**). Note that selecting a random-effects categorical model is identical to the mixed-effects model (Gurevitch and Hedges, 1993). If you select either a categorical analysis or a continuous analysis you must use the corresponding list box to select the column in the data sheet containing the model variable. MetaWin only displays *text* columns as potential grouping variables for categorical analyses, and *numeric* columns as possible independent variables for

continuous analyses. The label column cannot be used as a categorical variable. If a data column contains group information as a series of ordered numbers (e.g., 1,2,3, etc.), this column will be recognized as *numeric*. To use it as a categorical variable, you must exit from the summary analysis window, and change the column format from *numeric* to *text* in the data spreadsheet (see Section 4.1: pp. xx). For categorical analyses, MetaWin determines the number of groups based on the number of distinct values found in the selected data column. If case sensitivity was activated from the MetaWin options box (see Section 4.3), MetaWin will treat names that differ *only* by the case of one or more letters (e.g., “new york” and “New York”) as separate groups.

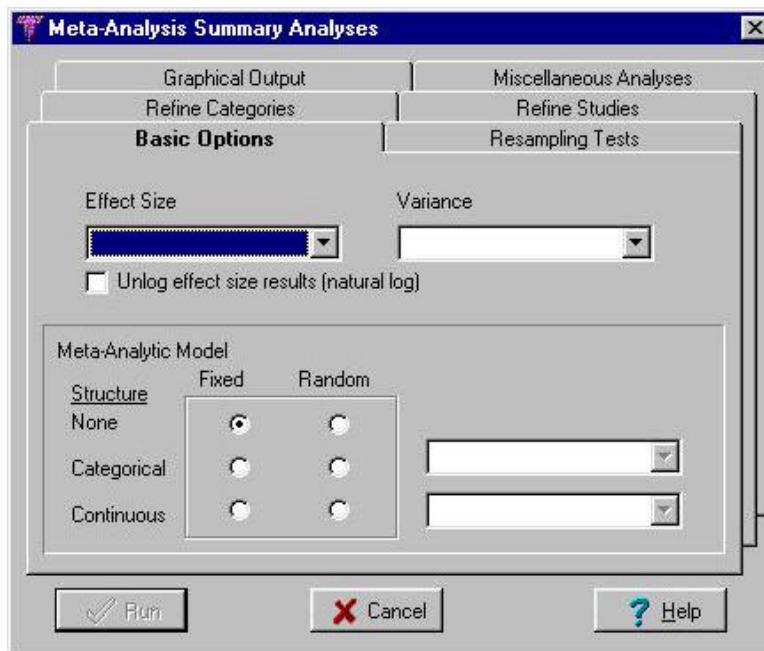


Figure 5.7. Basic Options tab of the summary analyses window.

Resampling Tests Tab

The Resampling Tests tab is where you choose the options for performing resampling statistics as companions to parametric estimates of significance. On this tab you may select whether or not MetaWin should perform resampling tests (default is to perform), which resampling tests you would like to perform, and how many iterations you want the test to be based on (Figure 5.8). If resampling tests are selected, bootstrap confidence intervals can always be performed around the cumulative mean effect size (see Section 3.3: pp. xx). Other resampling options depend on the data structure. For example, if a categorical analysis was chosen, you may perform randomization tests to assess the amount of heterogeneity between groups (Q_M). For this analysis, the bootstrap option will also calculate confidence intervals around group cumulative mean effect sizes. If you have chosen a continuous analysis, the significance of the slope and

intercept may be determined through randomization tests. For all resampling tests, you may choose how many iterations should be performed (19 to 64999). The significance levels from randomization are expressed as exact probability values based on the number of iterations selected (for details see Section 3.3).

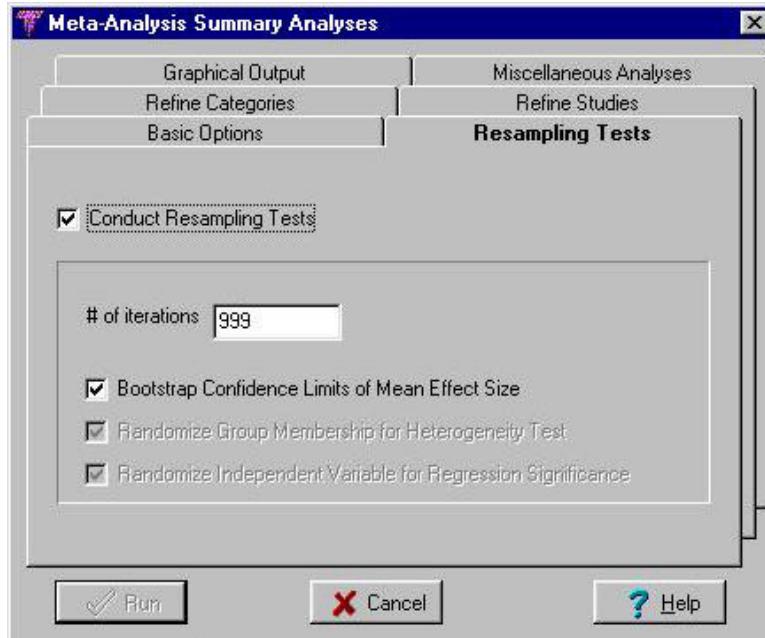


Figure 5.8. Resampling Tests tab of the summary analyses window.

Refine Categories Tab

The Refine Categories tab allows you to remove certain groups of studies from your summary analysis (Figure 5.9). This tab lists the groups in each category found in the data spreadsheet. Only *text* columns are displayed as possible categories (note that if a label column has been designated, it will *NOT* appear in the Refine Categories tab). By default, all groups in all categories are used in the summary analysis, and are shown as selected by having an “x” in their check box. Clicking on the box beside a particular group removes the “x” from that group’s checkbox. Any studies belonging to a deselected group will not be used in the summary analysis. In this way, you may customize your summary analysis to include only those studies of interest. This allows you to easily run analyses on a subset of your data. Note that if a categorical analysis was chosen and only one group has been selected, the summary analyses are performed as if there was *no* model structure (i.e., a one group analysis).

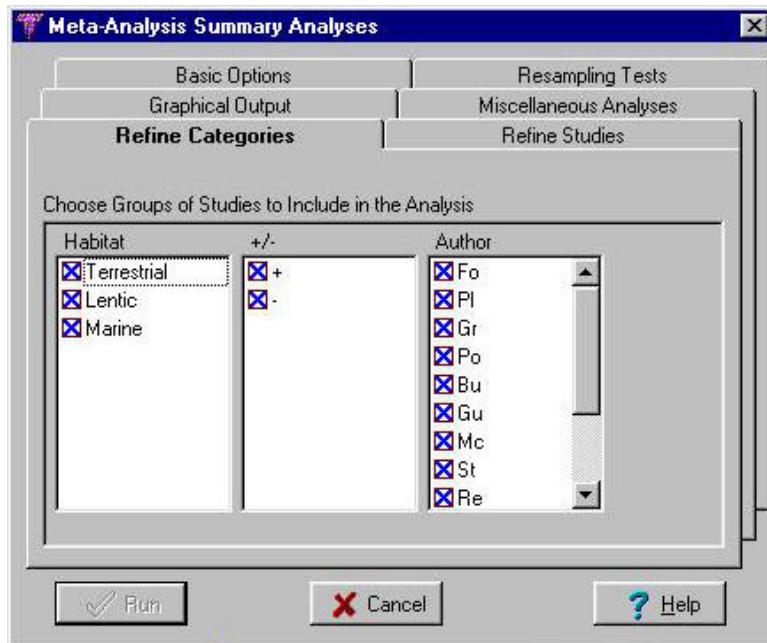


Figure 5.9. Refine Categories tab of the summary analyses window (data are a subset of the Gurevitch and Hedges, 1993 data set: see Appendix I).

Refine Studies Tab

The Refine Studies tab functions in a similar manner to the Refine Categories tab. From this tab you may add or remove individual studies from the summary analysis by clicking on that study's checkbox. This way, studies that are not of interest for a particular summary analysis may be removed, or studies that have been found to be statistical outliers may be eliminated from the analysis. In the Refine Studies tab, studies are listed according to their order in the data spreadsheet, and are represented as: Study 1, Study 2, etc. If a label column has been designated in the data spreadsheet, the names in this column will be used as study names in the Refine Studies tab (Figure 5.10). Clicking on the "Select All" button will include all studies in the summary analysis, and clicking on the "Deselect All" button will remove all studies from the analysis. Note that if you deselect all studies, the Run button will be deactivated. This option is useful only if you have many studies and only wish to select a few.

To perform a categorical analysis at least two studies must be selected from each selected group. If you deselect too many studies from a particular group, the summary analysis will be performed without that group. In this case, a warning message will appear informing you that one or more groups did not contain a sufficient number of studies. This information will also appear in the output window (see "Results from Summary Analyses Calculations" below).

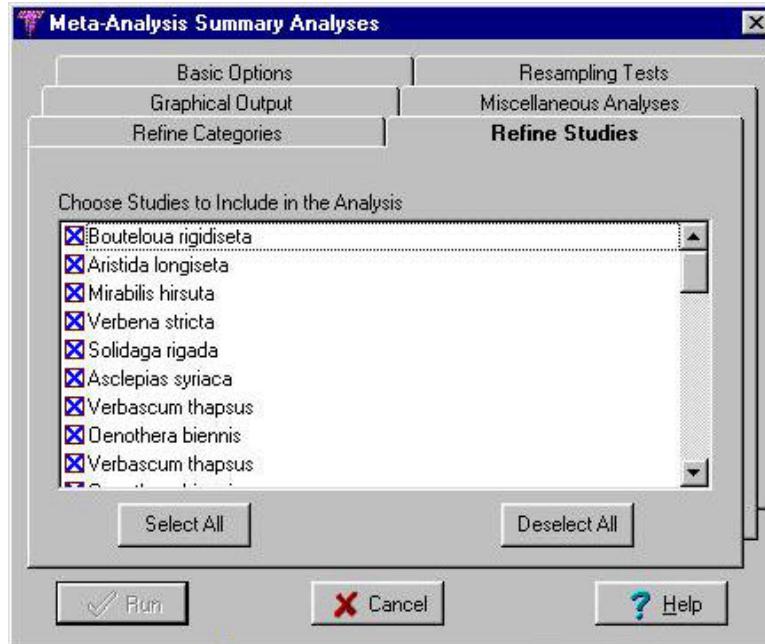


Figure 5.10. Refine Categories tab of the summary analyses window (data are a subset of the Gurevitch and Hedges, 1993 data set: see Appendix I).

Graphical Output Tab

The graphical Output tab allows you to choose to generate summary graphical presentation that corresponds to the type of summary analysis you have selected (Figure 5.11). There are three possible graphical summaries that can be generated: a plot of group cumulative mean effect sizes and their confidence intervals from a categorical analysis, a regression plot from a continuous analysis, and a plot of the results from a cumulative analysis. Note that you cannot select a graphical output before choosing the type of summary analysis you wish to perform: the options on this tab are only activated *after* you have chosen the summary analysis model. A brief description of each type of summary graphical output is found in the Section “Results from Summary Analyses Calculations” below.

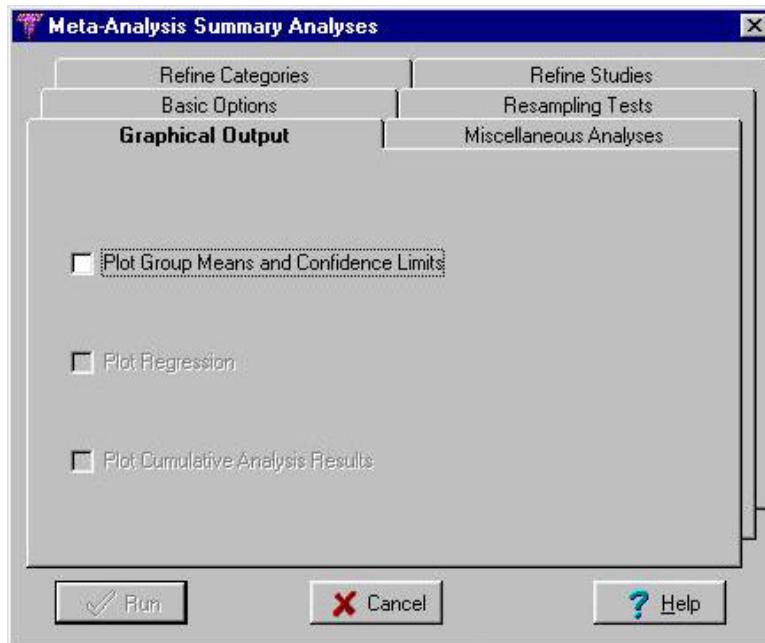


Figure 5.11. Graphical Output tab of the summary analyses window (in this figure, only the plot of group means is active).

Miscellaneous Analyses Tab

The final tab on the summary analyses window is the Miscellaneous Analyses tab. On this tab you can select whether or not you wish to perform a cumulative analysis, and whether or not you wish to perform fail-safe calculations (Figure 5.12). A cumulative analysis is a summary meta-analysis in which the individual studies are sequentially added to the pool of studies, and the summary statistics are recalculated (see Section 3.2). If a cumulative analysis is selected, you must determine how the studies should be ordered by selecting the appropriate column in the “Sort Cumulative Analysis by” list box. The default sorting order is the order of the studies in the data spreadsheet, but you may select any *numeric* column from the data spreadsheet as the sorting order (e.g., a column of publication date is often used). You may also choose to generate a graphical representation of the cumulative analysis output by selecting “Plot Cumulative Analysis Results” on the Graphical Output tab (see above). **Note:** Cumulative analyses may only be performed on summary models with no data structure or continuous data structure: the cumulative analysis routine will not work with categorical data.

You may also choose to perform fail-safe analyses methods from the Miscellaneous Analyses tab. Click on the Perform Fail-Safe Analyses checkbox to select these methods. For these analyses, you may also alter the alpha level (Rosenthal’s method) and the negligible effect (Orwin’s method) by changing the default values found in the

appropriate boxes (for a discussion of these statistics see Section 3.1: pp. xx). Note that results from fail-safe methods performed from the Miscellaneous Analyses tab may not be numerically identical to those calculated from the data exploration window (see Section 6.1). This is because the fail-safe calculations from the data exploration window are performed on *all* studies, while the fail-safe calculations from the Miscellaneous Analyses tab may be performed on a subset of studies (if studies were removed under the refine categories or refine studies tabs).

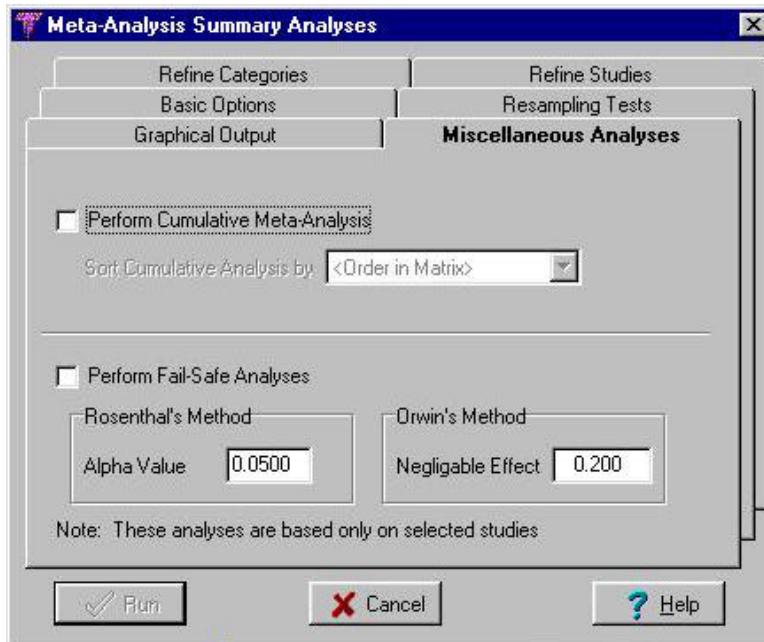


Figure 5.12. Miscellaneous Analyses tab of the summary analyses window.

Results from Summary Analyses Calculations

Once you have selected the appropriate data type and effect size for your data and have filled in the necessary list boxes, the RUN button is activated and you can calculate summary analysis statistics for your data. MetaWin saves all of the options you selected on the summary analyses window so that an identical analysis may be easily repeated. This is useful if you are running several similar analyses one after the other. If you change anything in the spreadsheet, however, all options will be cleared.

Results from summary analyses appear in the output window of MetaWin. There are two sections to the summary analyses output: a listing of the options that were selected for the summary analysis, followed by the statistical results. The list of options for the analysis includes the date and time of the analysis, the name of column containing the effect size and the name of the column containing the variance of the effect size. The type of statistical model is recorded, followed by any other options selected. The

statistical model line states whether the analysis is a fixed-effects model or random-effects model, and also states the type of data structure (i.e., categorical data or continuous data), if any. The other options selected for the analysis (e.g., resampling tests or cumulative analysis) are listed below the statistical model line. If resampling tests are selected, the number of iterations is stated. Finally, any individual studies or groups excluded from the analysis (through the Refine Categories or Refine Studies tabs) are listed below the other statistical options (as in Figure 5.14). Several examples of the summary analysis options section of the output are found in Figure 5.13.

```

MetaWin - [Output]
File Edit Analysis Windows Help
...FIXED-EFFECTS NO STRUCTURE MODEL...
Jul 7, 1999 11:40 AM
Meta-Analysis Results
Effect Size from column "d"      Variance from column "Var(d)"
Fixed effects model

...RANDOM-EFFECTS NO STRUCTURE MODEL...
Jul 7, 1999 11:40 AM
Meta-Analysis Results
Effect Size from column "d"      Variance from column "Var(d)"
Random effects model

...FIXED-EFFECTS CATEGORICAL MODEL...
Jul 7, 1999 11:40 AM
Meta-Analysis Results
Effect Size from column "d"      Variance from column "Var(d)"
Fixed effects model with a grouping variable (Habitat)

...RANDOM-EFFECTS CONTINUOUS MODEL WITH RESAMPLING TESTS...
Jul 7, 1999 11:40 AM
Meta-Analysis Results
Effect Size from column "d"      Variance from column "Var(d)"
Random effects model with a continuous variable (Nc)
Resampling tests generated from 999 iterations

```

Figure 5.13. MetaWin output window displaying options statements from several different summary analyses (text not generated by MetaWin is delineated by "...").

Below the options summary section of the output is the statistical results section. This section contains the results from all summary analyses performed on the data, and depends, in part, on which analysis was selected. Regardless of statistical model, several summary statistics are always reported. The total heterogeneity, Q_T , of the sample is always reported, and tested against a χ^2 -distribution (see Section 2.3: pp. xx).

The cumulative mean effect size, \bar{E} , for the entire sample is calculated, and its 95% confidence intervals (or whatever percentile was chosen in the Options Dialog Box, see Section 4.3) are reported. If resampling statistics were performed, the bootstrap

confidence intervals and bias-corrected bootstrap confidence intervals are also reported (see Section 3.3). The square root of the estimate of the pooled variance (see Section 2.3, **pp. xx**), the mean study variance, and the ratio between these two values is also reported. This ratio is an informal comparison of the amount of variance due to between-study differences versus the amount of variance due to sampling error within studies (Hedges et al., 1999). If the ratio is very large (i.e., the between-study variance is many times larger than the within-study variance), combining the studies in a summary analysis may not be warranted. In random-effects models, the pooled variance estimate is also reported. If this estimate is less than or equal to zero, the random-effects model does not apply and a fixed-effects model is calculated instead. The calculation of the pooled variance estimate is dependent on model structure (see Section 2.3, **pp. xx**).

For a fixed-effects model or random-effects model with no data structure (the simplest analyses), these are the only statistics that are calculated and reported. An example of the results from a fixed-effects model with no data structure is presented in Figure 5.14.

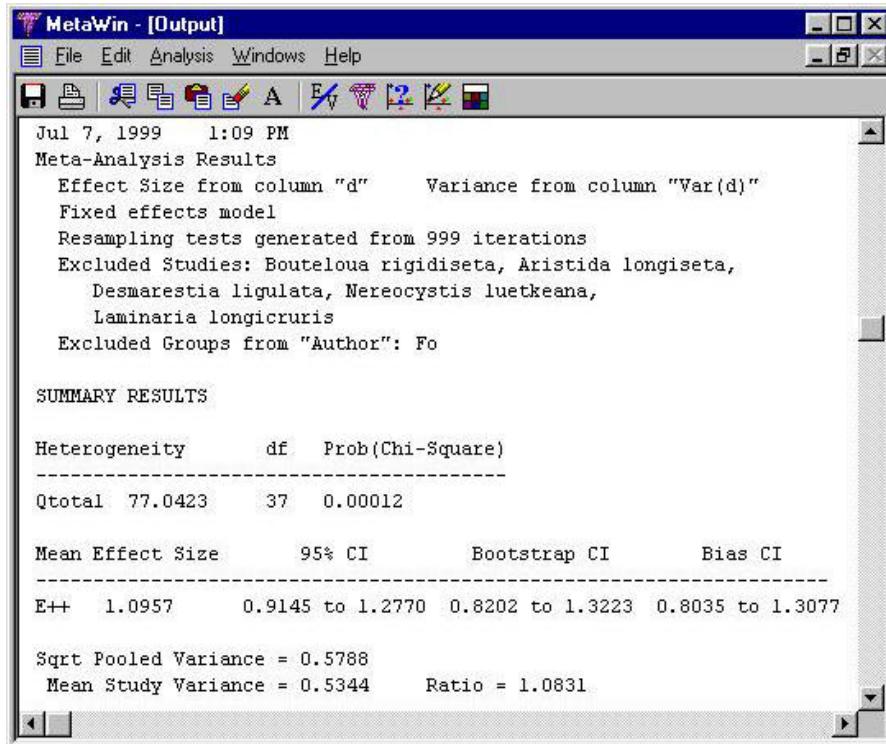
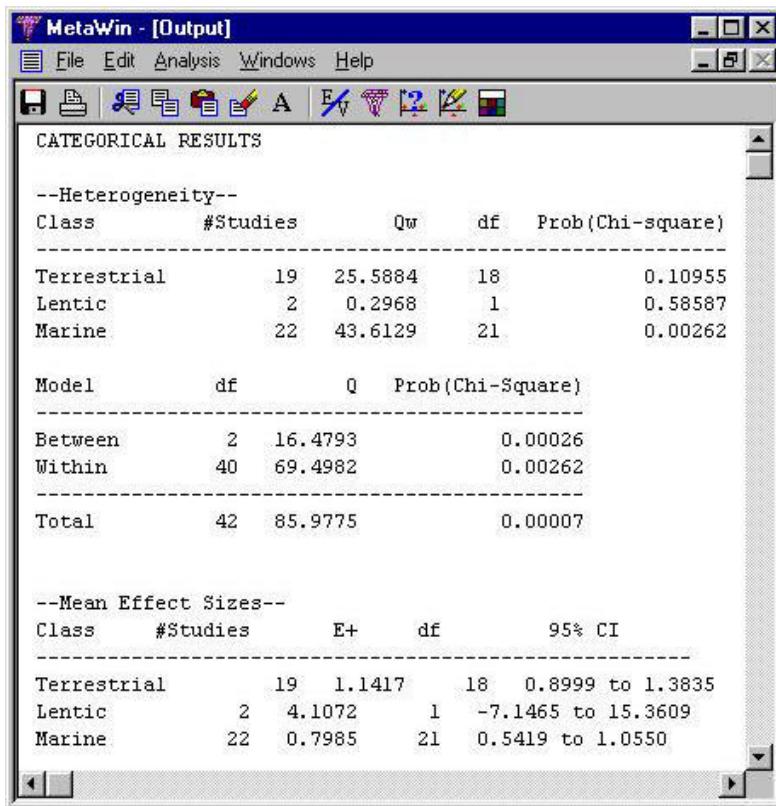


Figure 5.14. Results from a fixed-effects summary analysis with no data structure and resampling tests. Several studies and groups of studies have been eliminated from the analysis to demonstrate the output from a refined data analysis (data are from Gurevitch and Hedges, 1993: see Appendix I).

Results from a summary analysis of categorical data are slightly more complicated. Like the summary analysis of one-sampled data, the total heterogeneity, Q_T , and

cumulative mean effect size, \bar{E} , are calculated. For a categorical model however, the total heterogeneity can be partitioned into the variation in effect sizes explained by the categorical model, Q_M , and the residual error variation in effect sizes, Q_E . Both of these are tested against a χ^2 -distribution, and are presented in a heterogeneity table analogous to an ANOVA table. In addition, the cumulative mean effect sizes for each group of studies is calculated, and presented with their confidence intervals. Finally, for a fixed-effects categorical model, a table containing the within-group heterogeneity, Q_{Wj} , for each group is presented. If a random-effects categorical model is calculated, this table is replaced by a table containing the pooled variance for each group (for a discussion of these statistics see Section 2.3: pp. xx). An example of the results from a fixed-effects categorical model is presented in Figure 5.15.



The screenshot shows the MetaWin software interface with the title bar "MetaWin - [Output]". The menu bar includes File, Edit, Analysis, Windows, and Help. The toolbar contains icons for opening files, saving, printing, and other functions. The main window displays the "CATEGORICAL RESULTS" output. The output is divided into sections:

- Heterogeneity--**

Class	#Studies	Qw	df	Prob(Chi-square)
Terrestrial	19	25.5884	18	0.10955
Lentic	2	0.2968	1	0.58587
Marine	22	43.6129	21	0.00262

- Model**

Model	df	Q	Prob(Chi-Square)
Between	2	16.4793	0.00026
Within	40	69.4982	0.00262

- Total**

Total	df	Q	Prob(Chi-Square)
Total	42	85.9775	0.00007

- Mean Effect Sizes--**

Class	#Studies	E+	df	95% CI
Terrestrial	19	1.1417	18	0.8999 to 1.3835
Lentic	2	4.1072	1	-7.1465 to 15.3609
Marine	22	0.7985	21	0.5419 to 1.0550

Figure 5.15. Results from a fixed-effects categorical summary analysis (data are from Gurevitch and Hedges, 1993: see Appendix I).

If graphical output from a categorical analysis has been selected, a plot of the group means and their confidence intervals will be generated (Figure 5.16). A complete discussion of MetaWin graphical output is found in Section 6.2: pp. xx).

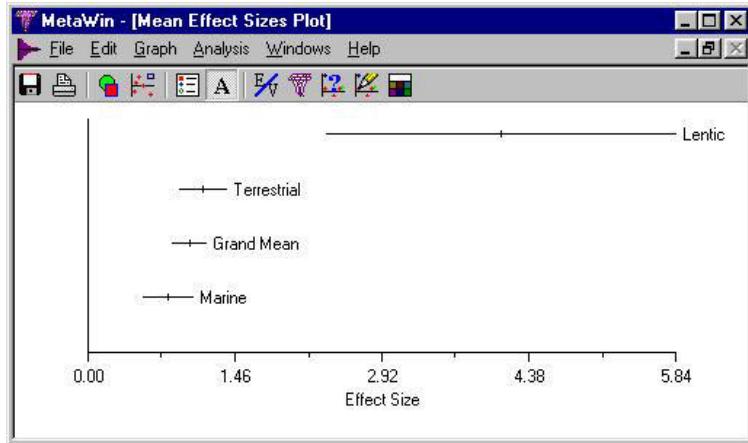


Figure 5.16. Graphical output from a fixed-effects categorical summary analysis (data are from Gurevitch and Hedges, 1993: see Appendix I).

Results from a summary analysis of continuous data are similar to those from a summary analysis of categorical data. For continuous data MetaWin calculates the cumulative mean effect size, \bar{E} , and the total heterogeneity, Q_T , and partitions the total heterogeneity into the variation in effect sizes explained by the regression model, Q_M , and the residual error variation in effect sizes, Q_E . The intercept and slope of the regression are calculated and tested against a normal distribution (see Secton 2.3: **pp. xx**). If graphical output from the continuous analysis is selected, a regression plot is also presented. The same summary statistics are calculated for both a fixed-effects and random-effects continuous model. An example of results from a random-effects continuous model is presented in Figures 5.17 and 5.18.

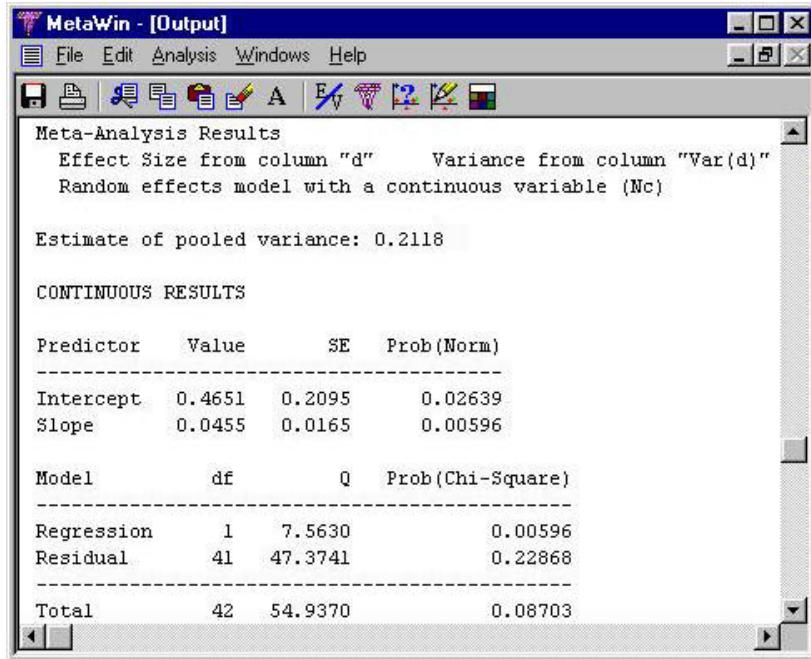


Figure 5.17. Results from a random-effects continuous summary analysis (data are from Gurevitch and Hedges, 1993: see Appendix I).

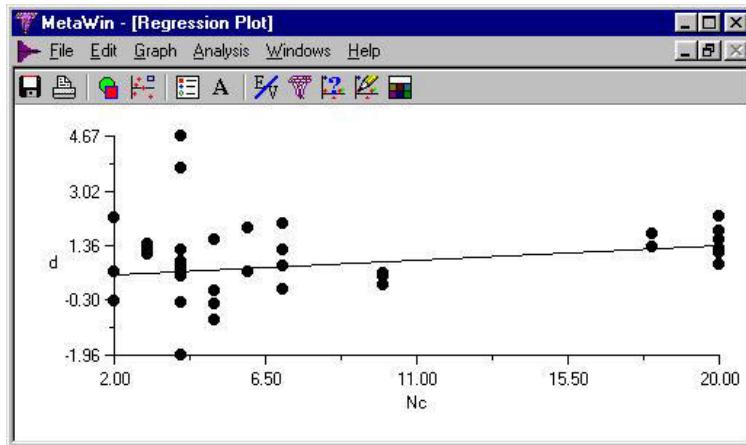


Figure 5.18. Graphical output from a fixed-effects continuous summary analysis (data are from Gurevitch and Hedges, 1993: see Appendix I).

For any summary analysis, if fail-safe calculations were selected, results from both Rosenthal's method and Orwin's method will be presented at the bottom of the results from the summary analysis (for more information see Section 3.1, pp. xx). Results from a cumulative summary analysis are presented somewhat differently from the analyses discussed above. Because cumulative summary analyses are really a series of sequential analyses, the results from this analysis are repetitious. For each step in the cumulative analysis, MetaWin reports the summary statistics, including the total

heterogeneity and the cumulative mean effect size for the sample. If graphical output was selected, MetaWin also generates a plot of the cumulative mean effect size and its confidence intervals from each step in the analysis (e.g., Figure 5.19). These are ordered from first analysis (bottom of graph) to last analysis (top of graph), and usually converge (for a discussion see Section 3.2: pp. xx).

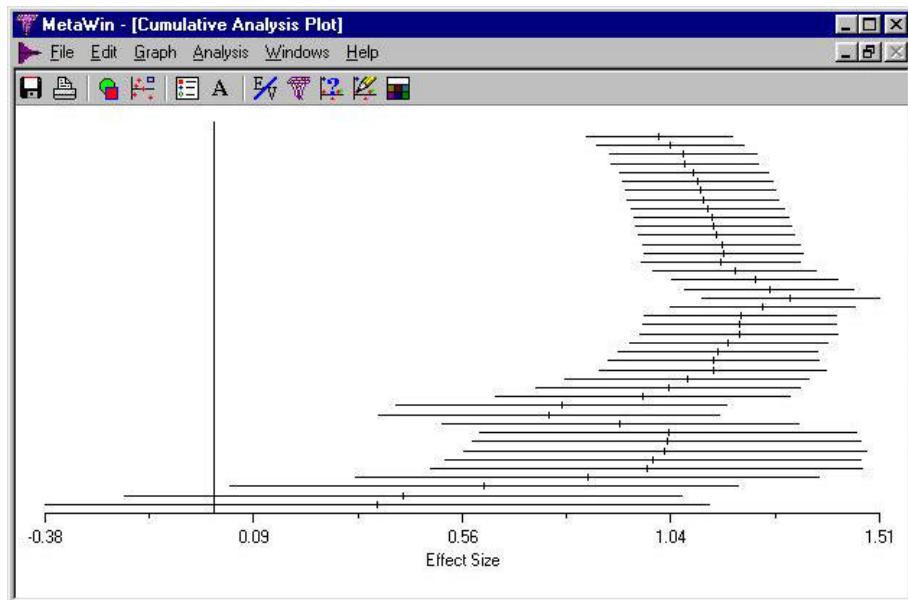


Figure 5.19. Graphical output from a cumulative summary analysis (data are from Gurevitch and Hedges, 1993: see Appendix I).

6. MISCELLANEOUS ANALYSES/OPTIONS

In the previous chapter we showed how MetaWin can be used to generate effect sizes and sampling variances for a set of studies, and how these can be statistically combined in a summary analysis. In this chapter, we discuss a set of “miscellaneous analyses” that are designed to supplement your summary meta-analysis and help you assess and understand the results you obtain from your data. In Section 6.1, we discuss methods for data exploration, including fail-safe calculations, methods for detecting publication bias, and various graphical data exploration techniques. These methods can be accessed through the Data Exploration window. In Section 6.2, we describe the Draw Graph window, and discuss the graphical techniques available in MetaWin that can be accessed from that window.

6.1 Data Exploration

The data exploration window is accessed by selecting Analysis | Data Exploration from the MetaWin menu, or by clicking the  button. There are four tabs on the data exploration window: Data, Graphical Methods, Publication Bias, and Fail-Safe. When the data exploration window is first opened, only the Data tab is active (Figure 6.1). On the Data tab you can select the columns containing the effect size, the sample size, and the variance for the data exploration methods. All of the methods require an effect size. Some of the graphical methods require variance, others sample size. The Publication Bias methods require variance; sample size is optional. The Fail-Safe methods require variance. The RUN button is not active until all necessary options have been selected.

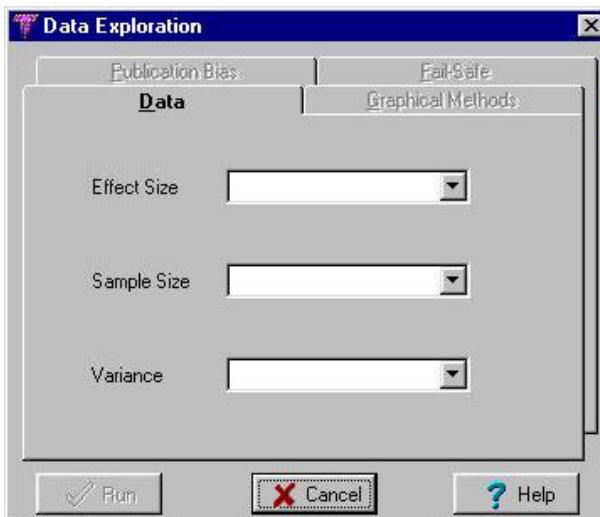


Figure 6.1. Data tab of the data exploration window.

Graphical Methods Tab

The Graphical Methods tab is where you select the options for generating exploratory data graphs (Figure 6.2). Three different graphical methods are available: funnel plots, weighted histograms, and normal quantile plots (for a description of these techniques, see Section 3.1, pp. xx). All three graphs may be generated if you specified a variance data column, but only funnel plots can be generated from sample size data. If you specified both a sample size *and* a variance, you may choose which to use for the funnel plot. A list box is available for selecting a categorical grouping variable: only *text* columns are displayed in this box (see Section 6.2 for more information on grouping variable options for graphics). If selected, groups will be displayed in all graphs (see “Results from Data Exploration” below).

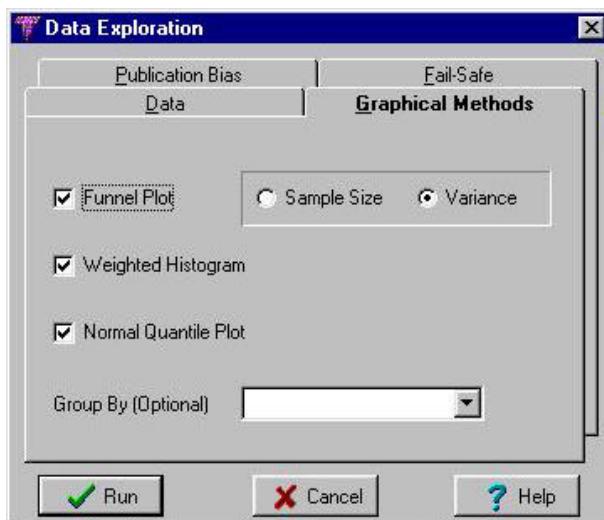


Figure 6.2. Graphical Methods tab of the data exploration window.

Publication Bias Tab

The Publication Bias tab (Figure 6.3) is where one selects the options for performing rank-correlation methods for detecting publication bias (Begg, 1994: for a description of these statistics, see Section 3.1, pp. xx). One can choose to perform rank correlation methods using either Kendall’s Tau or Spearman’s rank-order method or both methods. This test calculates the rank correlation between a standardized effect size and another variable. One can choose a standardized variance or sample size for the second variable in the box at the top. These methods require an effect size and a variance to be specified on the Data tab. If you wish to calculate the correlation with sample size, this third variable must also be specified on the Data tab.

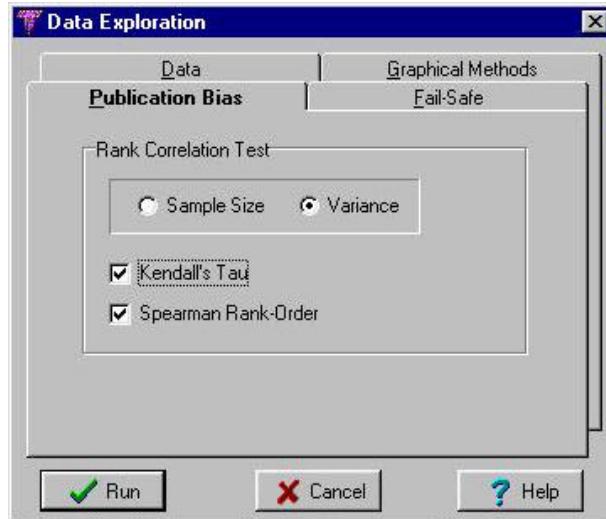


Figure 6.3. Publication bias tab of the data exploration window.

Fail-Safe Tab

On the Fail-Safe tab (Figure 6.4) you will find the options for performing Rosenthal's and Orwin's fail-safe calculations. You may choose the desired α level for Rosenthal's method and the negligible effect value for Orwin's method. A more complete description of these statistics is found in Section 3.1 (pp. xx). These methods require an effect size and variance to be specified in the Data tab.

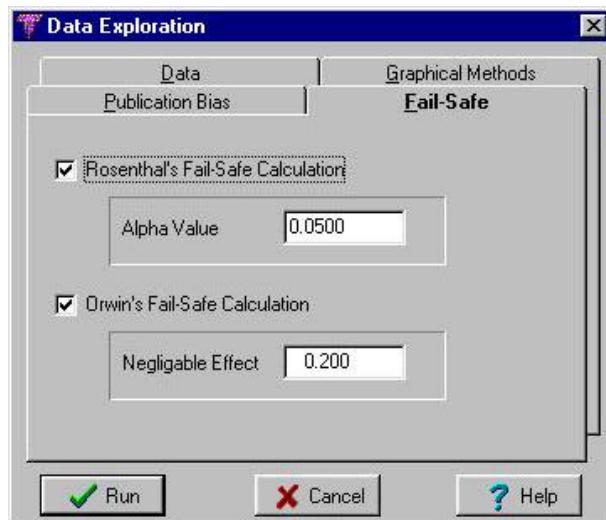


Figure 6.4. Fail-Safe tab of the data exploration window.

Results from Data Exploration

Once you have selected the appropriate data and options from the tabs on the data exploration window, the RUN button is activated and you can perform the exploratory analyses. Numerical results from the exploratory analyses appear in the output window of MetaWin (Figure 6.5). Any studies with missing or invalid values will automatically be excluded from the analysis and listed in the output window. For both the rank correlation methods, the correlation and its probability value are presented. A significant rank correlation may indicate a publication bias where larger effects in one direction are more likely to be published than smaller effect sizes (see Section 3.1, pp. xx). If fail-safe numbers were calculated, the fail-safe numbers are listed. Fail-safe numbers are an indication of the number of nonsignificant unpublished studies that, if added to the analysis, would change the summary results from significant to non-significant. Because large fail-safe values indicate that many unpublished studies are required, one may be more confident in summary results associated with large fail-safe values (for a discussion see Rosenthal, 1979; and Section 3.1, pp. xx). If selected, plots of exploratory graphical methods are generated in separate windows (Figures 6.6 – 6.8). A complete discussion of MetaWin graphical output is found in the next Section.

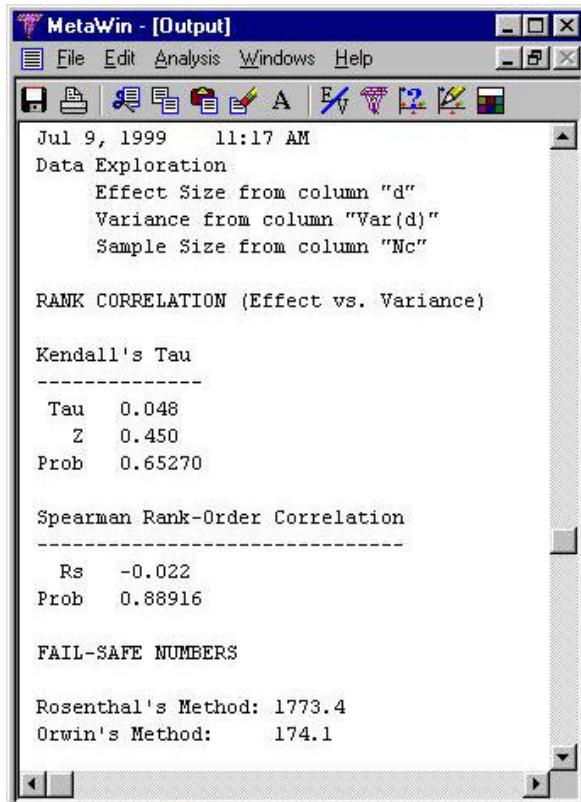


Figure 6.5. MetaWin output window displaying the results from exploratory data analyses using data from Gurevitch and Hedges, 1993 (see Appendix I: results from all methods are shown).

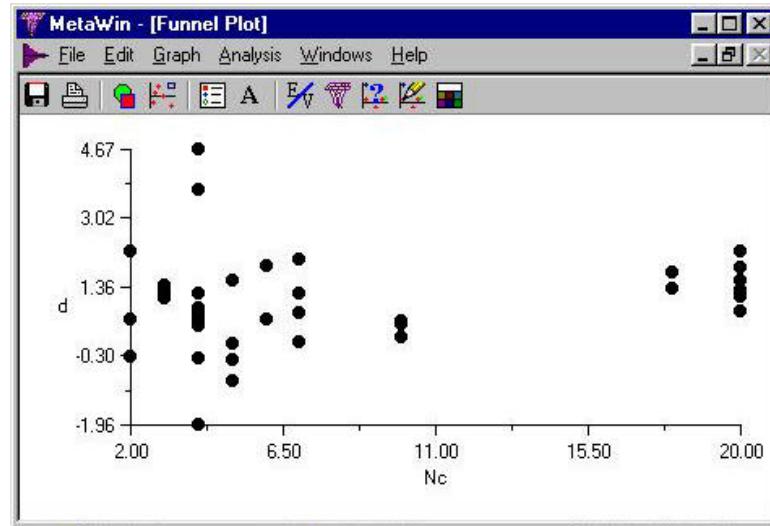


Figure 6.6. Funnel plot of effect size versus sample size (data from Gurevitch and Hedges, 1993: Appendix I).

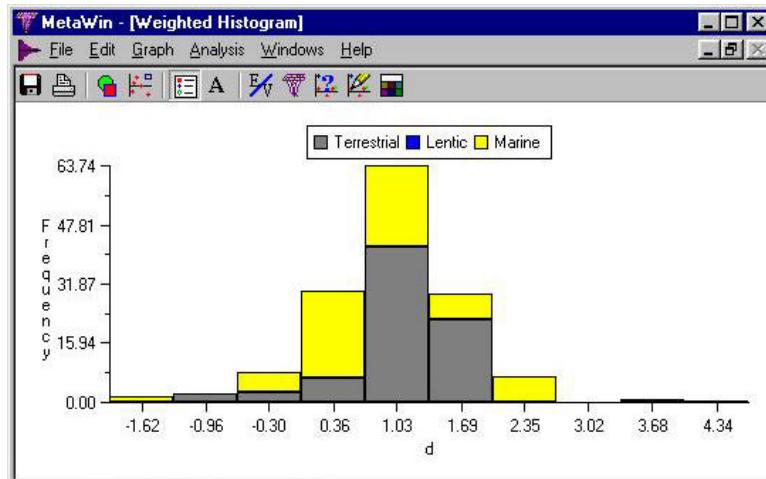


Figure 6.7. Weighted histogram with groups specified (data from Gurevitch and Hedges, 1993: Appendix I).

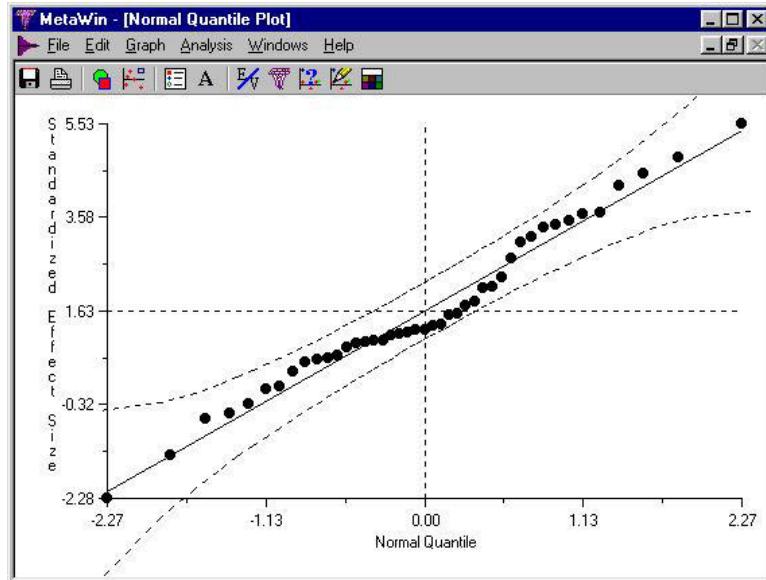


Figure 6.8. Normal quantile plot for Gurevitch and Hedges (1993) data (Appendix I).

6.2 Graphics

MetaWin 2.0 contains a number of graphical options. Some of these are associated with certain analyses (e.g., cumulative analysis plot) and can only be generated by performing that analysis. Others (e.g., scatter plot) can be accessed through the draw graph window, and are useful for visualizing your data. In this section, we discuss the different graphical methods available from the draw graph window. This is followed by a description of the general plot options that apply to all graphical output in MetaWin.

Draw Graph Window

The draw graph window is accessed by selecting **Analysis|Draw Graph** from the MetaWin menu, or by clicking the  button. There are two tabs on the draw graph window (Figure 6.9), the Graph Type tab and the Data Location tab. From the Graph Type tab, five different types of graphs that may be plotted: Scatter Plots, Radial Plots, Weighted Histograms, a plot of study Effect Sizes with confidence intervals, and Normal Quantile plots. Once a graph type has been selected, the Data Location tab is activated, and you may select the appropriate data for generating the graph. For all graph types, any studies with missing or invalid values will automatically be excluded from the analysis and listed in the output window. The options for each graph type are discussed below.

Group By: For every graph type the Group By option allows you to select a categorical variable specifying how to subdivide your data. The data will still be plotted on a single graph, but you have the option of using different symbols or colors to delineate the studies belonging to different groups. This can be very useful for detecting publication bias (Palmer, 1999; and Section 3.1, **pp. xx**).

Scatter Plot: If a scatter plot is selected, you may choose the columns for the X-variable and the Y-variable (Figure 6.9). Only *numeric* data columns are listed as possible variables. The scatter plot is a general plot, and can be used to visualize your data, or as one way to generate **funnel plots** (Figure 6.10: see Sections 3.1 and 6.1).

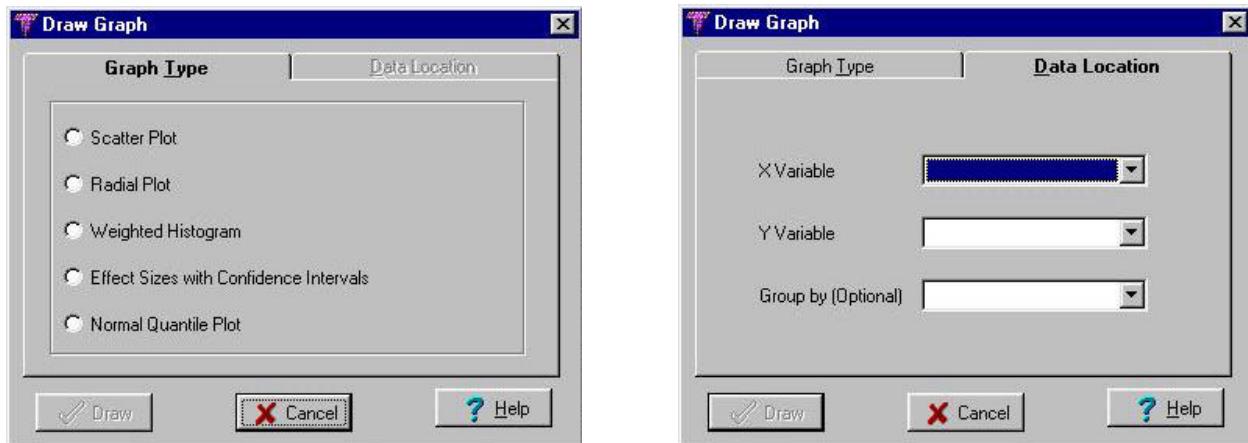


Figure 6.9. Graph Type tab of the draw graph window and the Data Location tab when a scatter plot is selected.

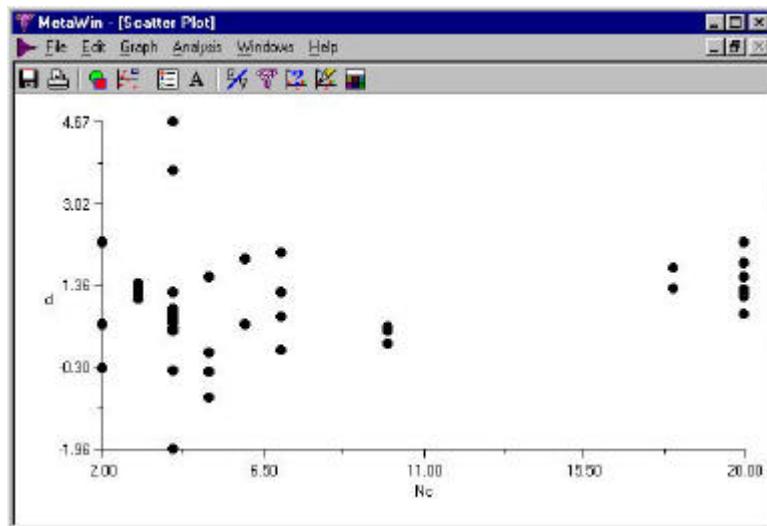


Figure 6.10. Example of a scatter plot of effect size versus sample size (data from Gurevitch and Hedges, 1993: Appendix I).

Radial Plot: A radial plot is similar to a scatter plot, but is used for the odds ratio effect size. On the Data Location tab (Figure 6.11) you select the column for the odds ratio and its variance. Only *numeric* data columns are listed as possible variables. The radial plot shows precision (square-root of the weight) on the X-axis versus a standardized effect size (effect * precision) on the Y-axis (Figure 6.12). There is a curve on the right side of the graph which has some interesting properties. Drawing a straight line from the origin through any point in the scatter intersects the curve at the OR for that study. Also, the regression through the origin for that scatter intersects the curve at the mean odds ratio. This graph is therefore a very succinct way of summarizing the results from a meta-analysis performed on $\ln OR$ (for a discussion see Galbraith, 1988).

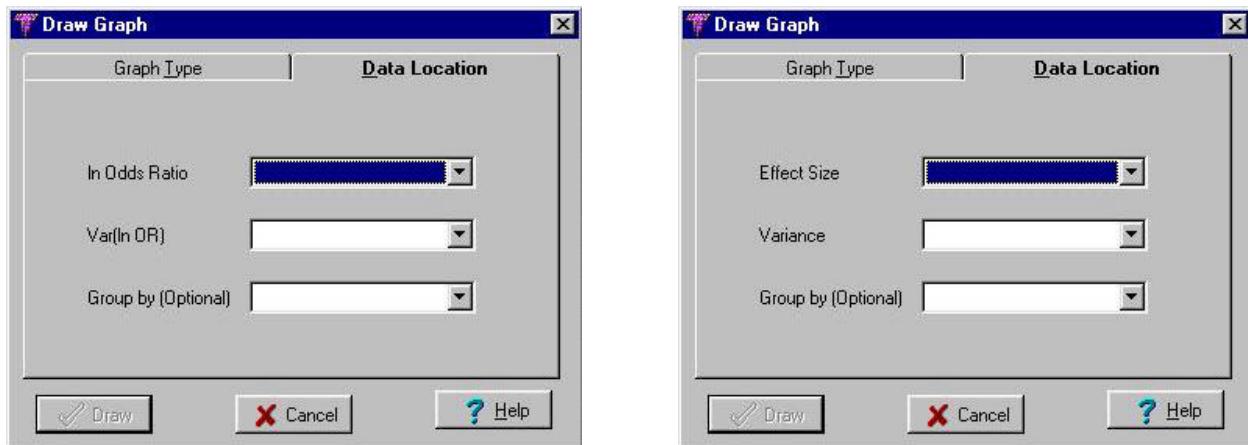


Figure 6.11. Data Location tab when a radial plot is selected (left), and Data Location tab when a weighted histogram, effect sizes and confidence intervals, or a scatter plot is selected (right).

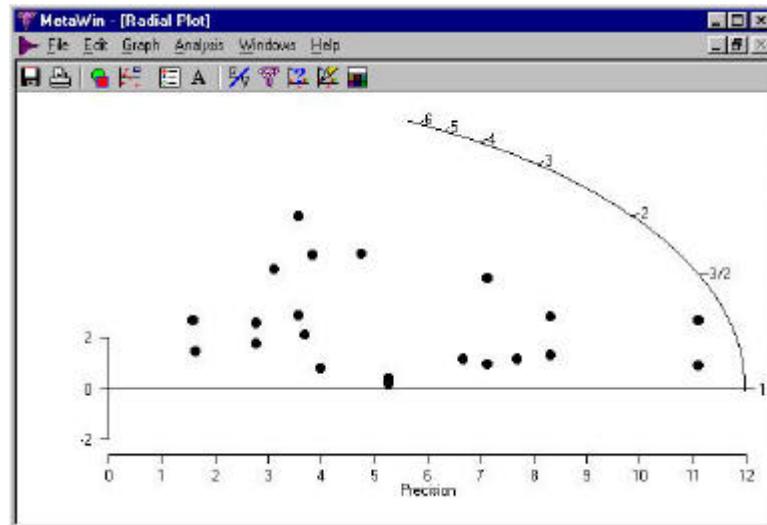


Figure 6.12. Example of a radial plot of standardized effect size versus precision (data from Berlin et al., 1989: Appendix I).

Weighted Histogram: If you select a weighted histogram, you may choose the columns for the effect size and the variance of the effect size from the list boxes on the Data Location tab (Figure 6.11). Only *numeric* data columns are listed as possible variables. A histogram is then plotted using a weighted frequency (weights = 1/variance) rather than a true frequency. Therefore, the weighted histogram shows the relative contribution of the data to each effect class, rather than the sample size for that effect class (Figure 6.13). If a conventional histogram is desired, simply create a column of 1's in the data spreadsheet and use this column as the variance. The initial number of histogram classes is automatically generated. You may choose a new number of classes by using the “# of Histogram Classes” option on the Graph menu.

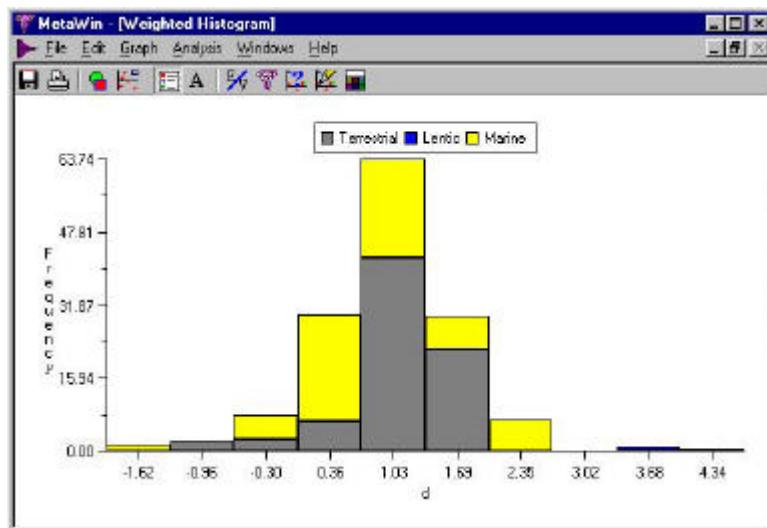


Figure 6.13. Example of weighted histogram with groups (data from Gurevitch and Hedges, 1993: Appendix I).

Effect Sizes and Confidence Intervals: If you choose an effect size graph, you may choose the columns for the effect size and the variance of the effect size from the list boxes on the Data Location tab (Figure 6.11). Only *numeric* data columns are listed as possible variables. This graph gives a general summary of the data in terms of effect sizes and their individual confidence intervals (Figure 6.14). The confidence interval width is based on the choice selected in the Options Dialog box (Section 4.3). The data in the plot may be ordered by their position in the data spreadsheet, or may be sorted by effect size or variance. The data may also be supersorted by group. Sorting options are found by selecting Graph | Sort Effects By (Figure 6.15).

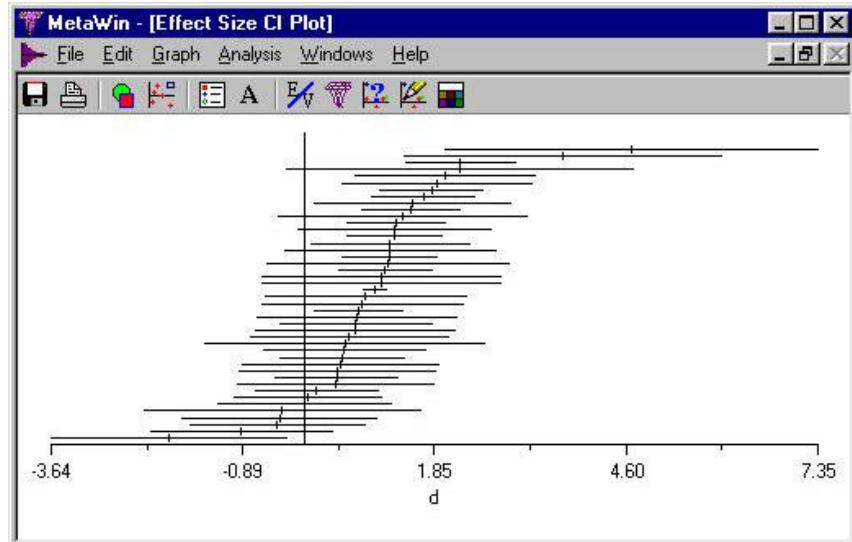


Figure 6.14. Example of effect sizes and 95% confidence intervals plot sorted by magnitude of effect size (data from Gurevitch and Hedges, 1993: Appendix I).

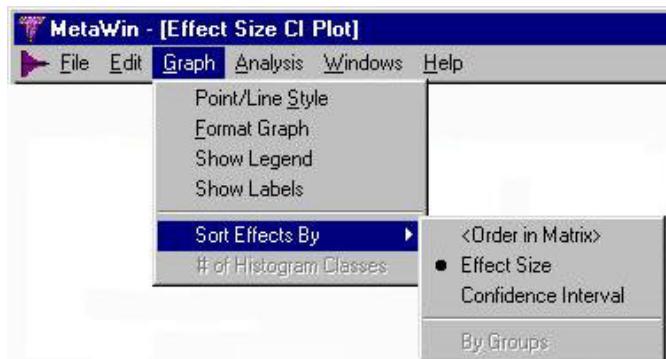


Figure 6.15. Sort Effects By options from the Graph menu.

Normal Quantile Plot: If you choose a normal quantile plot you may choose the columns for the effect size and the variance of the effect size from the list boxes on the Data Location tab (Figure 6.11). Only *numeric* data columns are listed as possible variables. In a normal quantile plot, the standardized effect size is plotted against the normal quantile values (Figure 6.16). If the distribution of effect sizes is similar to the distribution of normal quantiles, the data points will fall close to the line $X = Y$ (for more details, see Section 3.1, **pp. xx**).

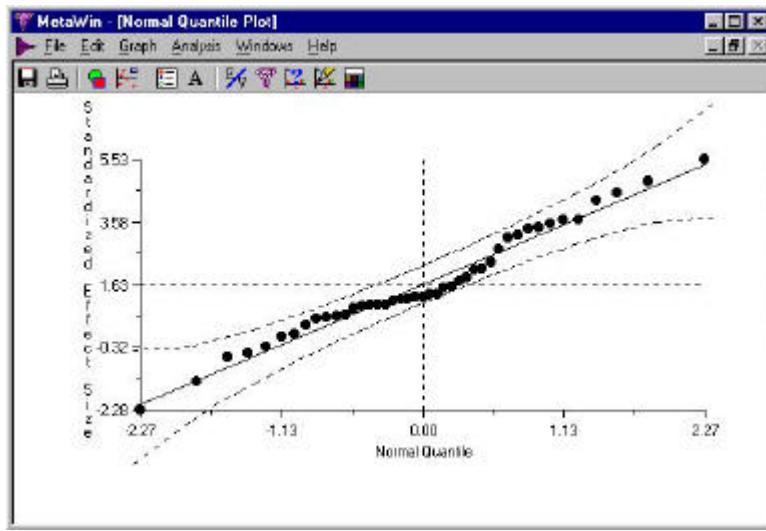


Figure 6.16. Example of a normal quantile plot of effect sizes versus normal quantiles (data from Gurevitch and Hedges, 1993: Appendix I).

The plots listed above can all be generated from the Draw Graph window. Below is a brief description of other plots that can be generated in MetaWin.

Funnel Plot: The funnel plot is a special case of the scatter plot. It can be created as part of the Data Exploration analysis (Section 6.1), or can be generated manually by drawing a scatter plot with the appropriate choice of X- and Y-variables.

Regression Plot: The regression plot can be generated from a summary analysis of continuous data (Section 5.2). It is a scatter plot with the regression line from the continuous analysis added. The Y-variable is effect size and the X-variable is the independent variable.

Group Cumulative Effect Sizes and Confidence Intervals: The group cumulative effect sizes and confidence interval plot is generated from a summary analysis of categorical data (Section 5.2). It is a special case of the effect sizes and confidence interval plot where group cumulative effect sizes are plotted rather than the effect sizes of individual studies.

Cumulative Analysis Plot: The cumulative analysis plot can be generated for a summary cumulative analysis (Section 5.2). It is a special case of the effect sizes and confidence interval plot where the results from successive cumulative effect sizes are plotted rather than the effect sizes of individual studies.

General Plot Options

All of the graphical methods in MetaWin have a set of general plot options. For example, all plots can be copied to the clipboard as either bitmaps or metafiles. All plots can be saved to a file by selecting File | Save from the menu. MetaWin plots can be saved as bitmap (*.bmp), Windows metafiles (*.emf; .wmf) or jpeg (.jpg) files. MetaWin plots can similarly be printed. All visual aspects of the plots may be changed. Options for changing visual aspects of the plots are found on the Graph Format window and the Point/Line Style window. These windows are accessed by selecting Graph | Format Graph or Graph | Point/Line Style from the menu (Figure 6.17), or by clicking on the  button or  button.

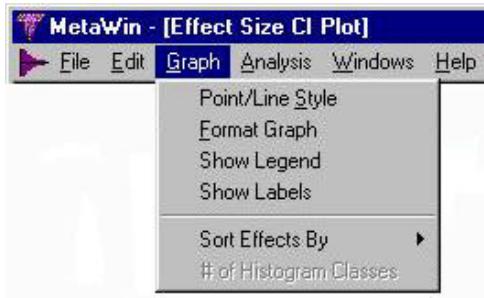


Figure 6.17. Commands found under the Graph menu.

Format Graph Window: On the Format Graph window (Figure 6.18) you will find all of the basic plot options for all types of MetaWin plots. This window allows you to customize any of the background elements of the plot. From this window you can manipulate options for the axes, the plot legend, the plot background, and data point labels. For both the X- and Y-axes, you can change the axis color, the axis style (e.g., solid, dash, dot, etc.), the width of the axis, the font of the axis label, the font of the axis numbers, the number of major and minor tick marks, the number of decimal places, and the minimum and maximum values of the axis. You can select whether or not to show the plot legend, change the font of the legend, and specify the legend location (top, left, right, or bottom). You can also choose to place a frame around the legend, and select its color and style. The background color of the plot may also be changed, and you can change the size of the X- and Y-margins (in pixels) for saving and printing plots. Finally, you can choose whether to show data point labels and can select the font of those labels.

Some of the formatting options discussed above may not apply to all graphs. For example, effect sizes and confidence interval plots do not have a meaningful Y-axis, so most of the Y-axis attributes are ignored.

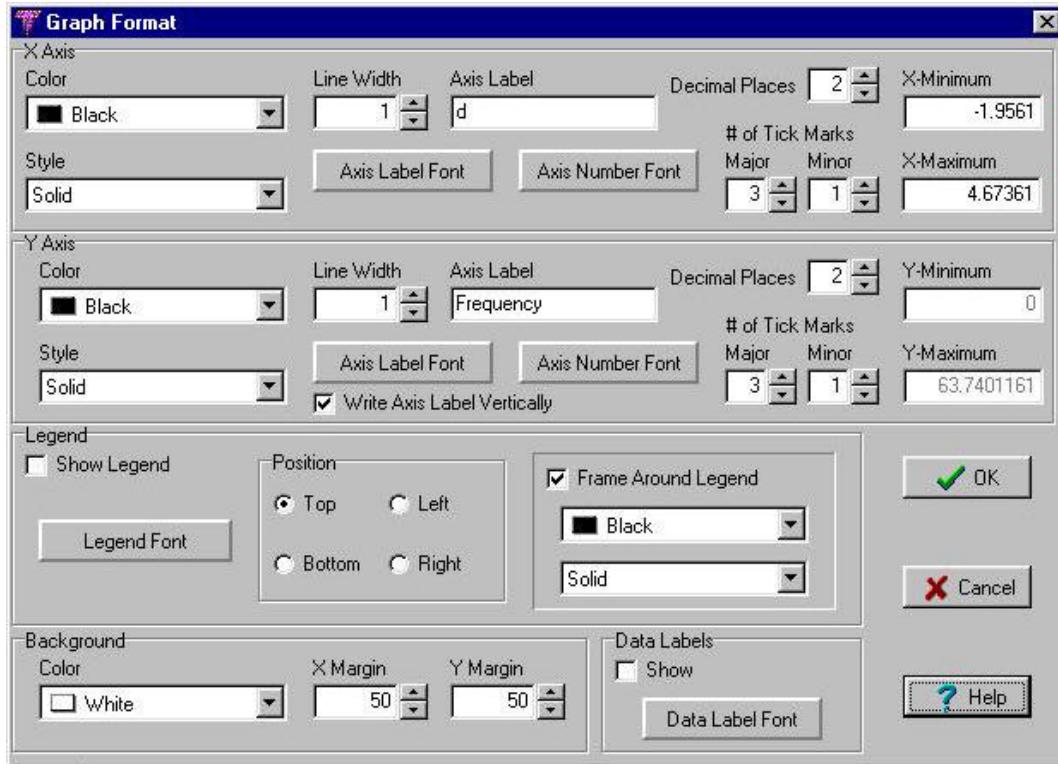


Figure 6.18. Format Graph window of MetaWin.

Point/Line Style Window: Plot options that are specific to a particular graphics type are found on the Point/Line Style window (Figures 6.19 – 6.21). This window allows you to customize the foreground options of the plot. The number and type of elements found in this window change with the type of graph. Scatter plots and their derivatives (funnel plots and regression plots) allow you to customize the symbol used to mark the data. Different groups (see above) may have different symbols. You can choose the type, color, size, and thickness of each symbol. Histogram plots allow you to customize the color and style of the bars. Effect sizes and confidence interval plots and their derivatives (group cumulative effect size plots and cumulative analysis plots) allow you to customize the color, style, and thickness of the lines. As always, different groups can have different colors and styles. Regression lines and confidence intervals and axes in the normal quantile plot can be similarly customized.

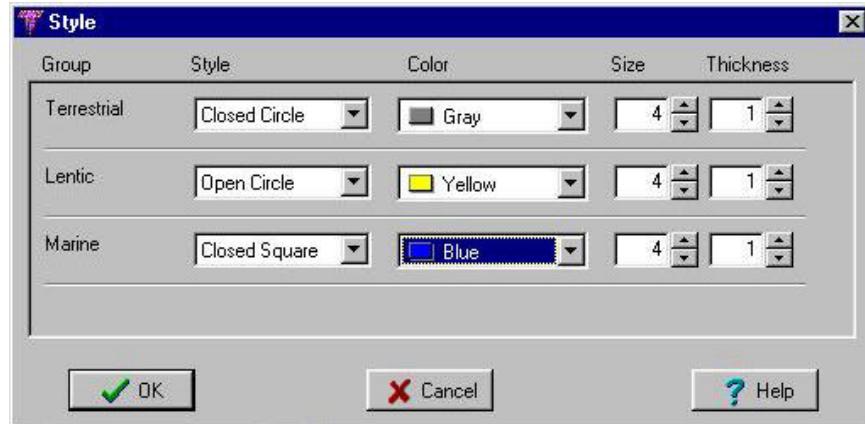


Figure 6.19. Point/Line Style window of MetaWin for a scatter plot with groups specified.

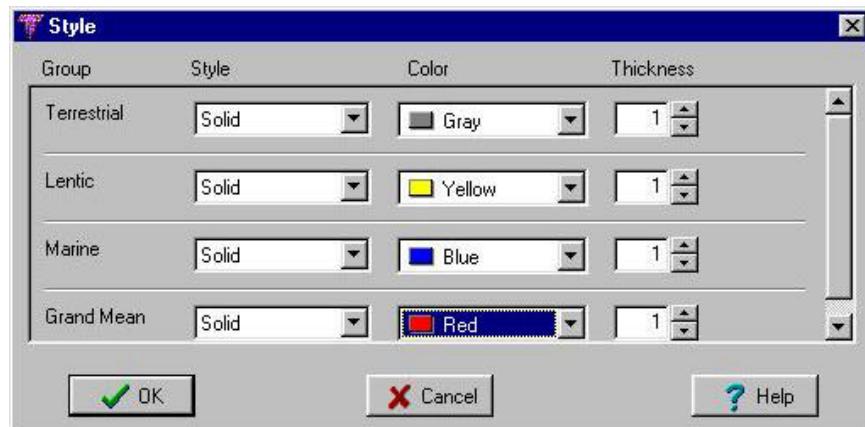


Figure 6.20. Point/Line Style window of MetaWin for an effect size and confidence interval plot.

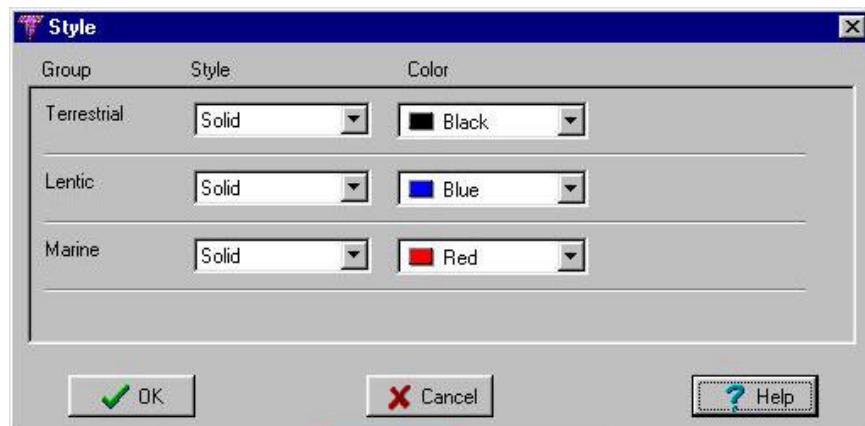


Figure 6.21. Point/Line Style window of MetaWin for a weighted histogram with groups specified.

7. METAWIN STATISTICAL CALCULATOR

7.1 What is MetaCalc?

MetaCalc is a stand-alone program which performs a number of useful statistical procedures which are often used in meta-analysis. MetaCalc is automatically installed on your computer as part of the MetaWin package. It can be run as an independent program or can be called directly from MetaWin by choosing Analysis | S-Calculator from the MetaWin menu or by clicking the  button.

The various procedures performed by MetaCalc are primarily useful in order to construct data for use in a meta-analysis. When conducting a literature review, one quickly finds that the data presented in the primary literature are often found in different forms. Some papers will present the results as correlation coefficients, others will present means and standard deviations, others will present means and standard errors, and others will present just the significance level. The functions in MetaCalc can be used to convert this bewildering array of data into the “raw” data necessary to calculate effect sizes in a meta-analysis (see Section 2.2).

MetaCalc is simple and straight-forward to use. It consists of a single window with three parts (Figure 7.1): a parameter section in the upper right, a function section in the upper left, and an output section in the bottom.

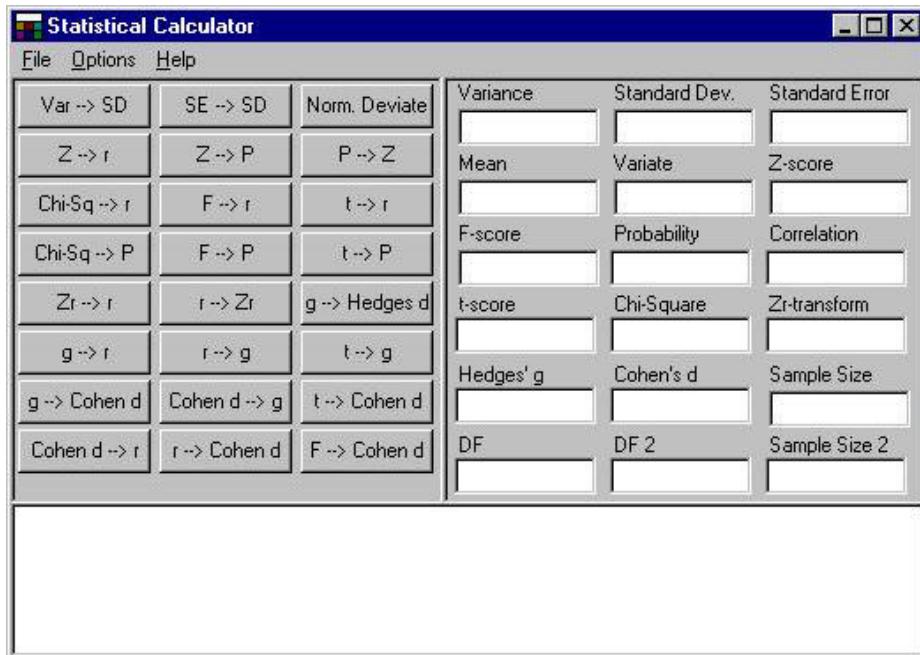


Figure 7.1. The MetaCalc window.

The parameter section contains input boxes for all of the various types of data that may be used by MetaCalc. The function section contains a button for each of the MetaCalc statistical procedures (each function is described in Section 7.2). In order to run a function, simply enter the required data into the correct boxes in the parameter section and press the function button. The results will appear in the output section at the bottom of the window. The results include a list of the parameters that were used in the calculation (prefaced by ">") and the resulting value. Data can be copied to and pasted from the clipboard using the standard control keys (Control-C and Control-V respectively).

MetaCalc has two user-controlled options, both of which are accessible from the Options menu. The first option is called "Put Result in Box," and refers to where the output from each function will be written. The output from each calculation is always written to the output section of the MetaCalc window. If the "Put Result in Box" option is checked, the result is also placed in the appropriate box in the parameter section. The second option is the number of decimal places. This allows you to control how many decimal places to present for all output. This can be a number from zero to fifteen.

7.2 Statistical Functions

Each of the statistical procedures performed by MetaCalc are described below. There are four types of procedures: variance functions, probability functions, correlation functions, and mean difference functions (there is some overlap between these categories). Below is a brief description of the functions that can be accessed through MetaCalc. For each function we give the name, a brief description, the equation (when applicable), and a list of the required input parameters (for more information on many of these and other similar functions, see Rosenthal, 1994).

Variance Functions

Var → SD

Required Parameters: Variance

This function will convert a *variance* to a *standard deviation*.

$$SD = \sqrt{Var} \quad (7.1)$$

SE → SD

Required Parameters: Standard Error (*SE*) and Sample Size (*n*)

This function will convert a *standard error* (also known as standard error of the mean) to a *standard deviation*.

$$SD = SE\sqrt{n} \quad (7.2)$$

Norm. Deviate

Required Parameters: Variate (*X*), Mean (\bar{X}), and Standard Deviation (*SD*)

This function will calculate a *standard normal deviate* (also known as a *Z-score*) from a *variate*, a *mean*, and a *standard deviation*.

$$Z = \frac{X - \bar{X}}{SD} \quad (7.3)$$

*Probability Functions***P → Z**

Required Parameters: Probability

This function will calculate the *standard normal deviate* associated with a one-tailed probability. A one-tailed probability is used to distinguish between positive and negative effects. Therefore, a probability of 0.025 yields a *Z-score* of -1.96, while a probability of 0.975 yields a *Z-score* of 1.96. This is based on the normal distribution (Sokal and Rohlf, 1995).

Z → P

Required Parameters: *Z-score*

This function will calculate the one-tailed probability associated with a *standard normal deviate*. A one-tailed probability is used to distinguish between positive and negative effects. Therefore, a *Z* of -1.96 yields a probability of 0.025, while a *Z* of 1.96 yields a probability of 0.975. This is based on the normal distribution (Sokal and Rohlf, 1995).

t → P

Required Parameters: *t-score* and *DF*

This function will calculate the one-tailed probability associated with a *t-score* with *DF* degrees of freedom, $t_{[DF]}$. A one-tailed probability is used to distinguish between positive and negative effects. Therefore, a *t* of -1.96 with a 20 degrees of freedom yields a probability of 0.032, while a *t* of 1.96 yields a probability of 0.968. This is based on Student's *t-distribution* (Sokal and Rohlf, 1995).

F → P

Required Parameters: F -score, DF , and $DF\ 2$

This function will calculate the probability associated with an F -score with DF degrees of freedom in the numerator and $DF\ 2$ degrees of freedom in the denominator, $F_{[DF,DF\ 2]}$. This is based on the F -distribution (Sokal and Rohlf, 1995).

Chi-Sq → P

Required Parameters: Chi-Square (c^2) and DF

This function will calculate the probability associated with a chi-square value with DF degrees of freedom, $c^2_{[DF]}$. This is based on the c^2 -distribution (Sokal and Rohlf, 1995).

*Correlation Functions***Z → r**

Required Parameters: Z-Score and Sample Size (n)

This function will calculate a *correlation coefficient* from a Z-score (Rosenthal, 1994).

$$r = \frac{Z}{\sqrt{n}} \quad (7.4)$$

t → r

Required Parameters: t -score and DF

This function will calculate a *correlation coefficient* from a t -score with DF degrees of freedom, $t_{[DF]}$ (Rosenthal, 1994).

$$r = \sqrt{\frac{t^2}{t^2 + DF}} \quad (7.5)$$

F → r

Required Parameters: F -score and DF

This function will calculate a *correlation coefficient* from an F -score with 1 degree of freedom in the numerator and DF degrees of freedom in the denominator, $F_{[1,DF]}$ (Rosenthal, 1994). DF should equal $n - 1$, where n is the sample size.

$$r = \sqrt{\frac{F}{F + DF}} \quad (7.6)$$

For F values with more than 1 degree of freedom in the numerator, the calculations and procedures are much more complicated. See Rosenthal and Rosnow (1985, 1991) for more details.

Chi-Sq → r

Required Parameters: Chi-Square (c^2) and Sample Size (n)

This function will calculate a *correlation coefficient* from a chi-square value with 1 degree of freedom, $c^2_{[1]}$ (Rosenthal, 1994).

$$r = \sqrt{\frac{c^2}{n}} \quad (7.7)$$

For c^2 values with more than 1 degree of freedom, the calculations and procedures are much more complicated. See Rosenthal and Rosnow (1985, 1991) for more details.

r → Zr

Required Parameters: Correlation (r)

This function will perform *Fisher's z-transformation* of a *correlation coefficient* (Sokal and Rohlf, 1995).

$$Z_r = \text{TanH}^{-1}(r) = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) \quad (7.8)$$

Zr → r

Required Parameters: *z-transform*

This function will convert *Fisher's z-transformation* back to a *correlation coefficient* (Sokal and Rohlf, 1995).

$$r = \text{TanH}(Z_r) \quad (7.9)$$

r → Cohen d

Required Parameters: Correlation (r)

This function will convert a *correlation coefficient* into a *standardized mean difference*, Cohen's d (Cohen, 1969; Rosenthal, 1994).

$$d = \frac{2r}{\sqrt{1-r^2}} \quad (7.10)$$

Cohen d → r

Required Parameters: Cohen's d , Sample Size (n_e), and Sample Size 2 (n_o)

This function will convert Cohen's d , a *standardized mean difference*, into a correlation coefficient (Cohen, 1969; Rosenthal, 1994).

$$r = \frac{d}{\sqrt{d^2 + \frac{(n_e + n_c)^2}{n_e n_c}}} \quad (7.11)$$

where n_e and n_c are the sample sizes of the experimental and control groups. If the sample sizes are equal (or if they are missing and we assume they are equal) equation 7.11 simplifies to

$$r = \frac{d}{\sqrt{d^2 + 4}} \quad (7.12)$$

MetaCalc uses the general form for unequal sample sizes (equation 7.11).

g → r

Required Parameters: Hedges' g , DF , Sample Size (n_e), and Sample Size 2 (n_c)

This function will convert Hedges' g , a *standardized mean difference*, into a correlation coefficient (Rosenthal, 1994).

$$r = \sqrt{\frac{g^2 n_e n_c}{g^2 n_e n_c + (n_e + n_c) DF}} \quad (7.13)$$

where n_e and n_c are the sample sizes of the experimental and control groups.

r → g

Required Parameters: Correlation (r), DF , Sample Size (n_e), and Sample Size 2 (n_c)

This function will convert a correlation coefficient into Hedges' g , a *standardized mean difference* (Rosenthal, 1994).

$$g = \frac{r}{\sqrt{1 - r^2}} \sqrt{\frac{DF(n_e + n_c)}{n_e n_c}} \quad (7.14)$$

where n_e and n_c are the sample sizes of the experimental and control groups.

Mean Difference Functions

t → Cohen d

Required Parameters: t-score, DF , Sample Size (n_e), and Sample Size 2 (n_c)

This function will calculate Cohen's d , a *standardized mean difference*, from a *t*-score with DF degrees of freedom, $t_{[DF]}$ (Rosenthal, 1994).

$$d = \frac{t(n_e + n_c)}{\sqrt{DF} \sqrt{n_e n_c}} \quad (7.15)$$

where n_e and n_c are the sample sizes of the experimental and control groups. If the sample sizes are equal (or if they are missing and we assume they are equal) equation 7.15 simplifies to

$$d = \frac{2t}{\sqrt{DF}} \quad (7.16)$$

MetaCalc uses the general form for unequal sample sizes (equation 7.15).

F → Cohen d

Required Parameters: F -score, DF , Sample Size (n_e), and Sample Size 2 (n_c)

This function will calculate Cohen's d , a *standardized mean difference*, from an F -score with 1 degree of freedom in the numerator and DF degrees of freedom in the denominator, $F_{[1,DF]}$ (Rosenthal, 1994). DF should equal $n - 1$, where n is the total sample size ($n_e + n_c$).

$$d = \frac{\sqrt{F}(n_e + n_c)}{\sqrt{DF} \sqrt{n_e n_c}} \quad (7.17)$$

where n_e and n_c are the sample sizes of the experimental and control groups. If the sample sizes are equal (or if they are missing and we assume they are equal) equation 7.17 simplifies to

$$d = \frac{2\sqrt{F}}{\sqrt{DF}} \quad (7.18)$$

MetaCalc uses the general form for unequal sample sizes (equation 7.17). For F values with more than 1 degree of freedom in the numerator, the calculations and procedures are much more complicated. See Rosenthal and Rosnow (1985, 1991) for more details.

t → g

Required Parameters: t -score, Sample Size (n_e), and Sample Size 2 (n_c)

This function will convert a t -score into Hedges' g , a *standardized mean difference* (Rosenthal, 1994).

$$g = \frac{t\sqrt{(n_e + n_c)}}{\sqrt{n_e n_c}} \quad (7.19)$$

where n_e and n_c are the sample sizes of the experimental and control groups. If the sample sizes are equal (or if they are missing and we assume they are equal) equation 7.19 simplifies to

$$g = \frac{2t}{\sqrt{N}} \quad (7.20)$$

where N is the total sample size ($n_e + n_c$). MetaCalc uses the general form for unequal sample sizes (equation 7.19).

g → Cohen d

Required Parameters: Hedges' g , DF , Sample Size (n_e), and Sample Size 2 (n_c)

This function will convert Hedges' g into Cohen's d (Rosenthal, 1994). They are different measures of *standardized mean difference*.

$$d = g \sqrt{\frac{n_e + n_c}{DF}} \quad (7.21)$$

Cohen d → g

Required Parameters: Cohen's d , DF , Sample Size (n_e), and Sample Size 2 (n_c)

This function will convert Cohen's d into Hedges' g (Rosenthal, 1994). They are different measures of *standardized mean difference*.

$$g = \frac{d}{\sqrt{\frac{n_e + n_c}{DF}}} \quad (7.22)$$

g → Hedges d

Required Parameters: Hedges' g , Sample Size (n_e), and Sample Size 2 (n_c)

This function will convert Hedges' g to Hedges' d . They are both measures of *standardized mean difference*. Hedges' d is an unbiased estimator of g .

$$d = g \left(1 - \frac{3}{4(n_e + n_c - 2) - 1} \right) \quad (7.23)$$

APPENDIX I: EXAMPLE DATA SETS

Throughout this manual we have used several example data sets to demonstrate the various meta-analytic methods available in MetaWin. Below is a brief description of these data sets. They are included with MetaWin and will be installed on your hard drive in a subdirectory of the program. We have chosen two different types of primary data for our example data sets. The first (from Gurevitch and Hedges, 1993) is an example of primary data containing means, standard deviations, and sample sizes for each study. The second (from Berlin et al., 1989) is an example of odds ratio data obtained from a series of two x two contingency tables. Note that this second data set is presented as effect sizes and their variances: the original contingency table data are not given. Each of these is discussed and presented below.

Gurevitch and Hedges (1993) Data Set

The data of Gurevitch and Hedges (1993) is an example of the type of primary data often available from the literature in the biological sciences. It is part of a larger study on the effects of competition in biological field experiments. The larger study (Gurevitch et al., 1992) included all measures of outcome (survivorship, reproductive output, density, etc.) in response to the manipulation of competitors for organisms in a wide range of trophic levels and systems. Detailed description of how studies were selected for inclusion in the summary analysis is presented in the original study (Gurevitch et al, 1992).

The example data set of Gurevitch and Hedges, 1993 (Table 17.1) show responses to competition by primary producers. This response was measured as recruitment, an increase in the number of individuals, or growth, an increase in the size of individuals. For each study, the experimental and control group mean response, sample size, and standard deviation are presented (Table A.1). Two additional columns are listed: one contains the species names and a second specifies the author of each study. This latter column contains two letter author codes, which are described in Table A.2.

In addition to the primary data, a column containing reversal markers for each study is found. Reversal markers are used to specify the direction of the expected effect for each individual study when studies have been measured in “different” directions. In this data example, the results of some studies are from experiments in which the competitor was removed from plots, where results from other studies are from experiments where competitor density was increased. Finally, the studies have been placed into one of three trophic level groups: terrestrial organisms, lentic organisms, or marine organisms. The categorical grouping variable for this information is called “Habitat.” The data file is called “gur_hedge.dta” and has been saved as text with tabs.

Table A.1. Example data set from Gurevitch and Hedges (1993).

Habitat	+/-	Nc	Ne	Xc	Xe	Sc	Se	Author	Species
Terrestrial	+	7	7	78.14	79.71	40.65	40.65	Fo	<i>Bouteloua rigidiseta</i>
Terrestrial	+	7	7	18.86	26	9.17	9.17	Fo	<i>Aristida longiseta</i>
Terrestrial	-	6	6	-1.8	-2.1	0.49	0.49	Pl	<i>Mirabilis hirsuta</i>
Terrestrial	-	5	5	-2.2	-2.8	0.224	0.447	Pl	<i>Verbena stricta</i>
Terrestrial	-	7	7	-2.1	-3	0.265	0.529	Pl	<i>Solidago rigada</i>
Terrestrial	-	6	6	-2.3	-4.2	0.49	1.225	Pl	<i>Asclepias syriaca</i>
Terrestrial	+	3	3	85.3	285.7	115.008	153.806	Gr	<i>Verbascum thapsus</i>
Terrestrial	+	3	3	0	3	0	2.425	Gr	<i>Oenothera biennis</i>
Terrestrial	+	3	3	0	2	0	2.078	Gr	<i>Verbascum thapsus</i>
Terrestrial	+	3	3	0	1.67	0	1.732	Gr	<i>Oenothera biennis</i>
Terrestrial	+	5	5	17	17	7.603	5.367	Po	<i>Plantago major</i>
Terrestrial	+	5	5	47	37	10.286	9.391	Po	<i>Plantago lanceolata</i>
Terrestrial	+	4	4	87	272	37.712	183.532	Bu	<i>Metrosideros polymorpha</i>
Terrestrial	+	18	20	-0.113	0.294	0.255	0.215	Gu	<i>Stipa neomexicana</i>
Terrestrial	+	20	20	-0.163	0.412	0.588	0.218	Gu	<i>Stipa neomexicana</i>
Terrestrial	+	18	20	0.14	0.632	0.38	0.359	Gu	<i>Stipa neomexicana</i>
Terrestrial	+	20	20	-0.184	0.259	0.326	0.238	Gu	<i>Aristida glauca</i>
Terrestrial	+	20	20	-0.075	0.354	0.487	0.182	Gu	<i>Aristida glauca</i>
Terrestrial	+	20	20	0.147	0.541	0.34	0.299	Gu	<i>Aristida glauca</i>
Lentic	-	4	4	281.11	-201.03	158.038	27.52	Mc	<i>Eleocharis acicularis</i>
Lentic	-	4	4	187.31	-155.32	80.163	41.252	Mc	<i>Juncus pelocarpus</i>
Marine	+	7	7	11.8	16	3.08	3.37	St	<i>Acropora spp.</i>
Marine	+	3	10	0.4	9.5	1.47	7.23	St	<i>Pocillopora verrucosa</i>
Marine	+	20	20	0	14.1	7.603	7.603	Re	<i>Pterygophora californica</i>
Marine	+	20	20	0	7.1	3.13	3.13	Re	<i>Macrocystis pyrifera</i>
Marine	+	20	20	0	1.4	1.789	1.789	Re	<i>Desmarestia ligulata</i>
Marine	+	10	10	82.2	94	29.093	9.171	Re	<i>Desmarestia ligulata</i>
Marine	+	10	10	8.3	10.5	14.546	11.068	Re	<i>Desmarestiia kurilensis</i>
Marine	+	10	10	0	20	42.691	42.691	Re	<i>Nereocystis luetkeana</i>
Marine	+	2	2	3.63	18.5	3.352	4.257	Jo	<i>Laminaria longicurvis</i>
Marine	+	2	2	0	0.25	0	0.354	Jo	<i>Laminaria longicurvis</i>
Marine	+	2	2	3.63	2.25	3.354	0.707	Jo	<i>Laminaria longicurvis</i>
Marine	+	4	4	0	34.8	0	58.2	Tu	<i>Rhodemela larix</i>
Marine	+	4	4	0	25.3	0	35.8	Tu	<i>Cryptosiphonia woodii</i>
Marine	+	4	4	5.4	23.6	10.88	47	Tu	<i>Phaeostrophion irregularare</i>
Marine	+	4	4	1.8	10.5	5.2	24.2	Tu	<i>Odonthalia floccosa</i>
Marine	+	4	4	0	10.3	0	17.4	Tu	<i>Mirocladia borealis</i>
Marine	+	4	4	0	8.7	0	17	Tu	<i>Fucus distichus</i>
Marine	+	4	4	0	5.7	0	14	Tu	<i>Iridaea heterocarpa</i>
Marine	+	4	4	10.8	5.4	15.8	8.8	Tu	<i>Bossiella plumosa</i>
Marine	+	4	4	21.25	37.25	9.54	22.02	Du	<i>Ralfsia pacifica</i>
Marine	+	4	4	40.25	20.25	8.78	9	Du	<i>Ralfsia pacifica</i>
Marine	+	5	5	15.8445	11.9533	10.787	6.24	Du	<i>Ralfsia pacifica</i>

Table A.2. List of author codes from Gurevitch and Hedges (1993).

Code	Author/s
Fo	Fowler, 1986
Pl	Platt and Weiss, 1985
Gr	Gross and Werner, 1982
Po	Pons and van der Toorn, 1988
Bu	Burton and Meuller-Dombois, 1984
Gu	Gurevitch, 1986
Mc	McCreary et al., 1983
St	Stimson, 1985
Re	Reed and Foster, 1984
Jo	Johnson and Mann, 1988
Tu	Turner, 1985
Du	Dungan, 1986

Berlin et al. (1989) Data Set

Berlin et al. (1989) compared two statistical methods for combining effect sizes from clinical trials: the Peto-modified Mantel-Haenszel method for combining odds ratios (using a fixed-effects model), and the DerSimonian and Laird method for combining rate differences (using a random-effects model). To compare the two methods, Berlin et al. (1989) combined the results from 22 previously published meta-analyses. The data are found in Table 1 of their original paper.

For our second example data set, we chose the Peto method summary data from Berlin et al. (1989). This data is comprised of the cumulative natural log *odds ratio* and variance for each meta-analytic study. In Berlin et al. (1989), only cumulative effects sizes and their standard errors are reported. For our purposes we have calculated the variance for each study as the square of the standard error. The data from Berlin et al. (1989) are reproduced here in Table A.3. The data file is called “berlinor.dta” and has been saved as text with tabs.

Table A.3. Example data set from Berlin et al. (1989).

lnOR	variance
0.2000	0.0625
0.0700	0.0361
0.5700	0.0729
0.2400	0.0081
0.0800	0.0081
0.1700	0.0225
0.0400	0.0361
0.6100	0.0196
0.3400	0.0144
1.1200	0.0441
1.3700	0.0676
0.1500	0.0169
0.1700	0.0225
0.8100	0.0784
0.9300	0.1296
0.6400	0.1296
0.8900	0.3721
1.6900	0.3969
1.5100	0.1024
1.9000	0.0784
0.1600	0.0144
0.1300	0.0196

APPENDIX II: WARNING AND ERROR MESSAGES

Because there are a large number of numerical calculations performed in MetaWin and MetaCalc, there are many places where things can go wrong. We have attempted to make these programs as bullet-proof as possible by creating a series of “catch loops” designed to detect potential problems before analyses are run. Many of these loops will generate either a Warning Message or an Error Message. Warning messages tell the user that MetaWin (or MetaCalc) has caught an invalid option, and has ignored it, while Error Messages inform the user that an invalid option that could not be ignored was detected. Catching such errors prevents the programs from crashing unexpectedly.

We have attempted to make the Warning and Error messages in MetaWin and MetaCalc as self-explanatory as possible. Nevertheless, some of these messages may be a bit obscure. Below is a list of all the Warning and Error messages that can be generated from MetaWin and MetaCalc. We have given each message a name, and have provided a description of the message and how you can correct the problem. The Warning and Error messages are divided into those that can be generated in MetaWin, and those that can be generated in MetaCalc.

METAWIN MESSAGES

Blank Group Names: Some of the entries in the grouping column were blank. These have been treated as a single group named “<<BLANK>>” If you did not want these studies included in the analysis, remove them using the Refine Categories option of the Summary Analysis window.

Conflict Between Categorical Structure and Cumulative Analysis: MetaWin will not conduct a cumulative analysis with categorical model structure. Turn off the cumulative analysis or choose a different model structure to continue.

Data Matrix Has Been Altered: The Summary Analysis options are dependent on the structure of the data matrix. If the matrix is unaltered, these options are preserved through multiple runs of the analysis. If any changes are made to the data, all of the options are reset.

Error Reading File: There are a number of reasons MetaWin may be unable to read a file. The most common of these are:

- 1) **Access Denied:** You may not have permission to read from this disk or directory.
- 2) **File in Use:** Make sure the file is not already being used by another program. Close any program accessing the file and try again.
- 3) **Invalid Text File:** You were attempting to load a non-text file as text.

Error Writing to File: There are a number of reasons MetaWin may be unable to write to an output file. The most common of these are:

- 1) **Disk Full:** Make sure you have enough free space to save the file and try again.
- 2) **Access Denied:** You may not have permission to write to this disk or directory. Try choosing another location.
- 3) **File in Use:** Make sure the file is not already being used by another program. Close any program accessing the file and try again.

File Contains No Data: MetaWin was unable to open the file because it contains no data.

File Does Not Exist: The file you tried to open from the previously used file list does not exist.

Groups Were Eliminated From Analysis: Some groups were eliminated from the categorical summary analysis because they did not contain at least two valid studies. Studies are considered invalid if they are missing data, if their variances are less than or equal to 0.

Invalid Axes Ranges: The minimum and maximum values for an axis must be logically consistent. You cannot choose a minimum value which is greater than or equal to the maximum value.

Invalid Data for Exploratory Analysis: There were no valid data points on which to perform data exploration. Studies are considered invalid if they are missing data, if their variances are less than or equal to 0, or if their sample sizes are less than or equal to 1.

Invalid Excel File: MetaWin can read Excel files, versions 2 through 9. The file you tried to import does not appear to be any of these types. **Note:** Excel for Windows95 is also known as Excel version 7; Excel97 is also known as Excel version 8; Excel2000 is also known as Excel version 9.

Invalid Lotus File: MetaWin can read Lotus 1-2-3 files, versions 1 through 3.x, and Symphony files, versions 1 through 2.x. The file you tried to import does not appear to be any of these types. MetaWin cannot read newer versions of Lotus 1-2-3 files at this time.

Invalid Reversal Column: There were invalid entries in the reversal column. When this error occurs, the reversal column is ignored during the calculation of effect sizes. The reversal column can be either numeric or text. Valid entries for the reversal column include numeric entries (-1 and 1, or blank), and text entries (-1 and 1, or - and +, or

blank). **Note:** For 1 and -1 in both numeric and text columns, the number of decimal places does not matter.

Mathematical Incompatibility in Calculation of Effect Size: Effect sizes could not be calculated for some studies because of mathematical incompatibilities with the desired effect size. No effect size was calculated for these studies. The specific incompatibility is dependent on the effect size being calculated. These include:

- 1) Samples sizes for Hedges' d , $\ln R$ must be greater than 1 and variances cannot be negative.
- 2) For $\ln R$, the control and experimental means cannot be equal to zero and must have the same sign.
- 3) No values can be negative for $\ln OR$, RD , and $\ln RR$.
- 4) Correlation coefficient must be between -1 and 1 for Fisher's z-transform and the sample size must be greater than 3.

Missing Values in Calculation of Effect Size: Some studies were missing required values for the proper calculation of an effect size. No effect size was calculated for studies with blank entries in the selected variable columns.

No Data Has Been Loaded: None of the analyses will run without data.

No Numeric Data: None of the analyses will run without numeric data. If your data are in text columns, change their format to numeric before running an analysis.

No Valid Data for Rank Correlation Test: There were not enough data points to perform the rank correlation test. MetaWin requires at least 2 points to perform these tests. Studies are considered invalid if they are missing data, if their variances are less than or equal to 0, or if their sample sizes are less than or equal to 1.

No Valid Points: MetaWin was unable to create a plot because there were no studies with valid plotting values given the current choices. A study will not be plotted if it is missing a value from one of the plot variables (e.g. Effect Size). A study will also not be plotted if it has an invalid variance (for Weighted Histograms, Radial Plots, Normal Quantile Plots, and Confidence Interval Plots only); the value in the variance column must be greater than zero to be considered valid. **Note:** The Normal Quantile plot requires a minimum of three valid points.

Not Enough Data for Categorical Model: There were not enough groups to perform a categorical summary analysis. MetaWin requires at least 2 valid groups to perform this test. Groups are considered invalid if they have fewer than 1 valid study within them (studies are considered invalid if they are missing data or if their variances are less than or equal to 0). MetaWin will conduct an unstructured analysis on the valid studies.

Not Enough Data for Continuous Model: There were not enough data points to perform a continuous summary analysis. MetaWin requires at least 3 valid studies to perform this test. Studies are considered invalid if they are missing data or if their variances are less than or equal to 0. MetaWin will conduct an unstructured analysis on the valid studies.

Out of Memory: MetaWin was unable to complete the desired function because of a lack of memory. Try closing extraneous programs, windows, and/or rebooting the computer. If the problem persists, try reducing the size of your data matrix.

Out of Resources: MetaWin was unable to complete the desired function because of a lack of windows resources. Try closing extraneous programs, windows, and/or rebooting the computer. If the problem persists, try reducing the size of your data matrix.

Spreadsheet Errors Were Encountered: MetaWin may have encountered errors within the spreadsheet you were attempting to import. You may want to check the spreadsheet carefully to make sure that it imported properly. If you are encountering problems or MetaWin is unable to import a worksheet properly, you may want to try saving the file in another format (perhaps without formulas) or try copy and pasting the data directly into MetaWin's spreadsheet.

Studies Were Eliminated From Exploratory Analysis: Some studies were eliminated from the exploratory analysis because their data did not fit the basic requirements of the analysis. Studies are considered invalid if they are missing data, if their variances are less than or equal to 0, or if their sample sizes are less than or equal to 1. Eliminated studies are listed in the output.

Studies Were Eliminated From Plot: Some studies were eliminated from the plot because their data did not fit the basic requirements of the selected plot. Studies are considered invalid if they are missing data or (for all plots other than the Scatter Plot) if their variances are less than or equal to 0. Eliminated studies are listed in the output.

Unable to Change Format of Text Column: MetaWin was unable to convert the text column into a numeric column because the text column contained non-numeric entries. A text column can be converted into a numeric column if it contains only numbers and blanks.

Unable to Paste Text into a Numeric Column: MetaWin was unable to properly paste the contents from the clipboard because you tried to paste text into numeric columns. All cells that could not be pasted into were left blank.

Unexpected Math Error: An unexpected math error has occurred. Try shutting down MetaWin and/or your computer and rerunning the application. If the problem persists, please contact the authors.

METACALC MESSAGES

Correlation Coefficient is too Extreme: This error will occur if you try to convert a correlation coefficient into a standardized mean difference (either Cohen's d or Hedges' g) and your correlation is equal to exactly -1 or 1 . In these cases, a correlation of -1 indicates a mean difference of negative infinity and a correlation of 1 indicates a mean difference of positive infinity.

Invalid Chi-Square: This error will occur when a function requiring a χ^2 is run and the *Chi-Square* box does not contain a real number greater than or equal to zero.

Invalid Cohen's d : This error will occur when a function requiring Cohen's d is run and the *Cohen's d* box does not contain a real number.

Invalid Correlation Coefficient: This error will occur when a function requiring a correlation coefficient is run and the *Correlation* box does not contain a real number between -1 and 1 , inclusive.

Invalid Degrees of Freedom: This error will occur when a function requiring degrees of freedom is run and the *DF* box does not contain a real number greater than zero. If two degrees of freedom measures are required, the same error will occur for the *DF 2* box.

Invalid F-score: This error will occur when a function requiring an *F*-score is run and the *F-score* box does not contain a real number greater than or equal to zero.

Invalid Hedges' g : This error will occur when a function requiring Hedges' g is run and the *Hedges' g* box does not contain a real number.

Invalid Mean: This error will occur when a function requiring a mean is run and the *Mean* box does not contain a real number.

Invalid Probability: This error will occur when a function requiring a probability is run and the *Probability* box does not contain a real number between 0 and 1 , inclusive.

Invalid Sample Size: This error will occur when a function requiring a sample size is run and the *Sample Size* box does not contain a real number greater than one. If two sample sizes are required, the same error will occur for the *Sample Size 2* box.

Invalid Standard Deviation: This error will occur when a function requiring a standard deviation is run and the *Standard Deviation* box does not contain a real number greater than zero.

Invalid Standard Error: This error will occur when a function requiring a standard error is run and the *Standard Error* box does not contain a real number greater than or equal to zero.

Invalid t-score: This error will occur when a function requiring a *t*-score is run and the *t-score* box does not contain a real number.

Invalid Variance: This error will occur when a function requiring a variance is run and the *Variance* box does not contain a real number greater than or equal to zero.

Invalid Variate: This error will occur when a function requiring a variate is run and the *Variate* box does not contain a real number.

Invalid Z-score: This error will occur when a function requiring a *Z*-score is run and the *Z-score* box does not contain a real number.

Invalid z-transform: This error will occur when a function requiring a *z*-transformation is run and the *Zr-transform* box does not contain a real number.

Probability is too Extreme: This error will occur if you try to convert a probability into a *Z*-score and your probability is equal to exactly 0 or 1. In these cases, a probability of zero indicates a *Z*-score of negative infinity and a probability of 1 indicates a *Z*-score of positive infinity.

Unexpected Math Error: An unexpected math error has occurred. Try shutting down MetaCalc and/or your computer and rerunning the application. If the problem persists, please contact the authors.

GLOSSARY

Between-Group Heterogeneity: The *Between-Group Heterogeneity*, Q_B , describes the variation between *Group Cumulative Effect Sizes* in a *Categorical Summary Analysis*, and is identical to the *Model Heterogeneity* for that analysis. See *Categorical Summary Analysis*; *Group Cumulative Effect Size*; *Model Heterogeneity*.

Between-Study Variance: See *Pooled Study Variance*.

Bias-Corrected Bootstrap Confidence Intervals: *Bias-Corrected Bootstrap Confidence Intervals* are nonparametric confidence intervals that correct for bootstrap distributions where more than 50% of the bootstrap replicates are either above or below the observed value. See *Bootstrap Confidence Intervals*; *Confidence Intervals*.

Bootstrap Confidence Intervals: *Bootstrap Confidence Intervals* are nonparametric confidence intervals used to estimate the range of uncertainty for a given test-statistic. They are calculated by choosing with replacement, X studies (from a sample of X), and then calculating the desired test-statistic. This procedure is repeated many times to generate a distribution of test-statistics, from which the lowest and highest 2.5% values are chosen to represent the lower and upper 95% confidence limits. See *Bias-Corrected Confidence Intervals*; *Confidence Intervals*.

Categorical Summary Analysis: A *Categorical Summary Analysis* is a modern meta-analysis where the individual studies can be categorized into more than one group. In this model, the overall *Cumulative Effect Size* and *Group Cumulative Effect Sizes* can be calculated and tested for significance, and the *Total Heterogeneity* in effect sizes, Q_T , is partitioned into variance explained by the model, Q_M , and residual error variance *not* explained by the model, Q_E . This partitioning of Q_T is similar to the partitioning of variance in ANOVA. See *Summary Analysis*; *Cumulative Effect Size*; *Error Heterogeneity*; *Group Cumulative Effect Size*; *Model Heterogeneity*; *Total Heterogeneity*.

Cohen's d : Cohen's d (Cohen, 1969) is an estimate of the *Standardized Mean Difference* similar to Hedges' g . Using the experimental and control group means and sample sizes, it is calculated as:

$$d_{\text{Cohen}} = \frac{\left(\bar{X}^E - \bar{X}^C \right)}{s}$$

where the pooled standard deviation is found from:

$$s = \sqrt{\frac{(N^E - 1)(s^E)^2 + (N^C - 1)(s^C)^2}{N^E + N^C}}.$$

See *Effect Size; Glass' D; Hedges' d; Hedges' g; Means and Standard Deviation Data; Response Ratio; Standardized Mean Difference*.

Combining Probability Levels: A series of quantitative methods for synthesizing the results of a set of primary research studies by statistically combining the exact probability values for those studies, or by combining test-statistics (e.g., Z-scores) based on the exact probabilities (e.g., Fisher, 1932; Pearson, 1933). These methods are often preferred over vote-counting techniques because they take into consideration the sample size information from individual studies. A disadvantage of these techniques is that they are unable to estimate the *magnitude* of the effect of interest. See *Modern Meta-Analysis; Research Synthesis; Vote-Counting*.

Confidence Intervals: *Confidence Intervals* are a method of estimating the uncertainty of a given statistic and represent the range of values that is expected to contain the true value of that test-statistic (Sokal and Rohlf, 1995). The parametric *Confidence Intervals* around a Cumulative Effect Size are found using the variance of the *Cumulative Effect Size*, and are calculated as:

$$CI = \bar{E} \pm t_{\alpha/2[n-1]} * s_{\bar{E}}$$

where $s_{\bar{E}} = \sqrt{\frac{1}{\sum_{i=1}^n w_i}}$ and t is the two-tailed critical value from Student's t -distribution.

Continuous Summary Analysis: A *Continuous Summary Analysis* is a modern meta-analysis where the *Effect Sizes* from individual studies can be explained by another independent variable. *Continuous Summary Analyses* perform a weighted linear regression analysis of *Effect Sizes* versus an independent variable, which is tested for significance. The *Total Heterogeneity* in effect sizes, Q_T , is also partitioned into variance explained by the model, Q_M , and residual error variance *not* explained by the model, Q_E . This partitioning of Q_T is similar to the partitioning of variance ANOVA. See *Summary Analysis; Cumulative Effect Size; Error Heterogeneity; Model Heterogeneity; Total Heterogeneity*.

Correlation Data: Often the data presented in a set of primary research studies are found in different forms, using different test-statistics. For such data, one may transform the various test-statistics into correlation coefficients, and use these to calculate effect sizes (Rosenthal, 1991). Typically, Fisher's z-transform is calculated from correlation coefficients. See *Fisher's z-Transformation*.

Cumulative Effect Size: The *Cumulative Effect Size* is a measure of central tendency for a set of studies found as the weighted sum of effect sizes divided by the sum of the weights. This is a weighted average, rather than an unweighted mean. For a set of n studies the *Cumulative Effect Size* is calculated as:

$$\bar{E} = \frac{\sum_{i=1}^n w_i E_i}{\sum_{i=1}^n w_i}$$

where E_i is the *Effect Size* for the i^{th} study and w_i is its weight. The weight is found as the reciprocal of the study sampling variance, or $w_i = 1/v_i$. *Cumulative Effect Sizes* are assessed statistically using *Confidence Intervals*. See *Confidence Intervals; Group Cumulative Effect Size*.

Cumulative Meta-Analysis: *Cumulative Meta-Analysis* is a series of *Summary Analyses* where studies are successively added to the analysis based on some predetermined order (Chalmers, 1991; Antman et al., 1992). For each set of included studies, the summary meta-analytic statistics (usually the *Cumulative Effect Size* and its variance) are calculated, another study is added, and the statistics are recalculated. The summary statistics are then placed in sequential order, and compared to one another to determine when a given result could have been discovered. See *Summary Analysis*.

Effect Size: An *Effect Size* is a statistical measure portraying the degree to which a given event is present in a sample (Cohen, 1969). The type of measure (e.g., *Standardized Mean Difference*) is called the effect, and its *magnitude* is considered an *Effect Size*. Different measures of effect size are calculated for different types of primary data. See *Correlation Data; Means and Standard Deviation Data; Standardized Mean Difference; Two x Two Contingency Data*.

Error Heterogeneity: The *Error Heterogeneity*, Q_E , is a quantity that describes the residual error variation in *Effect Sizes* not explained by the *Summary Analysis* model. Q_E is tested against a c^2 -distribution with $n - m$ degrees of freedom for a *Categorical Summary Analysis* and $n - 2$ degrees of freedom for a *Continuous Summary Analysis*. Q_E is sometimes referred to as the within-class heterogeneity, Q_W , for a *Categorical Summary Analysis* (see Hedges and Olkin, 1985). It is similar to the error variance of an ANOVA. See *Categorical Summary Analysis; Continuous Summary Analysis; Effect Size; Error Heterogeneity; Summary Analysis; Total Heterogeneity*.

Fail-Safe Tests: *Fail-Safe Tests* are statistical methods for estimating the magnitude of the *Publication Bias* known as the *File-Drawer Problem* (Rosenthal, 1979). These techniques calculate the number of nonsignificant, unpublished studies that need to be added to a *Summary Analysis* in order to change the results from significant to not

significant. A large fail-safe number indicates that many unpublished studies are required to change the statistical results, and thus one may be more confidence in the results from the *Summary Analysis*. See *Publication Bias*.

File-Drawer Problem: See *Publication Bias*.

Fisher's z-Transformation: *Fisher's z-Transformation* is a measure of *Effect Size* that can be calculated from data represented as correlation coefficients (Fisher, 1928). It is found as:

$$z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right).$$

Fixed-Effects Model: *Fixed-Effects Models* are *Summary Analyses* where it is assumed that there is one true *Effect Size* shared by all studies or one true *Effect Size* for each group of studies (Hedges, 1994; Gurevitch and Hedges, 1999). For fixed-effects summary analyses, all summary statistics are calculated using study weights defined as the reciprocal of the study variance, or $w_i = 1/v_i$. See *Random-Effects Model*.

Funnel Plot: A *Funnel Plot* is a graphical method for exploring meta-analytic data where the *Effect Size* of each individual study is plotted against its sample size (Light and Pillemer, 1984). If there are no biases in publication, three basic predictions can be made about this plot (Palmer, 1999): 1) studies with small sample sizes should have increased sampling error relative to those with larger sample sizes, 2) the *Cumulative Effect Size* should be independent of sample size, and 3) at a given sample size, individual studies should be normally distributed around the *Cumulative Effect Size*. The result is a scatter plot shaped like a funnel with the large opening at the smallest sample size. See *Publication Bias*.

General Linear Model: *General Linear Models* (GLM) are a class of statistical models that use matrix algebra to determine the relationship between a set of dependent and independent variables. For modern meta-analysis, the dependent variables are individual study *Effect Sizes* and the independent variables describe the underlying structure of the data (e.g., categorical data structure or continuous data structure). All models of *Summary Analyses* can be performed using a weighted GLM procedure (see Hedges and Olkin, 1985; 2000).

Glass' Δ : Glass (1976) developed an estimate of the *Standardized Mean Difference* which is calculated as:

$$\Delta = \frac{\left(\bar{X}^E - \bar{X}^C \right)}{s^C}$$

where \bar{X}^C and \bar{X}^E are the experimental and control group means and s^C is the control group sample size. See *Cohen's d; Effect Size; Hedges' d; Hedges' g; Means and Standard Deviation Data; Response Ratio; Standardized Mean Difference*.

Grouped Analysis: See *Categorical Summary Analysis*.

Group Cumulative Effect Size: When studies can be separated into more than one group, an estimate of the mean *Cumulative Effect Size* can be calculated for each group. This is a weighted estimate calculated as:

$$\bar{E}_j = \frac{\sum_{i=1}^{k_j} w_{ij} E_{ij}}{\sum_{i=1}^{k_j} w_{ij}}$$

where k_j is the number of studies in the j^{th} group, and w_{ij} and E_{ij} are the weight and effect size for the i^{th} study in the j^{th} group. The weight is found as the reciprocal of the study sampling variance, or $w_{ij} = 1/v_{ij}$. See *Cumulative Effect Size*.

Hedges' d: *Hedges' d* is an estimate of the *Standardized Mean Difference* that is not biased by small sample sizes (Hedges and Olkin, 1985). Using experimental and control group means, standard deviations, and sample sizes, it is calculated as:

$$d = g \left(1 - \frac{3}{4(N^C + N^E - 2) - 1} \right)$$

where g is *Hedges' g*. See *Cohen's d; Effect Size; Glass' D; Hedges' g; Means and Standard Deviation Data; Response Ratio; Standardized Mean Difference*.

Hedges' g: *Hedges' g* is an estimate of the *Standardized Mean Difference* that accounts for the fact that the sampling variances for both the experimental and control groups are not always equal (Hedges, 1981). Using the experimental and control group means, standard deviations and sample sizes, this *Effect Size* is calculated as:

$$g = \frac{(\bar{X}^E - \bar{X}^C)}{S}$$

where the pooled standard deviation is found from:

$$S = \sqrt{\frac{(N^E - 1)(s^E)^2 + (N^C - 1)(s^C)^2}{N^E + N^C - 2}}$$

If there are small sample sizes, Hedges' g can be biased. See *Cohen's d; Effect Size; Glass' D; Hedges' d; Means and Standard Deviation Data; Response Ratio; Standardized Mean Difference*.

Mantel-Haenszel: See *Odds Ratio*.

Mean Effect Size: See *Cumulative Effect Size*.

Means and Standard Deviation Data: A common type of data presentation from primary research studies in which the statistical results are represented as means, standard deviations, and sample sizes for both the experimental group and the control group. From these data, a *Standardized Mean Difference* can be often calculated. Several estimates of the *Standardized Mean Difference* have been developed, and are used as effect sizes. See *Cohen's d; Effect Size; Glass' D; Hedges' d; Hedges' g; Response Ratio; Standardized Mean Difference*.

Mixed-Effects Model: A *Mixed-Effects Model* is a *Random-Effects Model* for categorical data (Gurevitch and Hedges, 1993). See *Random-Effects Model; Pooled Study Variance*.

Model Heterogeneity: The *Model Heterogeneity*, Q_M , is a quantity that describes the variation in *Effect Sizes* explained by the *Summary Analysis* model. Q_M from a *Categorical Summary Analysis* is tested against a χ^2 -distribution with $m - 1$ degrees of freedom, while Q_M from a *Continuous Summary Analysis* is tested against a χ^2 -distribution with 1 degree of freedom. See *Categorical Summary Analysis; Continuous Summary Analysis; Effect Size; Error Heterogeneity; Summary Analysis; Total Heterogeneity*.

Model Structure: The *Model Structure* is defined by the set of independent variables used to describe the structure of the data in the *Summary Analysis*. When all of the studies come from one sample, the *Model Structure* is called NO model structure. Studies can be grouped in a categorical model, or an additional continuous variable can be used to explain variation in *Effect Sizes* in a continuous model. See *Categorical Summary Analysis; Continuous Summary Analysis; Summary Analysis*.

Moderator Variable: See *Categorical Summary Analysis*.

Modern Meta-Analysis: A quantitative method for synthesizing the results from a set of primary research studies that combine the measures of effects from individual studies and estimate the overall effect for that set of studies. This combined effect is then used to statistically assess whether there is evidence in favor of the hypothesis

being tested. See *Combining Effect Size; Probability Methods; Research Synthesis; Vote Counting*.

Nonparametric Variance: When Hedges' d is used as the estimate of individual study *Effect Sizes*, Adams et al. (1997) proposed a nonparametric weighting scheme that can be used in *Summary Analyses* with resampling tests. The nonparametric variance estimates use only the sample sizes from experimental and control groups rather than incorporating the *Effect Size* into the calculation, and are found as:

$$v_{np} = \frac{N_i^C + N_i^E}{N_i^C N_i^E}$$

where N_i^C and N_i^E are the sample sizes from the experimental and control group of the i^{th} study (see Adams et al., 1997). The reciprocal of the nonparametric variance is then used as the nonparametric weight.

Normal Quantile Plot: A *Normal Quantile Plot* is a graphical method for exploring meta-analytic data where the standardized *Effect Size* of each individual study is plotted against its normal quantile value (Wang and Bushman, 1998). If the distribution in *Effect Sizes* is similar to the normal distribution, these points should fall close to the line $X = Y$. Deviation from this line reveal how the distributions are different. See *Publication Bias*.

Odds Ratio: The *Odds Ratio* is a measure of effect size for *Two x Two Contingency Table* data which measures the odds of an event happening in the treatment group, relative to the odds of the same event happening in the control group (L'Abbé et al., 1987; Sokal and Rohlf, 1995). OR is calculated as:

$$OR = \frac{AD}{BC}$$

where A, B, C, and D are the number of observations in the treatment and control groups exhibiting and not exhibiting a response respectively. The natural log of RR is typically used in summary analyses. See *Effect Size; Rate Difference; Relative Risk; Two x Two Contingency Data*.

Omnibus Tests: See *Combining Probability Levels*.

One-Tailed Probability: For the transformation of several test-statistics to probability values (and vice-versa) a *One-Tailed Probability* is used. This way, one may distinguish between positive and negative effects.

Pooled Study Variance: The *Pooled Study Variance*, s^2_{pooled} , estimates the random component of variation in *Effect Sizes* in a *Random-Effects Model*. The calculation of s^2_{pooled} depends on the underlying structure of the data for the *Summary Analysis*, and it is an estimate of the amount of variance due to between-study differences. See *Categorical Summary Analysis*; *Continuous Summary Analysis*; *Random-Effects Model*; *Summary Analysis*.

Publication Bias: *Publication Bias* is the selective publication of articles showing certain types of results over those showing other types of results (Begg, 1994). The most commonly suspected publication bias is the tendency for journals to publish only those studies with statistically significant results. This type of *Publication Bias* (called the *File-Drawer Problem*: Rosenthal, 1979) will lead to an overestimate of the number of significant results on a given topic. See *Fail-Safe Tests*; *Funnel Plot*; *Normal Quantile Plot*; *Orwin's Fail-Safe Test*; *Rank Correlation Tests*; *Rosenthal's Fail-Safe Test*; *Weighted Histogram*.

Random-Effects Model: *Random-Effects Models* are *Summary Analyses* where studies differ from one another by sampling error (as in *Fixed-Effects Models*) and by a random component of variation in *Effect Sizes* between studies. This random component is called the *Pooled Study Variance*, s^2_{pooled} , and is incorporated into the formulation of weights for individual studies such that:

$$w_{i(\text{rand})} = \frac{1}{v_i + s^2_{pooled}}.$$

See *Fixed-Effects Model*; *Pooled Study Variance*.

Randomization Test: *Randomization Tests* generate a statistical distribution of test-statistics from the given data that are used in place of parametric distribution tests to determine the significance of that test-statistic. For each iteration of the randomization test, the original data are randomly reassigned to the structure defined by the model. A test statistic is then calculated using the randomly shuffled data. This process is repeated many times to generate a distribution of possible test-statistic outcomes. Finally, the actual test-statistic is compared to the frequency distribution of randomly-generated test statistics, and the proportion of randomly generated test-statistics more extreme than the actual value is taken to be the significance level for that data set (Edgington, 1987; Manly, 1997).

Rank Correlation Tests for Publication Bias: *Rank Correlation Tests for Publication Bias* are a statistical means of assessing *Publication Bias* that search for a relationship between the standardized Effect Size and sample size (Begg, 1994). Either Kendall's tau or Spearman's rho (Sokal and Rohlf, 1995) may be used in these tests. See *Publication Bias*.

Rate Difference: The *Rate Difference* is an effect size that measures the difference in rate scores between the treatment and control groups from a *Two x Two Contingency Table* (Normand, 1999). *Rate Difference* is calculated as:

$$RD = \frac{A}{A+C} - \frac{B}{B+D}$$

where A, B, C, and D are the number of observations in the treatment and control groups exhibiting and not exhibiting a response respectively. See *Effect Size; Odds Ratio; Relative Risk; Two x Two Contingency Data*.

Rate Ratio: See *Relative Rate*.

Regression: See *Continuous Summary Analysis*.

Research Synthesis: A *Research Synthesis* is a means of evaluating a hypothesis by summarizing the results from a series of previously published studies concerning that hypothesis. Though frequently narrative in nature, research syntheses can be quantitative and quite rigorous. See *Combining Probability Levels; Modern Meta-Analysis; Vote-Counting*.

Relative Odds: See *Odds Ratio*.

Relative Rate: The *Relative Rate* is a measure of effect size for *Two x Two Contingency Table* data which measures the rate of the treatment group relative to that of the control group (Greenland, 1987; L'Abbé et al., 1987). *RR* is calculated as:

$$RR = \frac{\frac{A}{A+C}}{\frac{B}{B+D}}$$

where A, B, C, and D are the number of observations in the treatment and control groups exhibiting and not exhibiting a response respectively. The natural log of *RR* is typically used in summary analyses. See *Effect Size; Odds Ratio; Rate Difference; Two x Two Contingency Data*.

Relative Risk: See *Relative Rate*.

Resampling Tests: *Resampling Tests* are nonparametric techniques that allow one to evaluate the significance of a given test-statistic. They are often useful when the data do not conform to the distributional assumptions of the parametric tests (Manly, 1997). *Resampling Tests* are performed by calculating a statistic from the original data, and

evaluating it by permuting the original data in some way, recalculating the test-statistics, and then repeating this procedure many times. This generates a distribution of possible test-statistics against which the original value is compared. See *Bias-Corrected Bootstrap Confidence Intervals; Bootstrap Confidence Intervals; Randomization Test*.

Response Ratio: The *Response Ratio*, $\ln R$, is the ratio of some measure of outcome in an experimental group to that of the control group (Hedges et al., 1999). The natural log of this measure is commonly used, and this *Effect Size* is calculated as:

$$\ln R = \ln\left(\frac{\bar{X}^E}{\bar{X}^C}\right) = \ln(\bar{X}^E) - \ln(\bar{X}^C)$$

See *Cohen's d; Effect Size; Glass' D; Hedges' d; Hedges' g; Means and Standard Deviation Data; Standardized Mean Difference*.

Risk Difference: See *Rate Difference*.

Risk Ratio: See *Relative Rate*.

Standardized Mean Difference: The *Standardized Mean Difference*, also known as a *Z-score*, is a simple measure of *Effect Size* found from the experimental and control group means and the sampling variance. See *Cohen's d; Effect Size; Glass' D; Hedges' d; Hedges' g; Response Ratio*.

Summary Analysis: A meta-analytic *Summary Analysis* is a means of statistically combining *Effect Sizes* from a set of individual studies to determine whether or not there is evidence supporting a particular hypothesis. For *Modern Meta-Analyses*, weighted *General Linear Models* are used in *Summary Analyses*.

Total Heterogeneity: A measure of the variation in *Effect Sizes* for a set of studies is the *Total Heterogeneity*. It is similar to the total sum of squares in an analysis of variance (ANOVA), and is calculated as:

$$Q_T = \sum_{i=1}^n w_i (E_i - \bar{E})^2$$

where n is the number of studies, and E_i and w_i are the *Effect Size* and weight for the i^{th} study. Q_T is tested against a χ^2 -distribution with $n - 1$ degrees of freedom to statistically assess whether there is heterogeneity among *Effect Sizes* for a given sample of studies. If data structure is incorporated into the summary analysis, Q_T can be partitioned into two quantities, one that describes the variation in *Effect Sizes* explained

by the model, Q_M , and one quantity that describes the variation in *Effect Sizes* not explained by the model, Q_E . This partitioning is similar to the partitioning of total sum of squares in and ANOVA, and is accomplished by:

$$Q_T = Q_M + Q_E$$

See *Categorical Summary Analysis*; *Continuous Summary Analysis*; *Effect Size*; *Error Heterogeneity*; *Model Heterogeneity*; *Summary Analysis*.

Two x Two Contingency Data: A data presentation commonly found from controlled experiments, where the results are categorical and can be placed in a two x two table. Most common in the medical literature, this data has two groups, treatment and control, and two outcomes, response and no response. Several different *Effect Sizes* can be calculated from data represented as *Two x Two Contingency Tables*. See *Effect Size*; *Odds Ratio*; *Rate Difference*; *Relative Risk*.

Vote-Counting: A quantitative method for synthesizing the results from a set of primary research studies by comparing the number of statistically significant outcomes to the number of nonsignificant outcomes (Light and Smith, 1971). The category with the highest proportion of studies is considered to be the most likely overall outcome for the data. See *Combining Probability Levels*; *Modern Meta-Analysis*; *Research Synthesis*.

Weighted Histogram: A *Weighted Histogram* is graphical method for exploring meta-analytic data where the height of the bar for each class in the histogram is made up of the combined weight of the studies which fall in that class, rather than their frequency (Greenland, 1987). The *Weighted Histogram* is often preferred to frequency histograms, because modern meta-analytic *Summary Analyses* are based on weighted models.

Within-Group Heterogeneity: For a *Categorical Summary Analysis* the variation in *Effect Sizes* for each group can be calculated. The *Within-Group Heterogeneity*, Q_{Wj} , describes the variation among *Effect Sizes* within each group, and may be tested against a χ^2 -distribution with $k_j - 1$ degrees of freedom (where k_j is the number of studies in the j^{th} group). See *Categorical Summary Analysis*; *Effect Size*; *Error Heterogeneity*; *Model Heterogeneity*; *Summary Analysis*; *Total Heterogeneity*.

Z-Score: See *Standardized Mean Difference*.

REFERENCES

- Adams, D. C., and C. D. Anthony. 1996. Using randomization techniques to analyze behavioural data. *Animal Behaviour*. 51:733-738.
- Adams, D. C., J. Gurevitch, and M. S. Rosenberg. 1997. Resampling tests for meta-analysis of ecological data. *Ecology*. 78:1277-1283.
- Antman, E. M., J. Lau, J. Jiminez-Silva, B. Kupelnick, F. Mosteller, and T. C. Chalmers. 1992. A comparison of results of meta-analyses of randomized controlled trials and recommendations of clinical experts: Treatments for myocardial infarction. *Journal of the American Medical Association*. 268:240-248.
- Balas, E. A., S. A. Baren, G. D. Brown, B. G. Ewigman, J. A. Mitchell, and G. T. Perkoff. 1996. Effect of physician profiling on utilization: Meta-analysis of randomized clinical trials. *Journal of General Internal Medicine*. 11:584-590.
- Bauchau, V. 1997. Is there a “file drawer problem” in biological research? *Oikos*. 79:407-409.
- Becker, B. J. 1994. Combining significance levels. In H. Cooper and L. V. Hedges (eds.). *The Handbook of research synthesis*. (pp. 215-230). Russell Sage Foundation, New York.
- Begg, C. B. 1985. A measure to aid in the interpretation of published clinical trials. *Statistics in Medicine*. 4:1-9.
- Begg, C. B. 1994. Publication bias. In H. Cooper and L. V. Hedges (eds.). *The Handbook of research synthesis*. (pp. 399-409). Russell Sage Foundation, New York.
- Begg, C. B., and J. A. Berlin. 1988. Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society, Series A*. 151:419-463.
- Begg, C. B., and M. Mazumdar. 1994. Operating characteristics of a rank correlation test for publication bias. *Biometrics*. 50:1088-1101.
- Berlin, J. A., N. M. Laird, H. S. Sacks, and T. C. Chalmers. 1989. A comparison of statistical methods for combining event rates from clinical trials. *Statistics in Medicine*. 8:141-151.
- Birnbaum, A. 1954. Combining independent tests of significance. *Journal of the American Statistical Association*. 49:559-574.

- Born, G. 1995. *The file formats handbook*. International Thomson Computer Press, London.
- Bradley, M. T., and R. D. Gupta. 1997. Estimating the effect of the file drawer problem in meta-analysis. *Perceptual and Motor Skills*. 85:719-722.
- Brind, J., V. M. Chinchilli, W. B. Severs, and J. Summy-Long. 1996. Induced abortion as an independent risk factor for breast cancer: A comprehensive review and meta-analysis. *Journal of Epidemiology and Community Health*. 50:481-496.
- Burton, P. J., and D. Mueller-Dombois. 1984. Response of *Metrosideros polymorpha* seedlings to experimental canopy openings. *Ecology*. 65:779-791.
- Bushman, B. J. 1994. Vote-counting procedures in meta-analysis. In H. Cooper and L. V. Hedges (eds.). *The Handbook of research synthesis*. (pp. 193-213). Russell Sage Foundation, New York.
- Chalmers, T. C. 1991. Problems induced by meta-analyses. *Statistics in Medicine*. 10:971-980.
- Chalmers, T. C., and J. Lau. 1993. Meta-analytic stimulus for changes in clinical trials. *Statistical Methods in Medical Research*. 2:161-172.
- Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. *Graphical methods for data analysis*. Duxbury Press, Boston.
- Clarke, G. M. 1998. Developmental stability and fitness: The evidence is not quite so clear. *American Naturalist*. 152:762-766.
- Cohen, J. 1965. Some statistical issues in psychological research. In B. B. Wolman (ed.). *Handbook of clinical psychology*. McGraw-Hill, New York.
- Cohen, J. 1969. *Statistical power analysis for the behavioral sciences*. Academic Press, New York.
- Connell, J. H. 1983. On the prevalence and relative importance of interspecific competition: Evidence from field experiments. *American Naturalist*. 122:661-696.
- Cooper, H. 1998. *Synthesizing research: A guide for literature reviews*. (3rd edition). Sage, Thousand Oaks.
- Cooper, H. and L. V. Hedges (eds.). 1994a. *The Handbook of research synthesis*. Russell Sage Foundation, New York.

- Cooper, H. and L. V. Hedges (eds.). 1994b. Research synthesis as a scientific enterprise. In H. Cooper and L. V. Hedges (eds.). *The Handbook of research synthesis*. (pp. 3-14). Russell Sage Foundation, New York.
- Coursol, A., and E. E. Wagner. 1986. Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology*. 17:136-137.
- Crowley, P. H. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. *Annual Review of Ecology and Systematics*. 23:405-447.
- Csada, R. D., P. C. James, and R. H. M. Espie. 1996. The “file drawer problem” of non-significant results: Does it apply to biological research? *Oikos*. 76:591-593.
- Dear, K. B. G., and C. B. Begg. 1992. An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science*. 7:237-245.
- DerSimonian, R., and N. Laird. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials*. 7:177-188.
- Dickersin, K., Y.-I. Min, and C. L. Meinert. 1992. Factors influencing publication of research results: Follow-up of applications submitted to two institutional review boards. *Journal of the American Medical Association*. 267:374-378.
- Dixon, P. M. 1993. The bootstrap and the jackknife: Describing the precision of ecological indices. In S. M. Scheiner and J. Gurevitch (eds.). *Design and analysis of ecological experiments*. (pp. 290-318). Chapman and Hall, New York.
- Dungan, M. L. 1986. Three-way interactions: Barnacles, limpets, and algae in a Sonoran desert rocky intertidal zone. *American Naturalist*. 127:292-316.
- Easterbrook, P. J., J. A. Berlin, R. Gopalan, and D. R. Mathews. 1991. Publication bias in clinical research. *Lancet*. 337:867-872.
- Edgington, E. S. 1972. An additive method for combining probability values from independent experiments. *Journal of Psychology*. 80:351-363.
- Edginton, E. S. 1987. *Randomization tests*. (2nd edition). Marcel Dekker, New York.
- Efron, B. 1982. The jackknife, the bootstrap, and other resampling plans. *Society of Industrial and Applied Mathematics*. CBMS-NSF Monograph 38.

- Fisher, R. A. 1928. *Statistical methods for research workers.* (2nd edition). Oliver and Boyd, London.
- Fisher, R. A. 1932. *Statistical methods for research workers.* (4th edition). Oliver and Boyd, London.
- Flather, M. D., M. E. Farkouh, J. M. Pogue, and S. Yusuf. 1997. Strengths and limitations of meta-analysis: Larger studies may be more reliable. *Controlled Clinical Trials.* 18:568-579.
- Fleiss, J. L. 1994. Measures of effect size for categorical data. In H. Cooper and L. V. Hedges (eds.). *The Handbook of research synthesis.* (pp. 245-260). Russell Sage Foundation, New York.
- Fowler, N. L. 1986. Density-dependent population regulation in a Texas grassland. *Ecology.* 67:545-554.
- Friedman, H. 1968. Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin.* 70:245-251.
- Galbraith, R. F. 1988. A note on graphical presentation of estimated odds ratios from several clinical trials. *Statistics in Medicine.* 7:889-894.
- Gibbons, J. G., I. Olkin, and M. Sobel. 1977. *Selecting and ordering populations: A new statistical methodology.* John Wiley and Sons, New York.
- Glass, G. V. 1976. Primary, secondary, and meta-analysis of research. *Educational Researcher.* 5:3-8.
- Goldberg, D. E., T. Rajaniemi, J. Gurevitch, and A. Stewart-Oaten. 1999. Empirical approaches to quantifying interaction intensity: Competition and facilitation along productivity gradients. *Ecology.* 80:1118-1131.
- Greenland, S. 1987. Quantitative methods in the review of epidemiologic literature. *Epidemiologic Reviews.* 9:1-30.
- Greenland, S. 1994. A critical look at some popular meta-analytic methods. *American Journal of Epidemiology.* 140:290-296.
- Greenland, S., J. J. Schlesselman, and M. H. Criqui. 1986. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *American Journal of Epidemiology.* 123:203-208.

- Gross, K. L., and P. A. Werner. 1982. Colonizing abilities of "biennial" plant species in relation to ground cover: Implications for their distributions in a successional sere. *Ecology*. 63:921-931.
- Gurevitch, J. 1986. Competition and the local distribution of the grass *Stipa neomexicana*. *Ecology*. 67:46-57.
- Gurevitch, J., and L. V. Hedges. 1993. Meta-analysis: Combining the results of independent experiments. In S. Scheiner and J. Gurevitch (eds.). *Design and analysis of experiments* (pp. 378-398). Chapman and Hall, New York.
- Gurevitch, J., and L. V. Hedges. 1999. Statistical issues in ecological meta-analyses. *Ecology*. 80:1142-1149.
- Gurevitch, J., L. L. Morrow, A. Wallace, and J. S. Walsh. 1992. A meta-analysis of competition in field experiments. *American Naturalist*. 140:539-572.
- Haddock, C. K., D. Rindskopf, and W. R. Shadish. 1998. Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods*. 3:339-353.
- Hauck, W. W. 1979. The large-sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics*. 35:817-819.
- Hauck, W. W. 1989. Odds ratio inference from stratified samples. *Communications in Statistics*. 18A:767-800.
- Hedges, L. V. 1981. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*. 6:107-128.
- Hedges, L. V. 1984. Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Education Studies*. 9:61-85.
- Hedges, L. V. 1992. Modeling publication selection effects in meta-analysis. *Statistical Science*. 7:246-255.
- Hedges, L. V. 1994. Statistical considerations. In H. Cooper and L. V. Hedges (eds.). *The Handbook of research synthesis*. (pp. 29-38). Russell Sage Foundation, New York.
- Hedges, L. V., J. Gurevitch, and P. S. Curtis. 1999. The meta-analysis of response ratios in experimental ecology. *Ecology*. 80:1150-1156.

- Hedges, L. V., and I. Olkin. 1980. Vote-counting methods in research synthesis. *Psychological Bulletin*. 88:359-369.
- Hedges L. V., and I. Olkin. 1985. *Statistical methods for meta-analysis*. Academic Press, New York.
- Hedges L. V., and I. Olkin. 2000. *Statistical methods for meta-analysis*. (2nd edition). (In Press).
- Hedges, L. V., and J. L. Vevea. 1996. Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics*. 21:299-332.
- Hedges, L. V., and J. L. Vevea. 1998. Fixed- and random-effects models in meta-analysis. *Psychological Methods*. 3:486-504.
- Hunt, M. 1997. *How science takes stock: The story of meta-analysis*. Russell Sage Foundation, New York.
- Hunter, J. E., and F. L. Schmidt. 1990. *Methods of meta-analysis: Correcting for sources error and bias in research findings*. Sage, Newbury Park.
- Hunter, J. E., F. L. Schmidt, and G. B. Jackson. 1982. *Meta-analysis: Cumulating research findings across studies*. Sage, Beverly Hills.
- Iyengar, S., and J. B. Greenhouse. 1988. Selection models and the file drawer problem. *Statistical Science*. 3:109-117.
- Johnson, C. R., and K. H. Mann. 1988. Diversity, patterns of adaptation, and stability of Nova Scotian kelp beds. *Ecological Monographs*. 58:129-154.
- Johnson, R. A., and D. W. Wichern. 1988. *Applied multivariate statistical analysis*. (2nd edition). Prentice Hall, Englewood Cliffs.
- Kleijnen, J., and P. Knipschild. 1992. Review articles and publication bias. *Arzneimittel-Forschung/Drug Research*. 42:587-591.
- L'Abbé, K. A., A. S. Detsky, and K. O'Rourke. 1987. Meta-analysis in clinical research. *Annals of Internal Medicine*. 107:224-233.
- Lane, D. M., and W. P. Dunlap. 1978. Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and*

- Statistical Psychology.* 31:107-112.
- Lau, J., E. M. Antman, J. Jimenez-Silva, B. Kupelnick, F. Mosteller, and T. C. Chalmers. 1992. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine.* 327:248-254.
- Levin, J., and B. Nalebuff. 1995. An introduction to vote-counting schemes. *Journal of Economic Perspectives.* 9:3-26.
- Light, R. J., and D. B. Pillemer. 1984. *Summing up: The science of reviewing research.* Harvard University Press, Cambridge.
- Light, R. J., and P. V. Smith. 1971. Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review.* 41:429-471.
- Louis, T. A., and D. Zelterman. 1994. Bayesian approaches to research synthesis. In H. Cooper and L. V. Hedges (eds.). *Handbook of research synthesis.* (pp. 411-422). Russell Sage Foundation, New York.
- Mann, C. 1990. Meta-analysis in the breech. *Science.* 249:476-480.
- Mann, C. 1994. Can meta-analysis make policy. *Science.* 266:960-962.
- Manly, B. F. J. 1997. *Randomization, bootstrap and monte carlo methods in biology.* (2nd edition). Chapman and Hall, New York.
- Mantel, N., and W. Haenszel. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute.* 22:719-748.
- McCreary, N. J., S. R. Carpenter, and J. E. Chaney. 1983. Coexistence and interference in two submerged freshwater perennial plants. *Oecologia.* 59:393-396.
- Microsoft Press. 1997. *Microsoft Excel97 developer's kit.* Microsoft Press, Redmond.
- Møller, A. P. 1997. Developmental stability and fitness: A review. *American Naturalist.* 149:916-932.
- Mosteller, F., and R. R. Bush. 1954. Selected quantitative techniques. In G. Lindsey (ed.). *Handbook of social psychology: Vol. 1. Theory and method.* (pp. 289-334). Addison-Wesley, Cambridge.

- Neter, J., W. Wasserman, and M. H. Kutner. 1989. *Applied linear regression models*. (2nd edition). Irwin, Homewood.
- Normand, S.-L. T. 1999. Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine*. 18:321-359.
- Orwin, R. G. 1983. A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*. 8:157-159.
- Osenberg, C. W., O. Sarnelle, and S. D. Cooper. 1997. Effect size in ecological experiments: The application of biological models in meta-analysis. *American Naturalist*. 150:798-812.
- Osenberg, C. W., and C. M. St. Mary. 1998. Meta-analysis: Synthesis or statistical subjugation? *Integrative Biology*. 1:37-41.
- 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.
- Palmer, A. R. 1999. Detecting publication bias in meta-analyses: A case study of fluctuating asymmetry and sexual selection. *American Naturalist*. (In Press).
- Pearson, K. 1904. Report on certain enteric fever inoculation statistics. *British Medical Journal*. 3:1243-1246.
- Pearson, K. 1933. On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*. 25:379-410.
- Platt, W. J., and I. M. Weis. 1985. An experimental study of competition among fugitive prairie plants. *Ecology*. 66:708-720.
- Pogue, J. M., and S. Yusuf. 1997. Cumulating evidence from randomized trials: Utilizing sequential monitoring boundaries for cumulative meta-analysis. *Controlled Clinical Trials*. 18:580-593.
- Pons, T. L., and J. van der Toorn. 1988. Establishment of *Plantago lanceolata* L. and *Plantago major* L. among grass. I. Significance of light for germination. *Oecologia*. 75:394-399.
- Potvin, C., and D. A. Roff. 1993. Distribution-free and robust statistical methods: Viable alternatives to parametric statistics? *Ecology*. 74:1617-1628.

- Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1989. *Numerical recipes in pascal*. Cambridge University Press, Cambridge.
- Raudenbush, S. W. 1994. Random effects models. In H. Cooper and L. V. Hedges (eds.). *The Handbook of research synthesis*. (pp. 301-321). Russell Sage Foundation, New York.
- Reed, D. C., and M. S. Foster. 1984. The effects of canopy shading on algal recruitment and growth in a giant kelp forest. *Ecology*. 65:937-948.
- Robins, J., S. Greenland, and N. E. Breslow. 1986. A general estimator for the variance of the Mantel-Haenszel odds ratio. *American Journal of Epidemiology*. 124:719-723.
- Rosenberg, M. S. 2000. The file-drawer problem revisited: A general weighted method for calculating fail safe numbers. *Submitted*.
- Rosenberg, M. S., D. C. Adams, and J. Gurevitch. 1997. *MetaWin. Statistical software for meta-analysis with resampling tests*. Version 1.0. Sinauer Associates, Sunderland.
- Rosenthal, R. 1978. Combining results of independent studies. *Psychological Bulletin*. 85:185-193.
- Rosenthal, R. 1979. The “file drawer problem” and tolerance for null results. *Psychological Bulletin*. 86:638-641.
- Rosenthal, R. 1984. *Meta-analytic procedures for social research*. Sage, Beverly Hills.
- Rosenthal, R. 1991. *Meta-analytic procedures for social research*. (revised edition). Sage, Newbury Park.
- Rosenthal, R. 1994. Parametric measures of effect size. In H. Cooper and L. V. Hedges (eds.). *The Handbook of research synthesis*. (pp. 231-244). Russell Sage Foundation, New York.
- Rosenthal, R., and R. L. Rosnow. 1985. *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge University Press, New York.
- Rosenthal, R., and R. L. Rosnow. 1991. *Essentials of behavioral research: Methods and data analysis*. (2nd edition). McGraw-Hill, New York.
- Rosenthal, R., and D. B. Rubin. 1978. Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*. 3:377-415.

- Schmidt, P., N. Scheer, and G. Berghaus. 1995. Marijuana use and driving: Simulated and actual driving tests for determination of the influence of marijuana: Literature review. *Kriminalistik*. 49:241-246.
- Schoener, T. W. 1983. Field experiments on interspecific competition. *American Naturalist*. 122:240-285.
- Shadish, W. R., and C. K. Haddock. Combining estimates of effect size. In H. Cooper and L. V. Hedges (eds.). *The Handbook of research synthesis*. (pp. 261-281). Russell Sage Foundation, New York.
- Smart, R. G. 1964. The importance of negative results in psychological research. *Canadian Psychology*. 5:225-232.
- Smith, M. L. 1980. Publication bias and meta-analysis. *Evaluation in Education*. 4:22-24.
- Smith, M. L., and G. V. Glass. 1977. Meta-analysis of psychotherapy outcome studies. *American Psychologist*. 32:752-760.
- Snedecor, G. W., and W. G. Cochran. 1989. *Statistical methods*. (8th edition). Iowa State University Press, Ames.
- Sokal, R. R., and F. J. Rohlf. 1995. *Biometry*. (3rd edition). Freeman, San Francisco.
- Sterling, T. D. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*. 54:30-34.
- Stimson, J. 1985. The effect of shading by the table coral *Acropora lyacinthus* on understory corals. *Ecology*. 66:40-53.
- Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams, Jr. 1949. *The American soldier, Vol. 1: Adjustment during army life*. Princeton University Press, Princeton.
- Tippett, L. H. C. 1931. *The methods of statistics*. Williams and Norgate, London.
- Turner, T. 1985. Stability of rocky intertidal surfgrass beds: Persistence, preemption, and recovery. *Ecology*. 66:83-92.
- VanderWerf, E. 1992. Lack's clutch size hypothesis: An examination of the evidence using meta-analysis. *Ecology*. 73:1699-1705.

- Vevea, J. L., and L. V. Hedges. 1995. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*. 60:419-435.
- Walden, J. 1986. *File formats for popular PC software: A programmer's reference*. John Wiley & Sons, New York.
- Wang, M. C., and B. J. Bushman. 1998. Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods*. 3:46-54.
- Wilkinson, B. 1951. A statistical consideration in psychological research. *Psychological Bulletin*. 48:156-158.
- Williams, P. A., E. H. Haertel, G. D. Haertel, and H. J. Walberg. 1982. The impact of leisure-time television on school learning: A research synthesis. *American Educational Research Journal*. 19:19-50.
- Winer, E., and P. M. Valley. 1988. Cracking the Lotus spreadsheet code. *PC Magazine*. December 13:297-314.
- Yusuf, S. 1997. Meta-analysis of randomized trials: Looking back and looking ahead. *Controlled Clinical Trials*. 18:594-601.
- Yusuf, S., P. Held, and C. Furberg. 1991. Update of effects of calcium antagonists in myocardial infarction or angina in light of the second Danish Verapamil infarction trial (DAVIT-II) and other recent studies. *American Journal of Cardiology*. 67:1295-1297.
- Yusuf, S., R. Peto, J. Lewis, R. Collins, and P. Sleight. 1985. Beta blockade during and after myocardial infarction: An overview of the randomized trials. *Progress in Cardiovascular Diseases*. 27:335-371.