Meta-Analysis Programs
by
Ralf Schwarzer

TABLE OF CONTENTS

Preface

This brochure not only gives a brief introduction to meta analysis, but also includes the manual for the meta-analysis programs. The program system consists of a number of routines that deal with probabilities, effect sizes d, and effect sizes r (correlations). The researcher can select these three options according to the kind of data base available. In addition, a data editor and several utilities are provided which allow one to make the necessary transformations. The program is initialized by typing META, and the user will be faced with the main menu where further information is supplied. Most procedures are self-explanatory, but before seriously running a meta-analysis it is recommended to consult the manual.

The program which was written in Turbo Pascal 5.0, is available either on 360K 5 1/4" disks formatted with MS-DOS 3.3 or on a 3 1/2" disk (720 K), and should run on all IBM-compatible microcomputers with 256K RAM or more. It is wise to copy all files to a hard disk. ***NOTE*** If you are using the 5 1/4" disk version the p-to-r.EXE file is on the second disk, and you must change the path within the program to make p to r conversions. This can be done by using the New Path option under the General Menu. Simply set the path to the drive where the second disk is located. (If the second disk is in drive B set the path to B:\.) After you are done, you must set the path back to the drive with the main disk. The *.EXE - files are necessary elements of the program, whereas the other files are example data sets. Files with the suffixes p, d or r are input for the meta-analyses for p-, d- or r-values, respectively. This manual is included as a WordStar 4.0 file and as a WordPerfect 5.0 file labelled MANUAL.WS4 and MANUAL.WP5 respectively.

There is also an ATARI ST version of the program designed by Uwe Czienskowski which requires some German. Inquiries should be made at the address below.

The program is activated by typing META (Enter).

Execution can be aborted by typing ^Break in case of an emergency but the menus include regular exit options, mostly ESC or Q.

Acknowledgements.

Several colleagues have been very helpful in detecting errors and suggesting improvements. They include Uwe Czienskowski, André Hahn, Dietmar Kleine, Anja Leppin, and Bettina Seipp. A few algorithms, especially for numerical transformations, have been taken from Mullen and Rosenthal (1985). The pull-down menus, the calculator, and a few other elements have been programmed with the help of source code provided by the Turbo Overdrive Package (Nescatunga Software). The editor is based on source code from Borland's Turbo Pascal Editor Toolbox.

License and Registration.

The program is not public domain but it is being distributed under the User Supported Software concept. Any user may copy and distribute the software as long as this is done free of charge. If you are a registered user you will receive update information. This is important because at the time you are reading this the program may already be obsolete. To obtain a copy of the latest version and to register send a formatted diskette with a US$10 check to the address below.

Disclaimer.

The program has been carefully tested with respect to published data in the meta-analysis literature. The results may sometimes differ slightly because the present algorithms are extremely accurate due to the software emulation of a numeric co-processor. But I make no warranty of any kind, and I shall not be liable for mistakes arising if this program fails to operate in the manner desired by the user. I will also not be liable for any damages arising from the use of or inability to use the program.

If you run into a problem, first consult this manual; only then request information from the author. Any further suggestions for improving the program are also welcome. Direct all correspondence to:

Prof. Dr. Ralf Schwarzer
Institut für Psychologie (WE 7)
Freie Universität Berlin
Habelschwerdter Allee 45
D - 14195 Berlin
G e r m a n y
FAX (Country Code + (0)30 838 5634

When this manual was written the program version 5.0 was valid. You may find more recent information in a text file called ReadMe.

Berlin, Fall 1989 *Ralf Schwarzer*

"The thirteenth stroke of a clock is not only false by itself, but casts grave doubts on the credibility of the preceding twelve." (Mark Twain, cit. Light & Pillemer, 1984, p. viii)

1. General Introduction to Meta-Analysis

1.1 Reviewing and integrating the research literature

Research results are known to be inconsistent. If 12 studies reach a specific conclusion there will be a thirteenth that draws the opposite one. As a researcher, one might feel tempted to conduct an additional study to solve the existing conflict but this would be premature and costly: (1) the information inherent in the available studies should be exhausted first, (2) the next researcher would again be confronted with 14 conflicting studies, and (3) if policy decisions were needed they would have to be postponed unduely. Instead, a detailed review is indispensable that allows to make a sound judgment on the average of the findings compiled so far and on the reasons for the prevailing inconsistency.

The research literature on most topics is sky-rocketing. Keeping pace with the overwhelming amount of incoming data in one's domain of interest is very time-consuming, if possible at all. For an evaluation of the "state of the art" of a specific field we often still have to depend on one or two literature reviews prepared by more or less ingenious scholars who have accumulated a great deal of studies, have summarized the heap of findings, and have drawn more or less valid conclusions, based on their own point of view. Narrative reviews, however, are considered to be subjective, to be scientifically unsound, and to be an inefficient way to extract useful information from the literature (Light & Pillemer, 1984, pp. 3ff).

The shortcomings of the traditional approach are evident, and therefore, attempts have been made to search for more promising methods of research integration and research evaluation. Glass (1976) was the first to introduce a novel perspective of dealing with the information overload problem by originating a comprehensive method that allows to estimate the average effect of treatments on outcome variables across a large number of studies. He coined the term "meta-analysis" and distinguished it from primary analysis and secondary analysis. Primary analysis is the original research which includes data collection, data processing, and publication of results, whereas secondary analysis requires a different investigator who, following the same research question, reanalyses the original data from either a different perspective or with different techniques. Meta-analysis, however, draws upon the summary statistics of a multitude of studies without having access to the original data.

According to Glass (1976) a meta-analysis aims at integrating a large number of results. Statistical methods are applied to summary statistics such as means or standard deviations found in the original studies, whereas raw data are subject to primary and secondary analyses only. The focus is not on statistical significance but on the size of treatment effects. Such an effect size is defined as the standardized mean difference between a treatment group and a control group in terms of an outcome variable. This unit-free measure of effect size ("d") had been proposed

previously by Cohen (1977, 2nd edition). The effect sizes of many studies are averaged in order to obtain an estimate of the most representative relationship between a treatment and an outcome.

The young history of meta-analysis started with two large-scale applications, one on the effectiveness of psychotherapy which covered 475 studies (Smith, Glass, & Miller, 1980), and one on the effect of school class size covering 77 studies (Smith & Glass, 1980). The ideas of Glass were received with great enthusiasm and were disseminated rapidly. Shortly after its introduction into research methodology the term meta-analysis was used in hundreds of theoretical articles as well as applications to advanced literature reviews. Although Glass had initially presented his programmatic outlook at a meeting of the American Educational Research Association, other disciplines such as psychology and the health sciences were immediately disposed to discuss, to apply and to improve this novel methodology.

One of the characteristics of Glass' approach lay in the inquiry of study features. Instead of simply averaging all available summary statistics, more emphasis was put on relating study features to outcomes and discovering the influence of specific factors inherent in the research design on the resulting effect sizes. If, for example, females and males differ with respect to their mean effect size, gender is identified as a moderator. This approach requires a coding scheme where as many study features as possible are entered in order to examine their potential moderating influence on effect sizes. Instead of integrating all summary statistics on the whole it is preferred to recognize subsets of data that produce distinctive mean effect sizes and, thus, proliferate additional knowledge about the research domain that was non-existent before applying meta-analysis. Study features could be the kind of population, age, source of publication, publication year, therapy type, technical quality of the study, etc. Glass suggested that these variables should be regressed on effect sizes in order to examine their interrelationships.

The definition of meta-analysis has become broader in recent years, and the label meta-analysis has been questioned by some authors who have coined alternative terms such as quantitative review, study synthesis, or research integration. Although these expressions might be somewhat more appropriate, the previous label has maintained its appeal and is used most often. In the present monograph the four terms are used interchangeably.

A classification of various meta-analytic strategies has been proposed by Rosenthal (1984). He distinguishes between meta-analytic comparisons and meta-analytic combinations of either probabilities or effect sizes (Figure 1).

	probabilities	effect sizes
comparison		
combination		

Figure 1
Classification of meta-analytic approaches (cf. Rosenthal, 1984)

The strategies proposed by Glass (1976) and many others include the fourth cell only: the integration of effect sizes. But it is also possible to integrate a number of significance tests -- a method discussed already some decades ago, but deemphazised by most authors. The comparison of either significance tests or effect sizes refers to the variation in the data instead of the average. Global comparison examines the degree of homogeneity, whereas specific comparisons are used when some studies are contrasted with others. Rosenthal broadens the scope of meta-analysis by pointing to stimulating alternatives of quantitative reviews and by providing handy formulas to carry them out.

1.2 Vote-counting: An inadequate attempt at quantification

Before advanced meta-analysis methodology was known, many reviewers have attempted to quantify the summary statistics found in the literature simply by counting positive and negative results. They established, for example, categories such as "significantly positive", "significantly negative" or "non-significant". If the findings supported the hypothesis a tally was credited to the first category, and if it contradicted the hypothesis a tally was assigned to the second category. Conclusions were then drawn by a straightforward comparison of the tallies. The shortcomings of this procedure can be illustrated by the following example of 8 experimental studies dealing with the impact of salt intake on blood pressure (Table 1a).

Only two out of eight findings (25%) favor the assumption that more salt intake increases blood pressure, two findings contradict this assumption, two findings tend to be positive but are not significant, and another two tend to be negative but are not significant either. From a traditional view, the conclusion would be to doubt any effect of salt intake on blood pressure because the majority of studies failed to support this assumption.

Table 1a
Fictitious example of the effect of salt intake on blood pressure

Study #	positive significant.	. positive non-significant	negative non-significant	negative significant
1	/			
2				/
3		/		
4				/
5	/			
6			/	
7		/		
8			/	

Table 1b
Fictitious example of the effect of salt intake on blood pressure (the coefficients are effect sizes)

Study #	positive significant	positive non-significant	negative non-significant	negative significant
1	.60			
2				-.10
3		.40		
4				-.10
5	.60			
6			-.30	
7		.40		
8			-.30	

From a meta-analysis viewpoint, however, the strength of association would be more meaningful (Table 1b). Computing the mean effect sizes of positive and negative findings and neglecting the level of significance yields .50 in favor of the hypothesis, but only -.20 opposed to it. The procedure has been simplified here for illustrative reasons, but the message is that vote-counting can be misleading. Hedges and Olkin (1980) have found that the statistical power of vote-counting is low, and that with increasing number of studies one fails to recognize a positive effect when in fact there is one (type II error). The number of tallies generates insufficient information: For example, to know that chemotherapy beats radiotherapy in 10 out of 12 cancer treatment studies is not to know whether chemotherapy wins by a nose or in a walkaway.

1.3 Combined significance tests

The idea of combining the results from independent tests is much older than the term meta-analysis, but only recently has this approach been popularized. It is possible to combine individual significance tests from a number of studies into one overall pooled test. The so-called Stouffer-method (see Rosenthal, 1984) which is briefly described here, is based on adding the standard normal deviation Z. Each individual probability *p* is converted to a Z score, and these Z scores are summed up across all studies. This sum is divided by the square root of the number of tests combined (k). The sum of normal deviates is itself a normal deviate and can be backtransformed into an overall p: The probability level associated with the sum of Z yields an overall level of significance. The complete procedure takes the form

$$p_i \rightarrow Z_i$$

$$Z_{(overall)} = \sum Z_i / \sqrt{k}$$

$$Z_{(overall)} \text{ ---> } p_{(overall)}$$

The advantage of such a procedure lies in the increased power of the overall comparison. If, for example, several tests consistently favor the research question but fail to reach the level of

significance, due to small sample sizes, the overall test would more easily become significant because the pooled sample size is much larger. On the other hand, if many studies are to be combined, sample size may become inflated, and the highly significant overall test does not provide useful information.

The procedure of combining significance tests shall be illustrated by a small example. Assume that four studies which were conducted to examine the effect of a new drug treatment on blood pressure, came up with conflicting conclusions. The meta-analyst tracked down the information provided in the following table.

Table 2
Summary statistics of the effect of drug treatment on systolic blood pressure (fictitious)

Study	Experimental Group			Control Group			*p (one-tailed)*
	M	*SD*	*N*	*M*	*SD*	*N*	
A	130	15	10	140	20	5	0.1662
B	120	12	40	140	15	20	0.0000023
C	140	20	30	150	25	30	0.04559
D	160	20	40	145	35	20	0.03981 $\rightarrow$ .960

We learn by this table that in Study A there was no significant difference between experimental and control groups, in Studies B and C the blood pressure readings of the experimental group were significantly decreased compared to the readings of the control group, and in Study D the opposite effect emerged. A vote count would only state that half of the studies confirmed the research question, one quarter disagreed with it, and one quarter reported no difference at all.

In order to obtain an overall result, the above-described method can be applied. For each *p* value the corresponding Z score is determined by looking up the table with standard normal deviates which can be found in every introductory statistics book. Notice that the *p* value of study D has to be reversed (1-p) because the mean difference is in the opposite direction. The four Z scores are Z_A=.9692, Z_B=4.5874, Z_C=1.6892, and Z_D=-1.7529. Their sum is divided by 2, and the result $Z_{(overall)}$ is 2.7438. The probability level of this total score can be tracked down in the statistics table again and yields $p_{(overall)}$=.003 (one-tailed). The conclusion is that the available studies, taken together, demonstrate that the new drug treatment significantly leads to lowered blood pressure levels. [In order to obtain the above results, the present computer program only requires input of the four *p* values; see data file ABCD.P on disk].

Procedures that combine significance tests have been criticized for several reasons. First, the focus is on an overall probability instead of on distributions. In meta-analysis, however, it is necessary to examine the variation of results because both positive and negative outcomes could

cancel each other out, and because conflicting results should represent a challenge to identify substantial sources of variation. Second, a *p* value just indicates the probability of an error in rejecting the null hypothesis if it were true (type I error). It does not provide, however, an estimate of the magnitude of treatment effects. In the present example, there is evidence that the drug is beneficial, but to what degree? Minute benefits might not outweigh the costs.

There are ways to obtain some of the missing information, however. First, variation in study outcomes can be assessed by a test of homogeneity. A formula is introduced by Rosenthal (1984, p. 77):

$$chi^2 = \sum (Z_i - \overline{Z})^2$$

$$df = k-1$$

For the present example, a chi^2 of 20.33 (ss) is computed with 3 degrees of freedom which indicates that the data are heterogeneous. This means that the meta-analyst should look for some systematic source of variation in reviewing the studies.

Second, as an estimate of magnitude an effect size r can be obtained by

$$r = Z_{(overall)} / \sqrt{N}$$

where *N* is the total sample size (Rosenthal, 1984, p. 31). In the present example, this estimate becomes r=.20 which gives an impression of how closely drug treatment versus no treatment is related to blood pressure.

Finally, there is a suggestion to deal with the "file drawer problem" (Rosenthal, 1984). It is possible that only few significant results are published and that the majority remains in the file drawers either because of reporting bias or publication bias. It is unknown how severe this bias really strikes, but it is possible to estimate the number of additional studies that would be required to reverse the overall *p* to a value higher than e.g., .05 (Rosenthal, 1984, p. 108; Wolf, 1986, p. 38):

$$N_{fs\ .05} = (\sum Z_i / 1.645)^2 - k$$

This number indicates how many no-effect findings would have to exist in the file drawers in order to invalidate a significant overall p. In the present example, 7 unpublished studies showing no effects would have to be tracked down in the file drawers to overturn the combined significance of p =.003 of the four available studies. With these 7 studies combined the overall p of all 11 studies would be .05.

In sum, the combination of significance tests can be a useful procedure if only probabilities are available to the meta-analyst. There are many further ways to compare and to combine tests (see Rosenthal, 1984). Effect sizes, however, are generally seen as a superior prerequisite for more

advanced meta-analyses (Fricke & Treinies, 1985; Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982; Light & Pillemer, 1984; Wolf, 1986).

1.4 Combination of effect sizes from experimental studies

The previous example data contain more information than has been used in our *p* values meta-analysis. Since the mean *M*, the standard deviation *SD* and the sample size *N* for both groups are provided, we can easily determine an effect size for each study. An effect size refers to the strength of a relationship between the treatment and the outcome variable. Most statistical tests such as the t - test can be transformed into an effect size *d* which, in this case, reflects the magnitude of the difference between two groups in standardized terms (Cohen, 1977). This coefficient is free of the original measurement unit. This is achieved by dividing the mean difference by the standard deviation of the measures. If, for example, one group has a mean of 25, the other has one of 20, and both have a standard deviation of 5, then the effect size is 5/5=1. In other words, the first group is one standard deviation superior to the second. [If no control group is available, effect sizes also can be determined as standardized pretest-posttest differences; see Kulik & Kulik, 1989, p. 254].

A common estimator of effect size is the standardized mean difference as proposed by Glass (1976). The mean M_c of a control group is subtracted from the mean M_e of an experimental group, and divided either by the control group standard deviation or by the pooled standard deviation of both groups:

$$g = (M_e - M_c) / SD$$

The latter strategy is generally preferred. In this case, *SD* is the square root of the weighted average of the two variances:

$$s^2 = ((n_e - 1)(s_e)^2 + (n_c - 1)(s_c)^2) / (n_e + n_c - 2)$$

In most of the literature the above effect size is called *d* instead of g, but in this monograph we follow the distinction made by Hedges and Olkin. The index *d* remains reserved for the unbiased effect size estimator d. Hedges and Olkin (1985, p. 80) show that *g* has a small sample bias. By defining a new estimator d, they remove this bias:

$$d = (1 - (3 / 4 * N - 9)) * g$$

Table 3 presents again the descriptive statistics of the four studies which examined the drug treatment effect on blood pressure. This time the focus is not on significance but on the magnitude of effects. Both effect sizes, *g* and d, are displayed (see data file ABCD.G on disk; *M*, *SD* and *d* need not be provided).

Table 3
Summary statistics of the effect of drug treatment on systolic blood pressure (fictitious)

Study	Experimental Group		Control Group		*g*	*d*

	M	*SD*	*N*	*M*	*SD*	*N*		
A	130	15	10	140	20	5	.5988	.5636
B	120	12	40	140	15	20	1.5315	1.5116
C	140	20	30	150	25	30	.4417	.4360
D	160	20	40	145	35	20	-.5794	-.5719

The average of effect sizes *g* turns out to be .498, the average of effect sizes *d* equals .485. If the effect sizes are weighted by their sample sizes, the population effect size would be delta=.472. All three parameters indicate consistently that the experimental group differs about half a standard deviation from the control group. So far the magnitude of the treatment effect has been estimated, which is more information than the mere significance we had found before.

However, is this result trustworthy? A closer look into the data shows that they are very heterogeneous. A test of homogeneity yields Q=25.06 (ss). The method provides more detailed information about the effect size variation, and before reporting more results the basic idea has to be described.

Although there are several ways to combine effect sizes from a series of experiments, the focus here will be on the "random effects model" (Hedges & Olkin, 1985). The variation of observed effect sizes can mirror partly the true variation of population effect sizes. If, for example, study characteristics such as different treatments, subject groups or stimulus materials account for some variation, then there may be several population effect sizes underlying the data. "Thus the observed variability in sample estimates of effect size is partly due to the variability in the underlying population parameters and partly due to the sampling error of the estimator about the parameter value" (Hedges & Olkin, 1985, p. 191). The program decomposes the observed effect size variance into both parts (see Formulas 7, 9 and 10, p. 194). The population variance is computed by subtracting the sampling error variance from the observed variance:

population variance = observed variance - sampling error

The percentage of observed variance made up by sampling error can be computed:

? % = sampling error * 100 / observed variance

This illustrates the degree of homogeneity or heterogeneity of the data set. If 100% of the observed variance is explained by sampling error (which is desired), the data are homogeneous. If, for example, 40% are explained by sampling error, then the residual variation of 60% is due to systematic factors. The meta-analyst should then look for moderator variables such as study characteristics which may account for this systematic variation. It is reasonable, therefore, to first examine the results from the "random effects model" to realize if the observed effect size variance is already exhausted. Otherwise the meta analyst may run several further analyses with specific data subsets according to some reasonable hypotheses. There is also a test of

homogeneity (p. 197) which serves as an additional indicator for a judgment on the heterogeneity status.

Within the "random effects model" a mean effect size delta is estimated (see p. 199). In case of homogeneity this would be the final result.

In the present example, the observed variance of effect sizes *d* is .73. This is decomposed into sampling error variance (.14) and population variance (.59). As can be seen, most of the observed variance remains unexplained because only 19% can be attributed to sampling error. The bulk of the observed variance is due to some systematic source of variance, and the meta-analyst should try to identify these sources, e.g., by looking at study features.

Sometimes it is helpful to group the effect sizes into subsets according to similarity. Then, one can inquire what these studies share in common in order to arrive at characteristics that had been overlooked previously when the coding scheme was set up. In this small example, three clusters emerged: Studies A and C were grouped together, and Study B as well as Study D represented "clusters" on their own. (For more details see chapter 2.4.8).

There is also a fail-safe *N* formula for *d* values developed by Orwin (1983) (see chapter 2.4.7). In this example, 6 no effect studies would have to exist in the file drawers in order to pull down the population effect size delta=.47 to a critical delta of .20. Which level of delta is seen as critical depends on the judgment of the meta-analyst.

1.5 Combination of effect sizes from correlational studies

Correlations are the best-known effect sizes. They describe the direction and strength of the relationship between two variables conveniently within a range of -1.0 and +1.0. Since most empirical studies focus on bivariate relationships and report correlations, suitable methods have been invented to combine effect sizes *r* from a multitude of studies. However, meta-analysis is more than simply averaging correlations, as will be shown in the following section.

Assume that we are interested in the association between drug abuse and delinquency, and that we have compiled eight studies where these two variables were related to each other. Six studies report Pearson correlations, one reports a t - test and the last a chi^2-test. The latter two were transformed into effect sizes *r* (see chapter 2.6 for formulas). Table 4 provides the fictitious data (see data file EIGHT.R on disk; the z values need not be provided).

Table 4
Relationship between drug abuse and delinquency

Study	Sample Size	r	Fisher's z
A	131	.51	.5627
B	129	.48	.5230
C	155	.30	.3095
D	121	.21	.2132
E	111	.60	.6931
F	119	.46	.4973
G	112	.22	.2237
H	145	.25	.2554

N = 1,023 $\bar{r}$ =.379 $\bar{z}$ =.410 $\rightarrow$ $\bar{r}$ =.388

In dealing with effect sizes r, the meta-analyst can take advantage of a convenient display technique which provides more information than a histogram: the stem-and-leaf display, introduced by Tukey (1977) within the framework of his "exploratory data analysis" (EDA). The present example yields the following display:

```
 -.9 |
 -.8 |
 -.7 |
 -.6 |
 -.5 |
 -.4 |
 -.3 |
 -.2 |
 -.1 |
 -.0 |
+.0 |
+.1 |
+.2 | 125
+.3 | 0
+.4 | 68
+.5 | 1
+.6 | 0
+.7 |
+.8 |
+.9 |
```

Figure 2
Stem-and-Leaf-Display for 8 Effect Sizes in Data Set: Eight.r

The effect sizes *r* range from .21 to .60 with .21, .22 and .25 in the first row, .30 in the next, .46 and .48 in the third, followed by .51 and .60. In case of many data this display gives an excellent overview. It is superior to a histogram because it provides the exact numbers.

Computing the average of a number of effect sizes *r* first requires a transformation of each individual *r* into Fisher's z. The average of the 8 z values is $\bar{z}$ = .41 which has to be back-transformed into $\bar{r}$ =.388. Some authors argue against the usual Fisher's z transformation and suggest to simply average the *r* values (e.g., Hunter et al., 1982). This would result in $\bar{r}$ =.379.

Another issue refers to weighting the effect sizes *r* with sample size. Most authors are in favor of this procedure because they believe that studies which employ large samples should get more credit than those which are based on small samples. Correlations are known to become more stable as sample size increases. A weighted average of the 8 effect sizes would be $\bar{r}_W$=.374 without Fisher's z transformation compared to $\bar{r}_W$=.383 with this transformation. The meta-analyst has to make transparent which option is selected. In case of doubt, the last coefficient seems to be the best estimator of the population effect size (the program provides all kinds of estimators).

The statistical procedures are mostly based on Hunter, Schmidt, and Jackson (1982) and Hedges and Olkin (1985). The aims of the meta-analysis of effect sizes *r* are the determination of (1) the population effect size and (2) the homogeneity. Given that all effect sizes belong to the same universe, it is assumed that each sample effect size *r* represents a deviation from its population effect size rho. Effects sizes of studies with large sample sizes should deviate less from the

population effect size than small *N* effect sizes. Therefore, in combining all effect sizes, it is fair to assign more weight to large *N* studies. Thus, the best estimate of the population effect size is the weighted average of all correlations.

A sampling error of an effect size from a study with a small *N* is relatively high, whereas a sampling error of an effect size from a large *N* study is small. When averaging correlations, one also averages the sampling errors. The sampling error of an individual correlation may be very high but the sampling error of the average correlation over many studies becomes small -- dependent on the total sample size. This happens because positive and negative sampling errors cancel themselves out in summation.
This is different, however, for the variance of correlations because the sign of the sampling error is eliminated by squaring the deviations. The variance across studies appears to be systematically larger due to sampling error. The relationship between sampling error variance and population variance is additive, and therefore, the observed variance of correlations can be regarded simply as the sum of both components.

The Schmidt-Hunter method allows a good approximation of the sampling error variance s^2_e by

$$s^2_e = ((1 - \bar{r}^2)^2 * k) / N$$

where $\bar{r}^2$ is the squared weighted mean of the effect sizes, k the number of studies, and *N* the total sample size (Hunter et al., 1982, 44).

After this is known one can easily determine the population variance s^2_{res} by subtracting the sampling error s^2_e from the observed variance s^2_r:

$$s^2_{res} = s^2_r - s^2_e$$

The population variance s^2_{res} is also called the residual variance, and its square root is called the residual standard deviation s_{res}.

It is desirable that all the observed variance s^2_r is accounted for by the sampling error s^2_e, and that the residual variance s^2_{res} would become zero. Then the "percentage of observed variance accounted for by sampling error" would be 100% as an indication of homogeneity. Very often, however, only a small percentage can be explained by artifacts, leaving a state of heterogeneity which requires further searches for moderator variables.

A population effect size can only be reliably interpreted if the underlying data set is sufficiently homogeneous. If most of the variance of the observed correlations is accounted for by sampling error, then this requirement is met. As a rule of thumb, Hunter et al. (1982) suggest, that at least 75% of the observed variance should be "explained" by artifacts such as sampling error. The remaining 25% of "unexplained" variance are seen as negligible: it would not be worth while to search for moderators.

The authors also provide a chi-square test of homogeneity with a very high statistical power. "Thus if the chi-square is not significant, this is strong evidence that there is no true variation

across studies; but if it is significant, the variation may still be negligible in magnitude" (Hunter et al., 1982, p. 47).

Finally, there is a third indicator of homogeneity which seems to be the most important one. The absolute amount of residual variance also counts. If there is almost no variance left after subtracting the sampling error variance, then homogeneity is assumed. ".. it is the actual amount of remaining variance that is important, not the percentage remaining..." (McDaniel, Hirsh, Schmidt, Raju, & Hunter, 1986). But how small has the actual amount of residual variance to be? As a rule of thumb one can consider a residual standard deviation as small if it does not exceed 25% of the population effect size (Stoffelmayr, Dillavou, & Hunter, 1983, p. 343).

In the present example, the observed variance of effect sizes is .01985, the variance due to sampling error is .00548, and the population variance is .01437 (all numbers based on algorithm with Fisher's z transformation). This means that only 28% of the observed variance is explained by sampling error. The residual standard deviation of .1199 is larger than ¼ of the population effect size. The test of homogeneity is chi^2 = 28 with 7 df which is significant. Thus, all three indicators agree upon heterogeneity.

In this situation, the meta-analyst goes back to the studies and searches for features that may have caused the systematic variation. Let us assume that no coding scheme has been used, and that the meta-analyst has no idea where to locate the substantial source of variation. Here it may be helpful to inspect groups of studies that share similar effect sizes. Grouping smaller numbers of studies into more homogeneous subsets is achieved by a cluster analysis. Moderator variables can be detected by employing this strategy and by inspecting the resulting subgroups. This can be called an "inductive moderator search". In addition to those potential moderators which are deducted from theory, empirical subsets of effect sizes may hint to a further moderator responsible for heterogeneous effect size distributions.

In the present example, two clusters emerged: Studies A, B, E and F versus Studies C, D, G and H (see chapter 2.5.7 for the cluster method). As a next step, these two data sets are separately analyzed in the same manner as the total data set above.

The first data set yields a population effect size of .51. The observed variance is .0059, the sampling error variance .0039, and the remaining population variance .002. This means that 66% of the observed variance is explained by sampling error which still looks heterogeneous. The two other indicators, however, suggest homogeneity. (See data file FOUR1.R on disk).

The second data set yields a population effect size of .25. The observed variance is .0013, the sampling error variance .0067, and no population variance remains. This means that 100% of the observed variance is explained by sampling error. All three indicators agree upon homogeneity. (See data file FOUR2.R on disk).

The two population effect sizes differ considerably, and the two population variances are much lower than the one for the total data set. This means that a moderator must exist (see chapter 2.5.4 for more details on moderator identification). The meta-analyst reviews the eight studies again and this time finds the clue: In the first set of studies, samples of adolescents have been

examined, whereas in the second set middle-aged adults have been the subjects under investigation. Young drug abusers who are jobless or lack money for other reasons more often have to rely on delinquent behavior to finance their habit than middle-aged adults with a secure job or other resources collected over the years.

1.6 Brief guidelines for meta-analysis

1.6.1 Finding and selecting studies

In some cases, relevant studies have been retrieved in the familiar and convenient environment of one's university library or have been located by chance. In order to avoid seriously flawed conclusions it is necessary to design the literature search in a more systematic manner.

First, the topic has to be clearly defined, and the criteria should be neither too broad nor too limited. In our work on social support and health, for example, we included social integration but excluded social conflict, and we included mortality but excluded psychological well-being (Schwarzer & Leppin, 1989a, 1989b; Leppin & Schwarzer, 1990). As can be seen, this remains a matter of judgment, and the meta-analyst has to make transparent how the research area has been defined and what the criteria for the selection of studies have been.

Searching the literature within a given range requires a systematic screening process which may start in one's university library but one also has to inquire for resources beyond it. Printed materials are, for example, the *Psychological Abstracts*, the *Sociological Abstracts*, and such periodicals as *Annual Review of Psychology*, and *Review of Educational Research*. It is indispensable to consult computerized data bases and retrieval systems such as SOCIAL SCISEARCH, ERIC, PsycInfo and MEDLARS. The German reader will find a good introduction to literature search in Fricke and Treinies (1985).

The references in these databases, however, are not the most recent ones. The meta-analyst should retrieve as much "gray literature" as possible, such as dissertations, conference papers and preprints of articles in order to minimize the "file drawer problem". The meta-analyst should not rely on published sources exclusively.

Not all of the studies retrieved can be used for a quantitative review. Some deal with the topic in a more theoretical or speculative manner without reporting effect sizes, probabilities, or summary statistics that could be converted into an effect size. The meta-analyst also will be reluctant to include an empirical study that is seriously flawed. If there is evidence that a study's effect size is based on a mistake, then its inclusion would pose a threat for the validity of the meta-analysis. This raises the difficult question of whether all available empirical studies that report the necessary information should be subject to the research synthesis, or whether only the methodologically best studies should be chosen. The latter position is taken by Slavin (1986). According to his "best-evidence approach" one should give little or no weight to studies of poor quality. This is in contrast to Glass, who is more liberal and favors the inclusion of all available studies (see Glass, McGaw, & Smith, 1981).

Studies vary in terms of quality, and this variation should be accounted for by a meta-analysis. The most appropriate way seems to be to make study quality a category of the coding scheme and

to relate it to the effect sizes. If it turns out to be a moderator then low-quality studies arrive at different conclusions than high-quality studies, and there would be no question which conclusions to follow.

1.6.2 Identifying and coding study characteristics

The database of a quantitative review has to be described in some detail. Therefore, the meta-analyst has to capture all the essential information such as publication year, publication source, instruments used, and type of sample. In addition, such information can be valuable when it comes to the search for moderators. If the results obtained are heterogeneous, the meta-analyst has to identify the reason for it and will turn to study features that might account for the unexplained variation. For the purpose of describing study features within categories, coding sheets are used. The beginning of such a coding scheme could look like the example in Table 5.

This is computerized and serves as a set of variables to be correlated with effect sizes. Some meta-analysts pay special attention to the detailed description of study features (Glass et al., 1981; Seipp, 1989a; Wittmann & Matt, 1986) but others deemphazise it. We have used a coding scheme with 23 categories and many subcategories (Schwarzer & Leppin, 1989a) but did not benefit from it very much. Creating such a database of study characteristics and relating it to effect sizes requires a great deal of time and effort and very often is not cost-effective. "Such coding can be 99 percent of the work in the research integration process. Yet this coding work may be entirely wasted" (Hunter et al., 1982, p. 32). If variation in the data turns out to be due to sampling error only, all the effort has been futile. Only in case of heterogeneity are the study features relevant for a moderator search. Sampling error is the major source of variation in meta-analytic data. According to our experience it amounts to 10-100%. On the other hand, a moderator usually accounts for a much smaller percentage -- if it can be identified at all. "Variance of study findings is only modestly predictable from study characteristics; in nearly all instances, less than 25% of the variance in study results can be accounted for by the best combination of study features..." (Glass et al,, 1981, p. 230).

For a parsimonious approach it is suggested to prepare a coding sheet with the available pertinent information, but before computerizing it, a meta-analysis should be run for the total as well as for some subgroups splitted by theoretical deliberations. Then, the state of affairs allows a better judgment of how many additional resources are worth investing in employing additional coders (for inter-coder reliability) and for computerizing the database. The narrower the defined topic and the smaller the resulting number of studies, the less vital the coding work is.

Table 5
Example for some categories of a coding scheme

Column	Feature
1 - 3	STUDY ID
4 - 5	YEAR OF PUBLICATION
01	1901
.	
89	1989
99	missing value
6 - 8	TYPE OF PUBLICATION
100	Journal
101	Health Psychology
200	Book
210	Book Chapter
300	Dissertation
400	Conference Paper
500	Research Report
9	GENDER OF SAMPLE
1	male
2	female
3	mixed
4	missing value
10 - 14	SAMPLE SIZE
etc.	

1.6.3 Reporting meta-analytic findings

It is desirable that a reader learns about all aspects of the meta-analytic process, but journal space is scarce and the meta-analyst has to face these constraints by selecting the most crucial findings. The author should make all decisions transparent such as the definition of the research area, the time period covered, the countries or languages included, the number of studies located compared to those selected, and the selection criteria. The database has to be described, including publication sources, samples studied, and instruments and statistics used. It is also useful to illustrate the total set of effect sizes visually by a histogram, a stem-and-leaf display or a funnel display (see Light & Pillemer, 1984).

The minimum information to be provided on the meta-analytic results include the number of effect sizes, the number of subjects, the population effect size, and measures of variation or homogeneity -- individually for each result. Usually, as a first step, all available effect sizes are subject to an overall analysis which turns out to yield heterogeneity and thus, requires a number of subanalyses. A meta-analysis, therefore, is in fact a series of meta-analyses of subordinate data sets. A result table for two such meta-analyses could look like the following (Table 6).

Table 6
Example: Gender-specific relationships between anxiety and self-concept

Study feature	k	N	$\bar{r}_w$	s_o	s_e	s_p
females	35	1,302	.42	.060	.040	.020
males	42	2,150	.37	.080	.020	.060

In this example, correlation studies have been investigated. There were 35 effect sizes *r* for a total of 1,302 females, which yield a weighted average (population effect size) of .42, an observed variance of .06, variance due to sampling error of .04, and remaining population variance of .02. That means that one third of the variance is not explained by sampling error.

More detailed result tables could be used which also include unweighted mean effect sizes, their credibility interval, and further indices of homogeneity (see Kleine, 1990; Seipp, 1989a, 1989b).

An appendix should be provided that lists all studies reviewed. If space is available, another appendix should be given that comprises all effect sizes cross-referenced with studies. This typically requires lots of remarks and footnotes because many effect sizes are the result of averaging and transformations. The authors of the original studies sometimes have a hard time recognizing effect sizes which they never had published in this way.

1.7 Criticism of meta-analysis

Meta-analysis has been criticized in a number of ways (see Glass, McGaw & Smith, 1981, for comments that the research synthesis is too dependent on published studies, while unpublished studies remain). First, it has been argued hidden in the "file drawers". Published findings are highly selective and, therefore, do not represent the "state of the art" accurately. Researchers are inclined to report findings that are statistically significant and to neglect those that are not. In addition to this reporting bias, journal editors tend to reject submitted manuscripts which do not include statistically significant findings, due to the high competition for journal space. This implies a publication bias. Manuscripts which were either rejected or not submitted in the first place may disappear in the "file drawers" or may be distributed as conference papers or "gray report literature" only. This *file drawer problem* leads to an overestimation of population effect sizes.

It is not proven, however, in how far this inference is valid. If, for example, an outcome statistic fails to become significant, this might be due to small sample size but the corresponding effect size may be large; if, on the other hand, statistics become significant, this might be attributed mainly to large sample size, and the corresponding effect size may be negligible. It is possible, therefore, that some findings in the "file drawers" include higher effect sizes than those in published studies. Nowadays, journal editors have become more aware of effect sizes and more

often disregard minute effects that are associated with "significant" findings based on hundreds or thousands of subjects.
The file drawer problem can not be understood as a criticism pertinent to meta-analysis only. Instead, it is a more general issue that applies to all kinds of literature reviews, qualitative as well as quantitative ones. Only those data that are available can be integrated . Meta-analysis achieved a great deal in having made this difficulty transparent and in searching for ways to deal with it. One suggestion made is to compute a "fail-safe *N*" (Rosenthal, 1984), which is the estimated number of non-significant results in the file drawers that would be required to turn a significant meta-analytic finding into a non-significant one. The fail-safe *N* is a unique way of pointing to the file drawer problem and making it more transparent but it cannot solve the problem itself. The meta-analyst has to consider as many unpublished sources of information as possible to cope with the imbalance of studies being either published or unavailable.

Second, meta-analysis has been criticized for paying too much attention to *studies of low quality*. If the entire pool of available studies serves as the object of an effect size integration, poor studies are assigned the same weight as good studies. If, for example, the treatment groups were not selected randomly, if inappropriate statistics were applied, or if the psychometric properties of the instruments were not satisfactory or were not reported at all, then such studies hardly deserve to be pooled with excellent studies. This problem can be partly solved if the meta-analyst includes the technical quality as part of the study features in the coding scheme. Thus, the influence of quality on effect sizes could be examined. If quality turns out to be a moderator, the subgroup results of low and high quality studies could be either separately reported or the meta-analyst could assign different weights to studies of different quality. Whatever is preferred, the meta-analyst has to make the judgments and decisions transparent to the reader. What has formerly come up as a criticism of meta-analysis has turned into an advantage: explicit methods of dealing with differences in study quality have been designed which are clearly superior to the more implicit ways in which narrative reviews deal with this issue.

Third, the reproach has been made that meta-analysis *mixes "apples" and "oranges"*. In a sense this is true because all reviews have to cover a variety of studies which necessarily differ in a number of characteristics. Reviews need to have an adequate scope in order to arrive at meaningful conclusions about a research domain. If "apples and oranges" are mixed the conclusions necessarily refer to "fruit". This issue remains to be a matter of degree. At the one extreme, only those studies would be integrated that represent solely a very narrow research question, for example inquiring about the influence of continuous teacher praise on first grade girls in four hours of arithmetic lessons. Each study here can be seen as an exact replication of the previous one, as it is the case in a planned series of experiments. Thus, merely very few studies can be summarized. At the other extreme, integration would aim at the entire collection of studies which refer to a broad research question, such as examining the impact of risk behaviors on illness. Smoking, drinking, poor nutrition, lack of exercise etc. are treated as risk behaviors, and blood pressure, time in hospital, self-reported disease symptoms, amount of medication etc. are taken together as indicators of illness. In their psychotherapy outcome study, for example, Smith and Glass (1977) mixed results from very different types of therapy and from very different kinds of outcome variables.

The question is how to deal with scope if the research question appears to be too broad for meaningful conclusions. The solution to this issue lies in establishing a hierarchy of constructs differing in scope. As Hedges (1986, p. 359) suggests, the meta-analyst may start out with a more general research question using broad constructs but then should turn to subordinate constructs which are more narrow in scope in order to avoid premature or overgeneralized conclusions. Findings related to disparate narrow constructs could cancel each other out and therefore have to be considered separately. A meta-analysis can arrive at a complex mosaique with different levels of different scopes. Awareness of the apples-and-oranges problem may help to identify a pattern of appropriate constructs. One set of studies would be nested under one specific category, while another set would be nested under another category.

Fourth, meta-analysis has been criticized for *lumping together nonindependent results*. When multiple outcomes are derived from the same studies, those studies are given more weight than others, and sample size would increase artificially. For example, in their psychotherapy outcome meta-analysis, Smith and Glass (1977) included 475 studies, but some of their computations were based on 1,766 effect sizes.

Usually, the number of effect sizes is much higher than the number of studies because authors report more than one summary statistic in their articles. The construct under investigation may be operationalized by several indicators each of which produces one effect size, or a variable may be measured at more than one point in time, or the author has conducted more than one experiment with the same subjects.

This issue refers to the unit of analysis. Shall we treat a single study or a single effect size as the unit of analysis? Facing this problem in a specific situation, the meta-analyst has to make a good judgment about the expected amount of bias which would affect the data, and then decide how to proceed. Most meta-analysts today do not agree with the position of Glass who liberally relied on the effect size as the unit of analysis, no matter how many studies they were derived from. Instead, it is suggested to use either the study or the sample within a study as the unit of analysis (see Kulik & Kulik, 1989). The latter procedure was also preferred in our own work (Kleine & Schwarzer, 1991; Leppin & Schwarzer, 1990; Schwarzer & Leppin, 1989a, 1989b). If, for example, a study included 100 females and 300 males, and separate statistics were available for both samples, two strategies were employed: (1) for the overall integration a weighted mean effect size was computed, and (2) for the gender-specific meta-analysis both effect sizes were taken separately. If, however, data were obtained at three points in time, the three effect sizes for females were averaged as well as the three effect sizes for males. An alternative to averaging would be to select one of the effect sizes randomly.

If multiple indicators had been used to measure a construct, it also might be adequate to average their effect sizes. In this situation, Rosenthal and Rubin (1986) suggested to compute mean effect sizes after accounting for the intercorrelation of these indicators, and they provided formulas for this purpose.

Another problem arises if several treatment groups are compared to a single control group, and the computation of all effect sizes draws multiply from this control group. Most meta-analysts neglect this very specific kind of dependency but Hedges and Olkin (1985) discuss possibilities to even account for this small bias statistically.

Another source of nonindependence lies in multiple publications by the same author. Study II may rely on the same sample as Study I did, or the same instruments and designs may have been used over and over again. In treating these studies as independent in a meta-analysis, the work of one author or research group receives the lion's share of effect sizes, and thus, may bias the results in favor of their outcomes. If such a bias becomes apparent, the meta-analyst is advised to reduce dependency by averaging those studies or defining other units of analysis.

In sum, the four major criticisms of meta-analysis have stimulated a great deal of creative thinking as well as the invention of novel strategies to cope with them. The points of criticism do not hold: either they refer generally to all kinds of literature reviews or they point to an inappropriate use of meta-analysis. Any statistical procedure can be misused, and meta-analysis is no exception. Today there is no longer a serious criticism that rejects meta-analysis methodology per se but there are warnings not to misuse the quantitative approaches to research synthesis.

2. The Meta-Analysis Programs

2.1 The Main Menu

When the program is called the main menu pops up. It contains five pull-down submenus (Figure 3).

GENERAL	*p* VALUES	*d* VALUES	*r* VALUES	UTILITIES
Editor	Meta-analysis	Meta-analysis	Meta-analysis	Conversion r
Directory	r-File	Cluster	Cluster	Effect Size d
Change Dir		r-File	StemLeaf	Signif. of Corr
Info				Weighted M,V,C
Calculator				t-Tests
Quit				

Figure 3
The Main Menu

Before starting serious work it is suggested to play around with the example files in order to become acquainted with the program system. Select the option "Directory" to view the available example files. Their suffix indicates whether they are set up for *p* values, *d* values or *r* values meta-analyses. The Editor allows to list and to modify the data as well as to create new data sets (see chapter 2.2). "Change Dir" gives access to other directories, but parts of the program rely on subroutines stored in the current directory. It is suggested, therefore, to reserve one directory on the harddisk for all meta-analytic work.

The option "Meta-Analysis" is provided three times, for p-, d-, and r-values, respectively. All other options are designed for additional features, such as cluster analyses or stem-and-leaf displays.

The Utilities menu does not require any data files, but manual input of statistics or coefficients such as t, F, chi^2, r, to be transformed, to be weighted, or to be compared.

2.2 Creating and modifying data files with the EDITOR

The present program system is not designed to accept input data from the keyboard. Instead, all data have to be typed into specific files on disk. The files are ASCII files, which implies that any appropriate editor can be used for data input, such as SideKick, WordStar (non-document mode), or the Turbo Pascal editor. In fact, the EDITOR provided with this set of programs is based on Borland's Turbo Pascal Editor Toolbox. Familiar keystrokes can be applied, therefore, when using this EDITOR. From the Main Menu select File/Editor. Those who are absolute novices will find the minimum information, how to save and how to quit, at the bottom of the screen. Do not use a Carriage Return (CR or Enter) after entering the last value.

There are three data structures depending on the strategy of meta analysis preferred, those for *p, d* and *r* values.

2.2.1 Probabilities

Example 1 for p values, structured as: # n p

```
9    100   0.005
2    60    0.0005
7    31    0.025
4    22    0.005
1    15    0.389
```

Preparing a data file for a meta analysis with probabilities requires the input of three columns: The study ID (any numerical identification that you have assigned to an empirical study), the sample size N, and the exact one-tailed probability p. [If sample size is not available assign any number, but it has to be the same for all studies.] In the above example, all research results have been in the same direction, favoring the research question. Assume, however, that the first two studies yielded significant results in the opposite direction, not supporting the research question. This will be the normal case in meta-analyses. To account for this, simply assign a minus sign to those *p* values, as demonstrated in the following example:

Example 2 for p values, structured as: # n p

```
9   100   -0.005
2    60   -0.0005
7    31    0.025
4    22    0.005
1   15    0.389
```

Of course, the correct notation would be p=.995 for the first, and p=.9995 for the second study, but for ease of use the program is designed to handle both ways. There are mostly some results that are in favor of the researcher's assumptions and others that are not. Either apply the correct

probabilities (1-p) in this instance, or simply assign minus signs to the *p* values of those results that do not support the research question.

2.2.2 Effect sizes *d* or *g*

Example for d values, structured as: # n_e n_c g r_{tt}

8	*100*	*100*	*1.2*	*0.9*
2	*200*	*200*	*0.6*	*0.8*
12	*300*	*300*	*0.9*	*0.8*
4	*400*	*400*	*0.8*	*0.9*

There is some confusion with respect to *g* or *d* as labels for effect sizes. They differ slightly. This manual refers to a *g* value as the standardized difference between two means, and to a *d* value as the corresponding unbiased effect size estimator (see below).

A data file for a meta analysis with effect sizes *d* requires five columns, as indicated. The sample sizes for the experimental group n_e and for the control group n_c are followed by the effect size. The estimator of effect size is based on the standardized mean difference as proposed by Glass (1976). The mean M_c of a control group is subtracted from the mean M_e of an experimental group, and divided by the pooled standard deviation of both groups:

$$g = (M_e - M_c) / SD$$

To obtain these *g* values one may use the subroutine "Effect sizes d" from the Utilities Menu. These *g* values will later be transformed by the meta-analysis subroutine into *d* values. According to Hedges and Olkin (1985) *d* is an unbiased estimator of the effect size. For details, see chapter 2.3 on the combination of effect sizes from a series of experiments. For technical reasons, the data set has to include a column with reliabilities r_{tt}. If not available, insert unity to assume unbiased measures. There is no specific treatment of missing values but coefficients of 0 are automatically transformed to 1.0.

2.2.3 Effect sizes r

Example for effect sizes r, structured as: # n $r_{(xy)}$ $r_{(xx)}$ $r_{(yy)}$

11	*100*	*0.23*	*0.9*	*0.8*
20	*200*	*0.20*	*0.9*	*0.7*
34	*300*	*0.19*	*0.8*	*0.9*
42	*400*	*0.21*	*0.8*	*0.8*

Correlations are in the third column, reliabilities for variables x and y in the fourth and fifth column, respectively. All five columns are required. If no reliabilities are available, these columns should contain a vector of 1.0. If, however, only some studies provide reliability information but others do not, the correction for attenuation procedure still may produce useful results. The program contains a specific missing value treatment based on the artefact distributions described by Hunter, Schmidt and Jackson (1982) and allows the necessary

corrections in case of incomplete reliability information. In this case, all available reliabilities have to be entered and all missing values have to be coded as 0 in columns 4 and 5, but for technical reasons, at least one reliability coefficient per column [e.g., 1.0] has to be provided in order to run the program.

Do not use a Carriage Return (CR or Enter) after the last value entry in the data set, because this would be interpreted as an additional empty data record.

2.3 Combination of probabilities

If empirical studies do not provide effect sizes or appropriate statistics which allow for transformation into effect sizes, the meta-analysis can be reduced to a combination of probabilities. *P* values are usually reported in the literature or can be easily obtained from statistical tables.

This program uses the Stouffer method of integrating one-tailed exact *p* values after transformation to the normal distribution *Z* (see Rosenthal, 1984). For each *p* the program computes the corresponding normal distribution *Z* and provides an unweighted as well as a weighted average Z. Their corresponding significance *p* is the main result of this kind of meta-analysis. The corresponding effect size *r* is computed as r=Z/sqrt(N). For the transformation of *p* to *Z* see Mullen and Rosenthal (1985, p. 134).

Figure 4 displays the results for the data set "Example1.p" which was shown in Chapter 2.2. Combining the probabilities of these five studies leads to a highly significant result (*p*=.0000022) indicating the very small likelihood that a *Z* this large or larger could be obtained if the null hypothesis were true. Although one of the five studies was not significant, the null hypothesis can be rejected for all five studies combined.

****
Results of Meta-Analysis for Probabilities

****
Filename : example1.p
Number of Studies: 5
Total Sample Size: N = *228*

Unweighted average Z-value Z = *4.7785*
Significance (one-tailed) p = *0.0000013*
Corresponding effect size r = *0.3164661*

Weighted average Z-value *Z = 4.6673*
Significance (one-tailed) p = *0.0000022*
Corresponding effect size r = *0.3090964*
Fail-Safe N*(p=.05) = 37.19*
Fail-Safe N*(p=.01) = 16.03*

Test of homogeneity *Chi-Square = 5.1911*
Degrees of freedom *df = 4*
Significance p = *0.268246*

****

Figure 4: *Results for data set "Example1.p"*

The Fail-Safe *N* informs about the number of nonsignificant "file drawer studies" necessary to invalidate a significant overall result at a certain predefined level. In this example 16 nonsignificant findings would be required to bring the overall *p* of .0000022 above a critical *p* of .01, and 37 nonsignificant findings would be required to bring it above .05. The formulas can be found in Rosenthal (1984, pp. 89, 94, 108).

```
**********************************************************
 Results of Meta-Analysis for Probabilities
**********************************************************
Filename : example2.p
Number of Studies :   5
Total Sample Size:                N = 228

Unweighted average Z-value  Z = -.4691
Significance (one-tailed)         p = 0.3195105
Corresponding effect size         r = -0.0310648

Weighted average Z-value          Z = -2.6980
Significance (one-tailed)         p = 0.0034877
Corresponding effect size         r = -.1786797
Fail-Safe                   N(p=.05) =  -4.59
Fail-Safe                   N(p=.01) =  -4.80

Test of homogeneity         Chi-Square  = 27.8055
Degrees of freedom          df = 4
Significance                p = 0.000014
**********************************************************
```

Figure 5: *Results for data set "Example2.p"*

In the second example (Figure 5) two *p* values were entered with a minus sign to indicate that their significant results were in the opposite direction (see chapter 2.2). Since these two entries made up the majority of the sample (n=160 out of N=228), there is a considerable difference between the unweighted (p=.319) and the weighted solution (p=.003). According to the first one, the combination of all five studies does not justify the rejection of the null hypothesis. According to the second one, combining all studies suggests rejection, favoring the unexpected direction because the *Z* value is negative.

Under some circumstances the Fail-Safe *N* can become negative and should be ignored. The test of homogeneity indicates high heterogeneity in this second example which derives from the minus sign manipulation leading to an extreme pattern of results.

Meta-analyses of effect sizes are superior to combinations of probabilities. The present program system also provides additional information for *r* value data sets such as cluster analyses and stem-and-leaf displays. It is, therefore, suggested that, after running the above analyses, the probabilities data set is converted to another data set of *r* values. Choose "p values/r-files" from the main menu to create a corresponding data set of effect sizes r. Then, use this new file with the suffix ".RRR" as input for the *r* values procedures. However, keep in mind that transformations of this kind may be biased sometimes (see Kulik & Kulik, 1989, p. 254).

2.4 Combination of effect sizes taken from a series of experiments

2.4.1 Introduction

This chapter describes how to run a meta-analysis for a number of experimental studies. It refers to the book by Hedges and Olkin (1985) which should be the main source for further consultations about this method. (For the German readership the book by Fricke and Treinies, 1985, is recommended). In the following section, some paragraphs are repeated from the general introduction of this manual.

A common estimator of effect size is the standardized mean difference as proposed by Glass (1976). The mean M_c of a control group is subtracted from the mean M_e of an experimental group, and divided either by the control group standard deviation or by the pooled standard deviation of both groups:

$$g = (M_e - M_c) / SD$$

The latter strategy is generally preferred. In this case s is the square root of the weighted average of the two variances:

$$s^2 = ((n_e - 1)(s_e)^2 + (n_c - 1)(s_c)^2) / (n_e + n_c - 2)$$

In most of the literature the above effect size is called *d* instead of g, but in this monograph we follow the distinction made by Hedges and Olkin. The index *d* remains reserved for the unbiased effect size estimator (see below). To easily compute these effect sizes *g* or d, the user can select the subroutine "Effect sizes d" from the Utilities Menu. Input of means and standard deviations (optional: variances) yield the desired effect sizes. Note, that the subsequent meta-analysis routine requires a data file with *g* values which will automatically be transformed to *d* values.

Hedges and Olkin (1985, p. 80) show that *g* has a small sample bias. By defining a new estimator d, they remove this bias:

$$d = (1 - (3 / 4 * N - 9)) * g$$

Although there are several ways to combine effect sizes from a series of experiments, the focus here will be on two methods: a) the weighted linear combination of estimators from different studies, and b) the "random effects model". These methods provide accurate estimates if the single effect sizes do not exceed the absolute value of 1.5 and if the respective sample sizes for each effect size are at least n=10.

2.4.2 The weighted integration method

The variance of a single effect size depends on its sample size (Hedges & Olkin, 1985, p. 86, Formula 14):

$$\text{est. } s^2(d_i) = \frac{n_e + n_c}{n_e * n_c} + \frac{d^2}{2(n_e + n_c)}$$

Effect sizes based on larger samples are more precise and, therefore, deserve more weight in a meta-analysis. The weighted integration method weighs -- in a complicated manner -- each effect size with its variance (p. 112, Formula 8).

$$d+ = \frac{\sum (d_i / \text{est. } s^2(d_i))}{\sum (1 / \text{est. } s^2(d_i))}$$

Example data are given by the authors (Hedges & Olkin, 1985, p. 114) which are saved as a disk file "Hedges1.d" for the convenience of the META user. Note, that the authors provide *d* values on page 114 but that the example file contains corresponding *g* values, necessary as META input (taken from Hedges & Olkin, p. 25).

A test of homogeneity (p. 123) serves to examine if all effect sizes can be considered to be samples from a common population of effect sizes. If it becomes significant, then the data set is rated as being heterogeneous.

2.4.3 The "random effects model"

The previous method is applied with the underlying assumption that the data set is homogeneous, i.e., all studies have the same population effect size, and the observed effect sizes differ only as a result of sampling error. In many situations, however, this is not the case, and therefore, a more general approach is preferred. The variation of observed effect sizes can mirror partly the true variation of population effect sizes. If, for example, study characteristics such as different treatments, subject groups or stimulus materials account for some variation then there may be several population effect sizes underlying the data. "Thus the observed variability in sample estimates of effect size is partly due to the variability in the underlying population parameters and partly due to the sampling error of the estimator about the parameter value" (Hedges & Olkin, 1985, p. 191). The program decomposes the observed effect size variance into both parts (see Formulas 7, 9 and 10, p. 194). The population variance is computed by subtracting the sampling error variance from the observed variance:

population variance = observed variance - sampling error

The percentage of observed variance made up by sampling error can be computed:

? % = sampling error * 100 / observed variance

This illustrates the degree of homogeneity or heterogeneity of the data set. If 100% of the observed variance is explained by sampling error (which is desired), the data are homogeneous. If, for example, 40% are explained by sampling error, then the residual variation of 60% is due to systematic factors. The meta-analyst should then look for moderator variables such as study characteristics which may account for this systematic variation. It is reasonable, therefore, to

first examine the results from the "random effects model" to realize if the observed effect size variance is already exhausted. Otherwise the meta-analyst may run several further analyses with specific data subsets according to some reasonable hypotheses. There is also a test of homogeneity (p. 197) which serves as an additional indicator for a judgment on the heterogeneity status.

Within the "random effects model" a mean effect size delta is estimated (see p. 199) including confidence intervals. In case of homogeneity this would be the final result.

2.4.4 Interpretation of computer output: An example

Example data are given by the authors (Hedges & Olkin, 1985, p.195) which are saved as a disk file "Hedges2.d" for the convenience of the META user. This is a data set with 11 records, each containing the study number, the sample size of the experimental group, the sample size of the control group, the effect size g, and a (fictitious) reliability coefficient. (Note, however, that the authors, in their table on p. 195, present unbiased estimators *d* while here the corresponding *g* values are input as requested by the program).

#	n_e	n_c	g	r_{tt}
1	*131*	*138*	*0.15845*	*0.90*
2	*40*	*40*	*-0.25647*	*0.81*
3	*40*	*40*	*0.26354*	*0.79*
4	*90*	*90*	*-0.04318*	*0.95*
5	*40*	*40*	*0.65532*	*0.91*
6	*79*	*49*	*0.50602*	*0.88*
7	*84*	*45*	*0.46073*	*0.89*
8	*78*	*55*	*0.58033*	*0.92*
9	*38*	*110*	*0.59104*	*0.79*
10	*38*	*93*	*0.39430*	*0.84*
11	*20*	*23*	*-0.05603*	*0.86*

When typing your own data with the "Editor" from the File Menu, do not finish your data set with an empty line. If reliability coefficients are not available, use 1.0 or 1 instead, pretending perfect measurement.

From the Main Menu select "d values" and "Meta-Analysis" and enter the name of a data file such as "c:\mypath\Hedges2.d". After a second the screen will be filled with the results.

The first two lines confirm the data set including number of effect sizes and total sample size. The next two lines refer to the unmodified effect sizes g, as entered, and present their arithmetic mean (.29582), the standard error SE (.0925) and the effect size variance (.09421) with its standard deviation SD (.3069). The following two lines give the same information based on the unbiased effect size estimator d.

File name: hedges2.d
Number of effect sizes: 11 Total sample size: 1401

Unweighted mean of effect sizes g = *0.29582* *SE = 0.09254*
Observed variance of effect sizes g = *0.09421* *SD = 0.30693*

Unweighted mean of adjusted effect sizes d = *0.29400* *SE = 0.09185*
Observed variance of adjusted effect sizes d = *0.09280* *SD = 0.30463*

----------"*Weighted Integration Method*"----------

Mean effect size *d+ = 0.28597 SE = 0.05532*
Significance Z *= 5.14931* p = *0.00000*
Variance *= 0.00306* *SD = 0.05532*
95% Confidence interval from 0.1771 to 0.3948
Homogeneity Q *= 23.1589* *df = 10* p = *0.01018*

----------"*Random Effects Model*"----------

Mean effect size DELTA *= 0.30075* *SE = 0.09009*
95% Confidence interval from 0.1242 to 0.4773
Significance Z *= 3.33826* p = *0.00042*
Observed variance *= 0.09280*
Error variance *= 0.04183*
Population variance *= 0.05097*
Homogeneity Q *= 23.13359* *df = 10* p = *0.01026*
Amount of variance explained by sampling error: 45.08 %

Figure 6
Screen display for combination of effect sizes d

The second third of the screen contains the results obtained with the "weighted integration method". The mean effect size d+ (.28597) is only slightly different here from the prior unweighted results. The significance of this coefficient is tested by the normal distribution *Z* and its corresponding probability p. The 95%-confidence interval of the mean ranges from .1771 to .3948. The coefficient Q of 23.1589 is significant (p = .01) and emphasizes the heterogeneity of the data set.

Moving on to the "random effects model", the mean effect size delta is .30 with a confidence interval from .124 to .477. The observed variance (.0928) is decomposed into sampling error variance (.0418) and population variance (.0509). That means that only 45% of the observed variance is made up by sampling error. The residual variance component of 55% must be due to systematic factors. The data set is heterogeneous and, therefore, the meta-analyst should search for moderators. By the way, due to the approximative character of the method it may happen that

the population variance becomes slightly negative. In this case the program interprets it as zero and, thus, the amount of explained variance becomes 100%.

2.4.5 The case of unreliability: Correction for attenuation

The optional printed output or the disk file output contains some additional information. If the data set has been completed with accurate reliability information it is worth while to examine the results after correction for attenuation. Each effect size *d* has been divided by the square root of its corresponding reliability. (Note that for technical reasons, the data set has to include a column with reliabilities. If not available, insert unity instead to assume unbiased measures. There is no specific treatment of missing values but coefficients of 0 are automatically transformed to 1.0). There will be a printout of two pages or, if disk file output is desired, two separate files with different suffixes, "filename.OUT" and "filename.REL", the latter one comprising the results after correction for attenuation. Inspecting the output of the present example reveals that all estimates, of course, are somewhat higher after this correction. The less reliable the measures are, the higher will be the "true" effect sizes or correlations. Such results should not be reported without disclosing the previous uncorrected results.

2.4.6 Equivalence to effect sizes *r* (correlations)

Additional information is given on the method suggested by Kraemer (1983). The author has described an algorithm that first transforms all effect sizes *g* into effect sizes *r* (correlations) and their Fisher's z counterparts, then determines the weighted mean of all z values which equals (after backtransformation) the population effect size rho. Finally, rho is backtransformed into the population effect size d.

********** Kraemer (1983) method **********

Mean effect size d = *0.30200*
95% Confidence interval from 0.1952 to 0.4097
Population effect size Rho = 0.14931
Variance of rho = 0.00073
95% Confidence interval from 0.0971 to 0.2007
Homogeneity Chi-square = 24.14141

In this example, the amount of the mean effect size *d* equals Hedges' delta. This illustrates that most methods converge to similar results. It also underscores the equivalence of meta-analyses using either effect sizes *d* or effect sizes r.

2.4.7 Estimating the "Fail-Safe N"

The file-drawer problem (Rosenthal, 1984) addresses the sampling bias in meta-analysis which poses a serious threat to the validity of results. It is assumed that an unknown number of studies with effect sizes of zero remain somewhere in file-drawers because they have either not been submitted for publication or have been rejected. The fail-safe N informs about the number of file-drawer studies required to bring the mean effect size down to a defined level. Orwin (1983) has adapted Rosenthal's formula for probabilities to effect sizes d:

$$N_{fs} = (N_{total} (mean_D - D_{crit})) / D_{crit}$$

Suggested critical *d* levels in meta-analysis are .80 (large), .50 (medium) and .20 (small).

*** Orwin"s Fail Safe N based on "random effects model" DELTA ***

Fail safe for critical delta of .20 = 5.54135
Fail safe for critical delta of .50 = -4.38346
Fail safe for critical delta of .80 = -6.86466

In this example, six file drawer studies with effect sizes of zero are necessary to reduce the population effect size delta of .30 to a mean effect size of .20. The results for critical values of .50 or .80 are meaningless here, because they exceed the empirical value of .30 anyway.

2.4.8 Cluster analysis for effect sizes d

The distribution of effect sizes in a number of experimental studies may be rather heterogeneous. If there is a hypothesis that particular study characteristics (such as old vs. new studies, male vs. female samples) contribute to the observed variation one will subdivide the data set accordingly and will run separate meta-analyses. If homogeneous subsets with different population effect sizes emerge then the hypothesized study characteristics are established as moderators. But in many circumstances the researcher has no idea which characteristic may be responsible for the remaining variation after some moderators have been found. Here an inductive approach may help, and one is advised to inspect the rank-ordered effect sizes directly to obtain an impression of possible groupings. In order to make a sound judgment it is even better to compute a cluster analysis.

Cluster analyses decompose a number of effect sizes into smaller subsets. Two clustering methods have been discussed by Hedges and Olkin (1985,p. 265-283), and the reader is referred to this source for more detailed information. Overlapping clustering is contrasted to nonoverlapping (=disjoint) clustering. The present program includes the latter method. It is based on the clustering method for approximatory standard normal variates. The u-values are determined differently for effect sizes and for correlations. In case of correlations, *r* is transformed into Fisher's z; This is multiplied by the square root of the sample size in the following way:

$$u = \sqrt{(n-3)} * z$$

In case of effect sizes g, a more complicated transformation is necessary to obtain the corresponding u-values (Hedges & Olkin, 1985, p. 272).

As a next step, the u-values are rank-ordered, and the gap between each pair of consecutive u-values is compared to a predetermined critical value. These critical values, taken from a table, differ for the 10%-, 5%- and 1%-levels of significance. The more liberal the significance level (e.g., .10), the more clusters will show up. A conservative significance level (e.g., .01) will result in fewer clusters. For each level of significance the printout of results contains all clusters with their rank-ordered effect sizes. The formulas for the critical values and the computer algorithm for the disjoint cluster analysis as well have been taken from Mullen and Rosenthal (1985). However, these formulas do only approximate the critical values published by Hedges and Olkin (p. 268). The fit is better for smaller numbers of studies.

Clusters usually show up in very heterogeneous data sets only. The data example chosen above did not yield any clusters. The data file "Example.d" results in the following screen display:

CLUSTERS AT 10 % LEVEL OF SIGNIFICANCE
****

CLUSTER 1:
StudyID= 1 Effect Size= 1.2000

CLUSTER 2:
StudyID= 3 Effect Size= 0.9000
StudyID= 4 Effect Size= 0.8000

CLUSTER 3:
StudyID= 2 Effect Size= 0.6000

The four effect sizes *g* have been rank-ordered and split into three groups according to similarity of values. The printout and the disk file output also contain the corresponding solutions for the 5% and 1% levels of significance, respectively.

If the results section for the 10%-level -- as shown quickly on the screen -- does not contain any disjoint clusters, it is not worthwhile to save an output disk file for more information because there will be no clusters at the 5%- and 1%-levels as well. A rank-order of all effect sizes, however, yields an impression of possible overlapping groupings. For that purpose it may appear useful to save the results on disk or to print a hardcopy of it.

As an add-on, this procedure contains the mean and standard deviation of the sample sizes usually to be reported in a research publication, and the product-moment correlation between the sample sizes and the effect sizes for further exploration.

Limitations: The implemented algorithm provides fair approximations, especially for a smaller number of effect sizes. The clustering works well for equal sample sizes, less satisfactorily

however, when sample sizes are very different, for which the decomposition into subsets is biased.

2.4.9 Creating an *r* value file

The present program system provides stem-and-leaf displays only for *r* value data sets. In case of need one can, after running the above analyses, convert the *g* data file to another data set of *r* values. Choose "d values/R-files" from the main menu to create a corresponding data set of effect sizes r. Then, use this new file with the suffix ".RRR" as input for the *r* values procedures.

2.5 Combination of effect sizes from correlational studies

2.5.1 Population effect size and population variance

This program runs a meta-analysis for effect sizes *r* obtained either directly from correlational studies or from other coefficients after transformation to r. (Some paragraphs in the following section are repeated from chapter 1.5.)

The statistical procedures are mostly based on Hunter, Schmidt and Jackson (1982) and Hedges and Olkin (1985). The aims of this meta-analysis are the determination of (1) the population effect size and (2) the homogeneity. Given that all effect sizes belong to the same universe, it is assumed that each sample effect size *r* represents a deviation from its population effect size rho. Effects sizes of studies with large sample sizes should deviate less from the population effect size than small *N* effect sizes. Therefore, in combining all effect sizes, it is fair to assign more weight to large *N* studies. Thus, the best estimate of the population effect size is the weighted average of all correlations.

A sampling error of an effect size from a study with a small *N* is relatively high, whereas a sampling error of an effect size from a large *N* study is small. When averaging correlations, one also averages the sampling errors. The sampling error of an individual correlation may be very high but the sampling error of the average correlation over many studies becomes small -- depending on the total sample size. This happens because positive and negative sampling errors cancel themselves out in summation.

This is different, however, for the variance of correlations because the sign of the sampling error is eliminated by squaring the deviations. The variance across studies appears to be systematically larger due to sampling error. The relationship between sampling error variance and population variance is additive, and therefore, the observed variance of correlations can be regarded simply as the sum of both components.

The Schmidt-Hunter method allows a good approximation of the sampling error variance s^2_e by

$$s^2_e = ((1 - r^2)^2 * k) / N$$

where ^{2} is the squared weighted mean of the effect sizes, k the number of studies, and *N* the total sample size (Hunter et al., 1982, 44).

After this is known one can easily determine the population variance s^2_{res} by subtracting the sampling error s^2_e from the observed variance s^2_r:

$$s^2_{res} = s^2_r - s^2_e$$

The population variance s^2_{res} is also called the residual variance, and its square root is called the residual standard deviation s_{res}.

It is desirable that all the observed variance s^2_r is accounted for by the sampling error s^2_e, and that the residual variance s^2_{res} would become zero. Then the "percentage of observed variance accounted for by sampling error" would be 100% as an indication of homogeneity. Very often, however, only a small percentage can be explained by artifacts, leaving a state of heterogeneity which requires further searches for moderator variables.

The residual standard deviation also serves as the multiplier in the formula for the confidence interval:

$$P (r - 1.96 * s_{res} < rho < r + 1.96 * s_{res}) = .95$$

When the observed variance is totally explained by sampling error, the confidence interval becomes zero.

2.5.2 Accounting for measurement error

In addition to sampling error one can account for measurement error. Most variables under study have not been measured perfectly, i.e., they have reliabilites below 1. Therefore, the observed scores differ from the true scores. Determination of the "true population effect size" and the "true population variance" requires a correction for attenuation procedure.

This procedure is based on formulas described by Hunter et al. (1982, pp. 76-80). Inputs of zero are interpreted by the program as missing values. The authors have developed artifact distributions based on the number of reliability coefficients available to the meta-analyst. It is therefore appropriate to look for correction for attenuation even when the reliability information is not complete.

The "true" population effect size is the weighted mean *r* of the effect sizes which have been corrected for attenuation. It is always higher than any other uncorrected means of effect sizes because it is not strained by measurement error. All statistics in this section have the same meaning as described above, but they relate to effect sizes after consideration of the artifact distributions.

A word of caution is necessary. Accounting for measurement error creates an illusionary situation. Running the procedures on the basis of very low reliabilities leads to very high estimates. Generally, it is superior to apply sound measures with high reliabilities in the first place instead of corrections for attenuation afterwards. But the meta-analyst cannot improve the original studies. In a review article, it is not sufficient to report the "true" estimates only; they should only be reported as additional information to the regular outcomes.

Remember that a complete set of reliability coefficients need not be entered but at least one reliability r(xx) and one reliability r(yy) have to be entered for technical reasons. In the latter case it does not make sense, however, to inspect the part of the output that deals with unreliability.

2.5.3 Homogeneity

A population effect size can only be interpreted reliably if the underlying data set is sufficiently homogeneous. If most of the variance of the observed correlations is accounted for by sampling error, then this requirement is met. As a rule of thumb, Hunter et al. (1982) suggest that at least 75% of the observed variance should be "explained" by artifacts such as sampling error. The remaining 25% of "unexplained" variance are seen as negligible: it would not be worth while to search for moderators.

The authors also provide a chi-square test of homogeneity with a very high statistical power. "Thus if the chi-square is not significant, this is strong evidence that there is no true variation across studies; but if it is significant, the variation may still be negligible in magnitude" (Hunter et al., 1982, p. 47).

Finally, there is a third indicator of homogeneity which seems to be the most important one. The absolute amount of residual variance also counts. If there is almost no variance left after subtracting the sampling error variance, then homogeneity is assumed. ".. it is the actual amount of remaining variance that is important, not the percentage remaining..." (McDaniel, Hirsh, Schmidt, Raju & Hunter, 1986). But how small does the actual amount of residual variance have to be? As a rule of thumb, a residual standard deviation which does not exceed 25% of the population effect size con be considered as small (Stoffelmayr, Dillavou & Hunter, 1983, p. 343).

In sum, there are three indicators of homogeneity:

(1) the residual standard deviation should be smaller than ¼ of the population effect size,
(2) the percentage of observed variance accounted for by sampling error should be at least 75%,
(3) the chi-square test should not become significant.
According to the above-cited literature, the relative importance of these three indicators seems to be in descending order (see also Seipp, 1989a).

2.5.4 Moderator search

If the meta-analyst has made the judgment that the data set is heterogeneous, he or she is advised to search for moderators that may account for the remaining systematic variation. This is performed by breaking down the data into at least two subsets with respect to a theoretically relevant variable (such as gender or socio-economic status). For these subsets separate meta-analyses are computed. In order to classify as a moderator, the following requirements have to be met: (1) the population effect size varies from subset to subset, and (2) the residual variance averages lower in the subsets than for the data as a whole (Hunter et al., 1982, p. 48).

Let us assume that the total residual variance has been .006, and three age-related subsets have yielded the following pattern of results (example taken from Seipp, 1989a, p. 135).

	young	*middle*	*old*
young	*.004*	*.07*	*.11*
middle		*.003*	*.04*
old			*.004*

In the diagonal, the residual variances of the three subsets are given. They are all lower than the comparison value of .006. It would be sufficient if the average of these three residual variances were lower than the total residual variance. Off-diagonal are the mean differences of .07, .11 and .04. As it is, the assumptions are met, and age can be regarded as a moderator in this example.

However, these requirements are usually hard to meet, and one often can be satisfied if at least one homogeneous subset has emerged which allows to estimate reliably the corresponding population effect size.

2.5.5 Significance of population effect sizes

There is no consensus in how to determine whether a population effect size differs significantly from zero. The Utility subroutine "Significance of Correlations" can be used for individual correlations but is misleading for averaged correlations. If, for example, the test is based on *N* as the total sample size, a type I error is likely; if it is based on k as the number of correlations, a type II error is likely.

There are two other ways of solving this problem, both of them making use of the residual standard deviation. First, the 95%-credibility interval gives an impression of the possible variation after accounting for sampling error. Second, there is a rule of thumb saying that the population effect size should be at least twice as high as the residual standard deviation (see Seipp, 1989, p. 130).

2.5.6 Computer output: An example

Figure 7 contains the brief screen output. The complete results can be sent to printer or to a disk file.

The unweighted average correlation (.2075) is given in order to allow a comparison with the population effect size (.204). Sometimes they differ considerably, and this would disclose the influence of sample sizes on the parameter estimation. If, for example, the least qualified study would include the largest sample size, it would hardly be justified to report the weighted solution only.

The observed variance (.00014), the sampling error variance (.00367) and the residual variance (-.00353) are computed. Variances cannot be negative, and as this example shows, such an estimate is a result of the approximative character of the procedures. Consequently, the residual standard deviation has been set to zero and the percentage to 100%.

```
**************************************************************************
File Name: example.r
Total N = 1000
Number of Studies: k = 4

Unweighted mean r = 0.20750
Population effect size (weighted mean r) = 0.20400  r-square = 0.04162

Observed variance of effect sizes      = 0.00014 obs. SD = 0.01200
 95% credibility interval of pop. effect size: from  0.204000  to  0.204000

 Variance due to sampling error      =  0.00367
 Population or residual variance     = -0.00353
 Residual standard deviation  =  0.00000
   <  than 1/4 of population ES = 0.051                          ---> homogeneous
 Percentage of observed variance accounted for by sampling error = 100.00 %      --->
homogeneous
 Test of homogeneity                 Chi-square = 0.15678        ---> homogeneous
 Degrees of freedom                  df = 3
 Significance                        p = 0.984245

 Mean standardized difference              g = 0.41676
**************************************************************************
```

Figure *7: Screen output for meta-analysis of* r *values*

Therefore, the 95% credibility interval became zero, and the population effect size has been reliably determined. All three indicators of homogeneity arrive at the same conclusion.

A mean standardized difference of .416 is computed. In case there is an independent variable x which influences a dependent variable y, it can be concluded that the effect of x on y has the strength of almost half a standard deviation.

Many users may be satisfied with the amount of information given by this brief screen output. The printed or filed output contains more detailed information. Some authors prefer to have all *r* values transformed into Fisher's z (e.g., Hedges & Olkin), while others do not (e.g., Hunter et al.). Some prefer unweighted, but most prefer weighted integrations of effect sizes either with or without correction for attenuation. In order to provide a maximum of information, the printed output is designed with the following structure:

1. Results without Fisher's z transformation
1.1 Unweighted and weighted analyses
1.2 Weighted analysis after correction for attenuation
2. Results with Fisher's z transformation
2.1 same as 1.1
2.2 same as 1.2

The analysis with Fisher's z transformation (Hedges & Olkin, 1985, pp. 229-236) repeats the above analyses with transformed effect sizes, providing also an additional population effect size and another test of homogeneity by Hedges and Olkin.

2.5.7 Cluster analysis for effect sizes r

The distribution of effect sizes in a number of studies may be rather heterogeneous. Grouping smaller numbers of studies into more homogeneous subsets is achieved by a disjoint cluster analysis.

Moderator variables can be detected by employing this strategy and by inspecting the resulting subgroups. This can be called an "inductive moderator search". In addition to those potential moderators which are deducted from theory, empirical subsets of effect sizes may hint to a further moderator responsible for heterogeneous effect size distributions.

CLUSTERS AT 1 % LEVEL OF SIGNIFICANCE
****

CLUSTER 1:
\# *40* r = *0.2350*

CLUSTER 2:
\# *1* r = *0.1500*
\# *41* r = *0.1500*
\# *27* r = *0.1151*
\# *34* r = *0.0850*
\# *43* r = *0.0407*
\# *7* r = *0.0328*
\# *32* r = *0.0050*

CLUSTER 3:
\# *3* r = *-0.6000*

****

Figure 8
Cluster Output for r *Values*

According to Hedges and Olkin (1985), all effect sizes are rank-ordered, and their differences are compared to critical values at the .10, .05 and .01 significance levels. The more liberal the significance level (e.g., .10), the more clusters will show up. A conservative significance level (e.g., .01) will result in fewer clusters. For each level of significance the printout of results contains all clusters with their rank-ordered effect sizes. The formulas for the linear approximations of critical values and the computer algorithm for the disjoint cluster analysis as well have been taken from Mullen and Rosenthal (1985).

2.5.8 Stem-and-leaf display for effect sizes r

In dealing with effect sizes r, the meta-analyst can take advantage of a convenient display technique which provides more information than a histogram: the stem-and-leaf display, introduced by Tukey (1977) within the framework of his "exploratory data analysis" (EDA). The previously used example with four effect sizes yields the following display:

```
-.9 |
-.8 |
-.7 |
-.6 |
-.5 |
-.4 |
-.3 |
-.2 |
-.1 |
-.0 |
+.0 |
+.1 | 9
+.2 | 013
+.3 |
+.4 |
+.5 |
+.6 |
+.7 |
+.8 |
+.9 |
```

Figure 9: *Stem-and-Leaf-Display for 4 Effect Sizes in Data Set: example.r*

The Y-axis or the "stem" is made up by the first digit of correlations from - .9 to + .9. The 4 second digits are in the "leafs". They are ordered according to size within each category. The correlations displayed here are +.19, +.20, +.21, and +.23. For small numbers of effect sizes this display is neither necessary nor impressive, but for large numbers it can be very helpful und illustrative to characterize the data base, as the following example shows.

```
-.9 |
-.8 |
-.7 |
-.6 | 0
-.5 | 7
-.4 | 012677
-.3 | 1144555667
-.2 | 012223445566777788
-.1 | 000112222223334444455556667777899999999
-.0 | 00111122222333333344444555555566666677777788888889
+.0 | 0001112223334445555566679999
+.1 | 001225899
+.2 | 258
+.3 |
+.4 |
+.5 |
+.6 |
+.7 |
+.8 |
+.9 |
```

Figure 10: *Stem-and-Leaf-Display for 164 Effect Sizes*

Figure 10 gives an example of 164 effect sizes. The lowest value in this example is -.60, the highest is +.28.

2.6 Utilities

2.6.1 Conversions to r

Not all studies found in the literature provide the appropriate effect sizes. Instead, some may report t-values, F-values, chi-squares or other statistics. In such cases a transformation is necessary (Figure 11).

```
**************************************************************************

 TRANSFORMATIONS

**************************************************************************

Select a coefficient to be transformed:


Pointbiserial correlation coefficient          ---> p)oint
t-value for 2 independent samples              ---> t)value
F-value for 2 or more independent samples      ---> F)value
Chi square value for contingency tables        ---> x)square
Four cells frequencies                         ---> c)ells
U-value (Mann-Whitney)                         ---> U)value
exact one-tailed probability p                 ---> e)xact p
effect size g (standardized mean difference)   ---> g)value
r to Fishers z transformation: r ---> z        ---> r) to z
Backtransformation             z ---> r        ---> z) to r
Normal distributon Z to probability p          ---> N)ormal

Quit Transformation Program                    ---> Q)uit

**************************************************************************
```

Figure 11: *Transformation menu*

Formulas can be found in Fricke and Treinies (1985), Mullen and Rosenthal (1985) and Rosenthal (1984). Some of them are printed below. The following algorithms were used to obtain an effect size r:

a) Pointbiserial Correlation

Some authors recommend to transform $r_{(pb)}$ to r, others prefer to use $r_{(pb)}$ as their effect size estimate.

$$r = 1.25 * r_{(pb)}$$

b) t value

$$r = \sqrt{[\ t^2\ /\ (t^2\ +\ df)\ ]}$$

c) F for 2 groups

$$t\ =\ \sqrt{F}$$ (continue with b)

d) F for more than 2 groups

$$r = SS_j / (SS_j + SS_i)$$ or:

$$r = F_j\ df_j\ / (F_j\ df_j + df_i)$$

SS_j is the Sum of Squares of factor j
SS_i is the Sum of Squares of the error
df_j is the degrees of freedom of the factor j
df_i is the degrees of freedom of the error

e) Contingency tables

$$r = \sqrt{[\ chi^2\ /\ (chi^2\ +\ N)\ ]}$$

f) Four cells frequencies

$$r = phi = |AD - BC|\ /\sqrt{[\ (A + B)(C + D)(A + C)(B + D)\ ]}$$

Make sure to assign the correct sign after transformation.

g) Mann-Whitney's U

$$r = 1 - 2 * U / (N_1 * N_2)$$

h) Effect size d

$$r = d /\sqrt{[\ d^2\ +\ 4\ ]}$$

If experimental and control groups have very unequal sample sizes, see → Rosenthal (1984, p. 25).

i) Probability p

$p \rightarrow Z$ (see Mullen & Rosenthal, 1985, 134)

$r = Z / \sqrt{N}$

In addition, *p* values can be derived from the normal distribution *Z* (Mullen & Rosenthal, 1985, 133), and *r* can be transformed into Fisher's z (p. 135).

2.6.2 Effect sizes d

This program computes a number of effect sizes and additional statistics for up to 10 groups. Input are the mean, the variance (or standard deviation), and the sample size for each group. Comparisons are examined for all samples, i.e., for k=10 groups there will be k * (k-1)/2=45 comparisons. Typically, one group is seen as the experimental group with a mean of M_e, variance of s_e^2 and sample size of n_e, the other as the control group. The screen will be filled with a pair of groups such as:

Group statistics:
Group 1 Mean= 8.5000 SD= 3.0000 Variance= 9.0000 N= 10
Group 2 Mean= 7.5000 SD= 4.0000 Variance= 16.0000 N= 100

Mean Difference D = 1.0000 Pooled Variance = 15.4167
Effect Size g *based on pooled variance = 0.2547*
Hedges` unbiased estimator d *= 0.2529*
Effect Size r *= 0.0730*
Homogeneity of Variance F = 1.7778

RESULTS [Formula for heterogeneous variances]:
t = 0.9713
Z *= 0.9692*
Significance [one-tailed] p *= 0.1662316*
Significance [two-tailed] p *= 0.3324632*
Omega Square = -0.0005 Effect Size r *= 0.0931*

RESULTS [Formula for homogeneous variances]:
t = 0.7679
Z = 0.7669
Significance [one-tailed] p *= 0.2215830*
Significance [two-tailed] p *= 0.4431659*

Omega Square = -0.0037 *Effect Size* r = *0.0737*

The pooled variance is determined by:

$$s^2 = ((n_e - 1)(s_e)^2 + (n_c - 1)(s_c)^2) / (n_e + n_c - 2)$$

The effect size *g* (according to Glass, 1976) is determined by:

$$g = (M_e - M_c) / SD$$

Usually the above estimator is called "d" in the literature, but according to Hedges and Olkin (1985), here *g* and *d* are distinguished. The coefficient *d* is defined as an unbiased estimator of effect size:

$$d = (1 - (3 / 4 * N - 9)) * g$$

The effect size *r* is derived from *g* by transformation see Rosenthal, 1984, p. 25):

$$r = g / \sqrt{(g^2 + 4)}$$

The F values for testing the homogeneity of variances is given by

$$F = s^2 \text{ max} / s^2 \text{ min}$$

and the user should consult the F table in a statistics book to determine whether the variances can be considered homogeneous. Dependent on this, two results sections follow. The t-test for two independent samples based on heterogeneous variances:

$$t = (M_e - M_c)/ \sqrt{(s_{e^2} / n_e + s_{c^2} / n_c)}$$

The t - test, when variances are homogeneous, is:

$$t = (M_e - M_c)/ \sqrt{((n_e + n_c) / n_e * n_c * s^2)}$$

where s^2 is the pooled variance.

These two results sections also display the omega^2 which is a measure of practical significance, and the effect size r, but this time derived from the t-tests.

2.6.3 Significance of correlations

When the user wants to know whether a correlation coefficient is significantly different from zero, or whether several coefficients from independent samples differ significantly from each other, this program provides the necessary information. Example:

Sample 1 :
r = 0.60 N = *80*

Results for sample 1 :
r = *0.6000*
N = *80*
Fishers z = 0.6931
t = 6.6238
Z *of normal distribution = 5.3666*
Probability p = *0.00000004*
95% interval from 0.438 to 0.724

Sample 2 :
r = *0.40* N = *50*

Results for sample 2 :
r = *0.4000*
N = *50*
Fishers z = 0.4236
t = 3.0237
Z *of normal distribution = 2.8284*
Probability p = *0.00233874*
95% interval from 0.137 to 0.610
Comparison of 2 correlations from independent samples:
Z of normal distribution = 1.4559
Probability p = *0.07270638*

2.6.4 Weighted means, variances and correlations

This is a utility program that combines values based on different sample sizes. For example, in determining the pooled variance for two groups, the following kind of screen display would result.

Input window:

Element1: Value = 12.5 N = *100*
Element2: Value = 13.8 N = *900*

Output window:

The weighted average of 2 elements is: 13.6710
The weighted standard deviation is: 3.6974
Total sample size: 1000

The two variances are input when the program prompts for the value. The resulting pooled variance is 13.67. The same routine can be used for correlations. In this case, the user will be informed about the Fisher's z transformation immediately when input data are typed in. Example:

Element1: Value = -0.3 N = *100* *Fisher`s z = -0.3095*
Element2: Value = -0.5 N = *200* *Fisher`s z = -0.5493*
Element3: Value = -0.7 N = *300* *Fisher`s z = -0.8673*

Fisher`s z backtransformation to r *has been used.*
The weighted average of 3 elements is: -0.5839

Total sample size: 600

2.6.5 t - tests

This option calls the same procedure as "Effect sizes d". Independent of meta-analyses it serves the purpose to compute pairwise t-tests based on means and standard deviations or variances.

References

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

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

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 10*, 3-8.

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

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*, 490-499.

Hedges, L. V. (1986). Issues in meta-analysis. *Review of Research in Education, 13*, 353-403.

Hedges, L. V. & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin, 88*, 359-369.

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

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

Kleine, D. (1990). Anxiety and sport performance: A meta-analysis. *Anxiety Research, 2*, 113-131.

Kleine, D. & Schwarzer, R. (1991). Angst und sportliche Leistung - Eine Meta-Analyse. *Sportwissenschaft, 20*, 9-28.

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta- analysis. *Journal of Educational Statistics, 8*, 93-101.

Kulik, J. A., & Kulik, C.-L. C. (1989). Meta-analysis in education. *International Journal of Educational Research, 13*, 221-340.

Leppin, A., & Schwarzer, R. (1990). Social support and physical health: An updated meta-analysis. In L. R. Schmidt, P. Schwenkmezger, J. Weinman, & S. Maes (Eds.), *Health Psychology: Theoretical and applied aspects*. London: Harwood (in press).

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

McDaniel, M. A., Hirsh, H. R., Schmidt, F. L., Raju, N. S., & Hunter, J. E. (1986). Interpreting the results of meta-analytic research: A comment on Schmitt, Gooding, Noe, and Kirsch (1984). *Personnel Psychology, 39*, 141-148.

Mullen, B., & Rosenthal, R. (1985). *Basic meta-analysis: Procedures and programs*. Hillsdale, NJ: Erlbaum.

Orwin, R. G. (1983). A fail safe N for effect size in meta-analysis. *Journal for Educational Statistics, 8*, 157-159.

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

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

Schwarzer, R., & Leppin, A. (1989a). Social support and health: A meta-analysis. *Psychology and Health: An International Journal, 3*, 1-15.

Schwarzer, R., & Leppin, A. (1989b). *Sozialer Rückhalt und Gesundheit: Eine Meta-Analyse* [Social support and health: A meta-analysis]. Göttingen: Hogrefe.

Seipp, B. (1989a). *Angst und Leistung in Schule und Hochschule: Eine Meta-Analyse.* [Anxiety and achievement: A meta-analysis]. Unpublished Dissertation. Düsseldorf, West Germany: University of Düsseldorf.

Seipp, B. (1989b). *Ansia e rendimento in situazione scolastica e universitaria: Una meta-analisi*. VII Colloquio, 28-30 Settembre 1989. Università degli Studi di Napoli Federico II, Facoltà di Lettere e Filosofia.

Slavin, R. E. (1986). Best-evidence synthesis: An alternative to meta-analysis and traditional reviews. *Educational Researcher, 15*, 5-11.

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

Smith, M. L., & Glass, G. V. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. *American Educational Research Journal, 17*, 419-433.

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

Stoffelmayr, B. E., Dillavou, D., & Hunter, J. E. (1983). Premorbid functioning and outcome in schizophrenia: A cumulative analysis. *Journal of Consulting and Clinical Psychology, 51*, 338-352.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wittmann, W. W., & Matt, G. E. (1986). Meta-Analyse als Integration von Forschungsergebnissen am Beispiel deutschsprachiger Arbeiten zur Effektivität von Psychotherapie. *Psycholo-gische Rundschau, 37*, 20-40.

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage.

Appendix A

History of the META-Programs

Differences between version 3.0 and version 4.0

The META programs have been upgraded from Turbo Pascal 3 to Turbo Pascal 4 and have received a number of substantial and technical improvements, among others:

1.The program now reads ASCII data files. Instead of the previous filer a full-screen editor is built in which is compatible to the WordStar (non-document mode), SideKick, and Turbo Pascal editors. Thus, however, old data sets are no longer compatible. Use the external procedure "OldNew" to convert your old data sets to ASCII data files.
2. It supports input pathes.
3. It allows to write results to an output file on disk.
4. It supports color monitors.
5. The d values routine now includes the "variance partitioning approach".6. A routine on effect sizes and t-tests has been added to compute effect sizes and additional statistics from a number of group means and variances.

Differences between version 4.3 and version 4.4

1. META 4.4 supports a numerical coprocessor, but does not require one.
2. A bug in the cluster procedures for 1% and 10% significance levels has been found and repaired. Now also the study ID is printed.
3. The cluster programs compute the correlation between effect sizes and sample sizes.
4. There are three additional utility programs: - OldNew transforms data sets created with Version 3.x to ASCII data sets required by the updated versions.
   - p-to-r converts a data set of p values into a data set of r-values.
   - g-to-r converts a data set of g values into a data set of r values.

Differences between version 4.5 and version 4.6

The subroutine "Effect Sizes and t-Tests" has produced a wrong g value, due to a missing square root. Thanks to Bettina Seipp for identifying this mistake!

Differences between versions 4.6 and 5.0

The program has acquired a completely new structure including a new main menu with pull down submenus. It allows to check the directory entries, to change the directory and to use a pop-up calculator.

Some other minor technical improvements have been added.
The manual has been completely revised.

Appendix B

Error Handling

1. The results contain one sample more than has been input.

Check the number of cases (studies) that your data set has compared to the k which is presented in the results section. Your data set may contain an empty record as the last input line. Delete that line.

2. The program or subprogram quits without doing its job.

- Check all individual sample sizes: the minimum N is 4 subjects in a study.

- If r values are used: are you sure that you have provided at least one reliability for columns r_{xx} as well as r_{yy}?

- Does your data set exceed 500 effect sizes? If so, please request another program version by indicating what precisely you are going to compute. We will try to tailor the program to your specific needs.

Appendix C

Help Section

This appendix shall answer the question "What to do when?"

1."There are multiple outcomes in a study and I want to combine them to one effect size."

Computing a weighted average is the most frequent way to solve this problem but there are also others (cf. Rosenthal & Rubin, 1986). Select "Utilities/Weighted M, V & Corr" from the main menu and type in your data. You will get an average, weighted by sample sizes in case the subsamples differ.

2. "How can I get some of the results directly into my research report without retyping?"

If results such as the stem-and-leaf display are to be reported, they can easily be stored on disk as an ASCII-file and later be retrieved by any word processing software. Use the disk option instead of the print option.