Citation:

Schulze, R., Holling, H., Großmann, H., Jütting, A. & Brocke, M. (2003). Differences in the results of two meta-analytical approaches. In R. Schulze, H. Holling & D. Böhning (Eds.), *Meta-analysis: New developments and applications in medical and social sciences* (pp.21-39). Hogrefe & Huber.

2

Differences in the Results of Two Meta-Analytical Approaches

Ralf Schulze
Heinz Holling
Heiko Großmann
Andreas Jütting
Michaela Brocke

Psychologisches Institut IV
Westfälische Wilhelms-Universität Münster

Summary

Two meta-analytical approaches for the analysis of correlation coefficients as effect sizes are distinguished. The approaches differ mainly with respect to the effect size being aggregated (r vs. Fisher's z), the weights used in aggregation, estimators for the standard error of the aggregate, and computational procedure for the homogeneity test. The performance of the approaches is compared with regard to bias of estimators, coverage rates of confidence intervals, and Type I error of the homogeneity tests. To comparatively evaluate the approaches, a simulation study with a varying number of studies, number of subjects per study, and population correlations was conducted. The situations in the simulation study are restricted to homogeneous cases. The results show that, overall, the approach as proposed by Hedges and Olkin (1985) as well as Rosenthal (1991) is preferable to the alternative approach of Hunter and Schmidt (1990) for the situations under study.

2.1 INTRODUCTION

As almost any other statistical method, meta-analysis has its own history of developments. Many of its procedural details emerged from adaptations of the method to specific problems of application. This can easily be observed by comparing the major book publications by the various early protagonists of the method in psychology (Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982; Rosenthal, 1991)[1].

The development of meta-analysis was at least partly motivated in the late 1970s and early 1980s by a widespread dissatisfaction with the state of the social sciences and psychology in particular (see Hunter & Schmidt, 1990). The situation was characterized by a large number of published studies on various subject matters apparently showing heterogeneous results. There were only few areas of research for which clear conclusions about the effectiveness of interventions or the quality of models to explain and predict human behavior could be drawn from the literature. Facing this state of affairs, researchers from educational, clinical, and industrial/organizational (I/O) psychology began to develop methods to systematically integrate research findings across studies to overcome deficiencies associated with more narrative methods of literature reviews. Interestingly, these developments were done in parallel in subdisciplines of psychology.

Glass and coworkers were the first to publish a comprehensive treatment of the topic (Glass et al., 1981) with a focus on the evaluation of educational and clinical research questions. Accordingly, their main interest was the development of methods for cumulating results from experimental designs. The most prevalent effect sizes in this are of research were therefore (standardized) mean differences between groups.

In contrast, one of the focal research questions in I/O psychology, and personnel selection in particular, has been the question of whether the validities of personnel selection procedures are situation specific or can be generalized across situations. The validities were ordinarily assessed by the correlations of results from procedures to select applicants with criterion measures like supervisory ratings, for example. Thus, the main concern here was to develop procedures to accumulate the effect size r from a series of studies (Hunter et al., 1982).

In addition to differences in emphasis on effect size families, there also emerged a plethora of further differences between meta-analytical methods, like the introduction of the 75% rule for the detection of heterogeneity in the approach by Hunter and Schmidt (1990), which is unique to their procedures. Furthermore, other researchers developed methods to aggregate study results like p-values or other outcomes of significance tests (Rosenthal & Rubin, 1979), for example, and summarized their developments in their own treatment of

[1]The history of meta-analysis does not begin with these references and is much older, as Hunt (1997) and Olkin (1990) have pointed out.

the topic (Rosenthal, 1991). Additionally, some researchers focused more on the statistical steps of meta-analysis (Hedges & Olkin, 1985).

As a result of these attempts to establish a comprehensive and elaborate set of methods and procedures for the purpose of integrating research findings, there emerged distinguishable *approaches* to meta-analysis. These approaches differ with respect to a series of attributes and are associated with different areas of application, at least in psychological research. Despite attempts to systemize approaches (e. g., Bangert-Drowns, 1986) along characteristics like units or outcomes of analysis, for example, the approaches as described above still seem to prevail in different subdisciplines of psychology.

As a new and largely statistical method, meta-analysis diffused astonishingly fast into psychological research practice and was quickly adopted by the research community. This gave rise to a rapid growth of number of articles that used meta-analysis instead of narrative reviews to summarize the state of the art on a research question. It is noteworthy in this context that in applications of meta-analytical methods, researchers almost exclusively used the respective approach of their field. The approach by Hunter and Schmidt, for example, strongly dominated in I/O psychology.

The story of meta-analysis differs between psychology and other disciplines like medicine, where researchers were more reluctant to use meta-analysis (for a critical assessment of the method, see Feinstein, 1995, for example). Comprehensive treatments of meta-analytic methods (e.g., Sutton, Abrams, Jones, Sheldon, & Song, 2000) also became available in medicine much later than in psychology. Furthermore, the focus of expositions in psychology and medicine differs with respect to effect sizes of main interest, specialized techniques, and many other attributes.

Thus, although there is a general purpose of application common to all methods of meta-analysis, a large number of procedures and techniques exist. This supports the view that meta-analysis should not be regarded as a single method but as a conglomerate of methods to integrate research findings encompassing statistical as well as non-statistical steps. Despite existing differences between approaches, there are also efforts to point out a general structure of the statistical procedures to aggregate effect sizes (Shadish & Haddock, 1994). But even when this general structure can be regarded as accepted, there still remain more subtle differences between approaches. Such differences might influence the meta-analytic results and therefore also the substantive conclusions drawn from these results.

In this chapter, we focus on approaches of meta-analysis developed in psychological research that are designed for the aggregation of the correlation coefficient as an effect size. As a consequence, procedures for the aggregation of standardized mean differences or other effect sizes will not be of concern here. The specific statistical procedures of the relevant approaches will first be presented in the next section. After an analysis of their properties and accentuation of theoretical differences, the results of a simulation study will be presented in the subsequent section. The aim is to make a comparative evaluation of the approaches under scrutiny with respect to their statistical performance.

2.2 COMMON META-ANALYTICAL APPROACHES IN THE SOCIAL SCIENCES

As has been described in the preceding section, there are several approaches of meta-analysis in psychology that differ with respect to a large number of attributes. Of these, the procedures developed in the context of educational research by Rosenthal (1991), in I/O psychology by Hunter et al. (1982), and with a more general statistical focus by Hedges and Olkin (1985) are of main concern here. Because the methods summarized in the book by Rosenthal (1991) were preceded by several journal publications in collaboration with Rubin (Rosenthal & Rubin, 1979, 1982), this approach will be labelled as Rosenthal-Rubin (RR) approach. For applications in I/O psychology, Hunter and Schmidt published a successor of their first book publication (Hunter & Schmidt, 1990), which has become the main reference in this field of research. Their approach will therefore conveniently be labelled as Hunter-Schmidt (HS) approach. Finally, the meta-analytical procedures summarized in the book by Hedges and Olkin (1985) will be abbreviated in the following as HO approach.

The three approaches cannot be comprehensively evaluated here in all of the steps they propose for meta-analysis, partly because they are not equally specific. We therefore focus on the step of statistical aggregation of effect sizes (correlations) to arrive at an estimate of the mean effect size, estimates for confidence limits for the mean effect size, and the homogeneity test. These steps are element of almost every published meta-analysis in the social sciences but represent only a core of the statistical procedures. The elaborate artifact corrections, for example, proposed and advocated by Hunter and Schmidt (1990) are not considered here because other approaches do not specify alternative procedures or do not recommend using corrections for unreliability of measures (Rosenthal, 1991).

The three approaches were also distinguished in a previous comparison of these methods (Johnson, Mullen, & Salas, 1995), which we took as a starting point for our evaluation. However, Johnson et al. specified the approaches in a form that differs from the specification presented in detail below.

One major difference to the specification of Johnson et al. (1995) is that the approaches of HO and RR are not distinguished here. The reasons for this are twofold. First, the HO approach is specified by Johnson et al. (1995) as using the d-statistic as effect size. With correlation coefficients as effect sizes, this would require to convert the correlations to standardized mean differences before aggregation by using an appropriate transformation. This is exactly what Johnson et al. have done in their evaluation of the approaches. Unfortunately, we cannot see any reason why one would in general want to convert a database consisting entirely of r to d. More importantly, we cannot see any indication in the work of Hedges and Olkin for a recommendation to do that. Instead, Hedges and Olkin (1985) present specific formulas for correlations in meta-analysis which we will use in our simulation study. Second, Johnson et al. presented the mean effect size estimator for the RR approach to use study sample sizes (N_i) as weights. We note, however, $N_i - 3$ being recommended

by Rosenthal (1991) as weights, the same as in the HO approach for correlation coefficients. This also has an effect on the standard error for the mean effect size estimator which becomes the same as in the HO approach. Furthermore, although Rosenthal and Rubin (1979) have indeed presented procedures to summarize significance levels, they do not strictly advocate using these methods for the case of interest here. By consulting Rosenthal's work (Rosenthal, 1991) it becomes evident that for the present purposes the approaches by HO and RR are in fact identical. Thus, in contrast to Johnson et al. (1995) we do not distinguish them in the following.

We next turn to a specification of the remaining two approaches. Differences between them can best be explained by considering the basic formulas shown in Table 2.1.

Table 2.1 Basic Formulas for the HO/RR- and HS-Approach (Homogeneous Case)

	HO/RR	HS
Estimator for MES	$\bar{z} = \frac{\sum_{i=1}^{k} \left(\hat{\sigma}_{z_i}^{-2}\right) z_i}{\sum_{i=1}^{k} \hat{\sigma}_{z_i}^{-2}}$	$\bar{r} = \frac{\sum_{i=1}^{k} N_i r_i}{\sum_{i=1}^{k} N_i}$
Variance of MES	$\hat{\sigma}_{\bar{z}}^2 = \left(\sum_{i=1}^{k} N_i - 3k\right)^{-1}$	$\hat{\sigma}_{\bar{r}}^2 = \frac{1}{k}\left(\frac{\sum_{i=1}^{k} N_i \left(r_i - \bar{r}\right)^2}{\sum_{i=1}^{k} N_i}\right)$
Homogeneity test	$Q = \sum_{i=1}^{k} \left(N_i - 3\right)\left(z_i - \bar{z}\right)^2$	$Q = \frac{\sum_{i=1}^{k} \left(N_i - 1\right)\left(r_i - \bar{r}\right)^2}{\left(1 - \bar{r}^2\right)^2}$

Note. HO/RR = Hedges-Olkin and Rosenthal-Rubin approach, HS = Hunter-Schmidt approach, MES = Mean Effect Size.

The HO/RR and HS approach can both be used to summarize a database consisting of a total of k correlations. Whereas by using the HS approach the untransformed correlations r are taken to arrive at an estimate for the mean effect size (MES), the correlations have to be transformed in the HO/RR ap-

proach by applying the following transformation

$$z_i = .5 \times \ln\left(\frac{1 + r_i}{1 - r_i}\right) = \tanh^{-1}(r_i) \tag{2.1}$$

The transformation given in Equation 2.1 was first introduced by Fisher (1915) in the context of deriving the sampling distribution of the correlation coefficient and is mostly labelled as Fisher's z. It is important to note that an estimator for the approximate sampling variance of the transformed correlations z_i is given by $\hat{\sigma}^2_{z_i} = 1/(N_i - 3)$. Thus, the standard error for the transformed correlations *only* depends on the sample size N_i. Whereas the standard error of the correlation coefficient depends on sample size and the population correlation ρ, Fisher's z stabilizes the variance in the sense that it is independent of ρ. By inspecting the formula for the estimator of the mean effect size in the HO/RR approach in Table 2.1, it can be seen that the inverse variances of the estimator ($\hat{\sigma}^{-2}_{z_i}$) are used as weights in the aggregation process. These weights are optimal in the sense that they minimize the variance of the estimator of the MES (Hedges & Olkin, 1985). To compute a mean correlation coefficient for a set of studies, the inverse transformation given by

$$r = \frac{\exp(2z) - 1}{\exp(2z) + 1}. \tag{2.2}$$

is usually applied to $\bar{z}$.

The variances given in the second row of Table 2.1 can be used to conduct a significance test for the MES and also to construct confidence intervals. The formulas differ between the approaches because of the aforementioned use of Fisher's z in the HO/RR approach and untransformed correlations in the HS approach. Additionally, the estimator for the variance in the HS approach differs from the one specified by Johnson et al. (1995). As Schmidt and Hunter (1999) have pointed out, Johnson et al. used a wrong formula for that purpose. By using an incorrect estimator for the standard error, most of their results on the differences between approaches were invalidated.

Confidence intervals for the MES can be constructed for both approaches by using the information given in the first two rows of Table 2.1, assuming a normal sampling distribution for the MES, and applying standard procedures. Whereas on the basis of theoretical results (see Fisher, 1915; Hotelling, 1953) the normal distribution can safely be assumed in the HO/RR approach, the distribution of the correlation coefficient is known to be non-normal for non-zero population correlations ρ and small to moderate N_i. Therefore, problems may result in the HS approach for confidence intervals and statistical testing of the MES, especially when N_i and k are small and ρ is large. In our evaluation of the approaches we follow general (Wilkinson & Task Force on Statistical Inference, 1999) and specific (Schmidt & Hunter, 1995) recommendations to focus on confidence intervals and not null hypothesis testing.

In the last row of Table 2.1, the formulas for conducting a homogeneity test in both approaches are given. Although both formulas are apparently differ-

ent, they follow the same structure in that the squared differences of the (transformed) effect sizes are weighted by the inverse of their variance and summed over k studies. The result is a statistic ordinarily designated Q that follows a χ^2 distribution with $k - 1$ degrees of freedom in the homogeneous case (see also Chapter 10 by Böhning & Dammann as well as Chapter 1 by Hartung, Argaç, & Makambi, this volume).

A still open question is how the approaches HO/RR and HS should be classified with respect to random vs. fixed effects models of meta-analysis. The HO/RR approach as presented in this chapter and most often used in practice, clearly represents a fixed effects model and is classified as such by its authors (Hedges & Olkin, 1985). One indication of the fixed effects model is that there is no variance component being estimated and used in the formula for the variance of the MES (see Table 2.1). In the random effects procedures specified by Hedges and Olkin (1985) a variance component is estimated and used to compute the weights applied in aggregation. The classification of the HS approach is not as simple as for the HO/RR approach. The authors are inconsistent in their own classification by stating that their methods use the fixed effects model (Hunter & Schmidt, 1990, p. 405) but also that the methods are random effects models (Hunter & Schmidt, 2000, p. 275). Furthermore, other authors also do not seem to agree (cf. Erez, Bloom, & Wells, 1996; Field, 2001; Hedges & Olkin, 1985). We note that for the classification it is essential to assess the assumptions of an approach about population parameters, whether they are constant (fixed) or possibly variable (random). Although not obvious from the procedures of the HS approach as presented in this chapter, it is an integral part of the general HS approach that – for several reasons – population correlations can be variable and are best considered as a random variable. However, Hunter and Schmidt (1990) do not present separate procedures for the different models as Hedges and Olkin (1985) do. Instead, the statistical procedures as shown in Table 2.1 for the HS approach can be applied in homogeneous as well as heterogeneous situations, that is, when the fixed or the random effects model, respectively, is appropriate. Of particular interest in this context is the variance of the MES as shown in the right column of Table 2.1. This is the formula generally recommended in the HS approach (Schmidt & Hunter, 1999) because it is supposed to hold for the heterogeneous case and "serves equally well when study effect sizes are homogeneous" (Osburn & Callender, 1992, p. 116).

To summarize, with respect to the statistical procedures of the approaches that are appropriate for the homogeneous case as presented in this chapter, there are two main differences that lead to further differences in details of the procedures. First, in the HO/RR approach Fisher's z is used whereas in the HS approach untransformed correlations are aggregated. Second, in the HO/RR approach the inverses of the (estimated) variances are applied as weights in the aggregation process whereas in the HS approach the sample sizes are used.

In the following simulation study we will comparatively evaluate the performance of the approaches in only the homogeneous case, that is, when there is only one constant population correlation ρ common to all studies.

On the basis of the outlined differences of the approaches we expect the following congruences and divergences in results in homogeneous situations for different N, k, and ρ:

1. The estimates of mean effect sizes of both approaches will be biased in cases when $\rho \neq 0$. The HO/RR will show an upward bias and the results for the HS approach will be biased downwards. However, biases will in general be negligibly small, except for cases in which N is very small.
2. The absolute biases will be larger for the HO/RR approach but the absolute difference between biases of the approaches will be small.
3. The performance for the confidence intervals as assessed by the coverage of the true parameter value will be better for the HO/RR approach when N and k are small. In other cases both approaches will show similar performance.
4. The performance for the homogeneity test will be better for the HO/RR approach when N and k are small. In other cases both approaches will show similar performance.
5. We will not be able to replicate most of the results of Johnson et al. (1995) but our results will be in general agreement with those reported by Field (2001).

Predictions concerning the bias of the estimators of the MES are based on the theoretical results given in the seminal paper by Hotelling (1953). He gives estimates for biases of the transformed and untransformed correlation coefficient which can be used to deduce predictions one and two. However, we also note that previous Monte-Carlo studies on the bias of Fisher's z and r have found a smaller bias for Fisher's z which contradicts prediction two (e.g., Corey, Dunlap, & Burke, 1998; Field, 2001).

The expected superiority of the HO/RR approach for confidence intervals is based on a faster asymptotic of the distribution of the statistic ($\bar{z}$) to the normal distribution in the HO/RR approach as compared to r in the HS approach, for which convergence of the sampling distribution to the normal distribution is remarkably slow. Accordingly, the asymptotic behavior of the Q-statistic is also assumed to be better for the HO/RR procedure. Furthermore, for confidence intervals in the HS approach the mean sampling error of the correlations in studies is used, which may be influenced by what Hunter and Schmidt (1990) call "second order sampling error", that is, inaccuracies in estimation when N and/or k are small.

Finally, we note that there are two kinds of asymptotics relevant for the predictions. First, as the sample size of studies N grows larger but the number of studies k remains constant, results are expected to converge to theoretical predictions derived from large sample theory of the statistics. Second, as k grows larger but N remains constant, results for the estimators need not converge to true values being estimated. Thus, we expect larger N to have a more profound effect on the results in comparison to an increase in k.

2.3 SIMULATION STUDY

To comparatively evaluate the two approaches, a C++ program was written to perform all computations in the situations under study. The procedures to generate the database for applying the computational procedures of the approaches follow the descriptions given by Corey et al. (1998). Details of the computational procedure are reported in Schulze (in press). As already indicated, the main parameters varied in the Monte-Carlo study are number of studies k to be aggregated, number of subjects per study N, and the population correlation coefficient ρ. In the following subsection, the design of the study is described in more detail, and results are presented in the subsequent subsection.

2.3.1 Design and Procedure

The levels used for the number of studies were $k = 8$, $k = 32$, and $k = 128$. They span a wide range of k to explore the results for the approaches in a more typical case ($k = 32$) (see Cornwell, 1988) as well as extreme cases. The same is true for the number of subjects which was varied across the following levels: $N = 16$, $N = 64$, $N = 128$, and $N = 256$. For the population correlation only positive values were used because results were expected to be similar in the negative range of values. The levels of ρ used in the Monte-Carlo study were: $\rho = 0$, $\rho = .10$, $\rho = .30$, $\rho = .50$, $\rho = .70$, and , $\rho = .90$. Again, these values were chosen to explore the performance of the approaches across a wide range of values. The more typical values in psychological research are in the range from 0 to .50. However, there are also research questions in psychology for which correlations to be aggregated can be much higher as in studies on the reliability of a measurement instrument. Thus, very high values were also included in our study.

All levels of the three design features were fully crossed, that is, we distinguished a total of $3 \times 4 \times 6 = 72$ situations. Within these situations all levels were held constant in the simulation procedure. For example, for the situation $k = 8$, $N = 16$, and $\rho = 0$, the procedures of the approaches outlined in Table 2.1 were both applied to databases of 8 studies, all of which had a constant N of 16 subjects and the true correlation underlying all of the observed 8 effect sizes was zero. The computations for all of the 72 situations were repeated 10,000 times and means across these iterations were computed for the statistics of interest. By holding ρ constant within situations, we investigated the homogeneous case for which the fixed effects methods of meta-analysis are appropriate.

2.3.2 Evaluation Criteria

The performance of the approaches with respect to estimation of ρ is straightforward. We computed the estimates for both approaches in all situations and

compared them to the known true values. Deviations from the true values indicate bias of the estimators.

For the confidence intervals, the number of intervals covering the population correlation ρ were counted in all iterations and divided by the number of iterations (10,000). Thus, the coverage probability was estimated by the coverage rates. This is similar to the procedure used by Brockwell and Gordon (2001). All confidence intervals were computed with a prescribed coverage probability of 95%, so that the expected coverage rate is .95 for all situations. Because high coverage rates can come at the cost of long interval widths, the mean widths were also computed by the difference of the mean upper and lower limit across all iterations. This will enable a comparison of the approaches with respect to estimated coverage probabilities when coverage rates are (almost) equal, for example. Ceteris paribus, the approach showing smaller intervals shows better performance. Additionally, estimated confidence interval widths may also be indicative for causes of potential deficiencies of coverage rates.

The homogeneity tests for both approaches and in all situations tests were conducted with $\alpha = .05$, and the rate of significant test results was assessed. Thus, we evaluated whether Type I error rates of the tests for both approaches conformed to the prescribed significance level of the tests.

2.3.3 Results

The results will be presented in separate tables for the estimates of the mean effect size, coverage rates as well as interval widths of the confidence intervals, and Type I error rates for the homogeneity tests. All tables will have the same general structure showing the results for the two approaches in two blocks of columns subdivided by levels of k. Levels of N are shown in blocks of rows where blocks represent levels of ρ.

The estimates of the mean effect sizes for the two approaches are shown in Table 2.2. There are several results for the mean effect size estimates to be highlighted. First, overall the biases are small for most combinations of ρ, k, and N. However, when N is very small (16), biases are not within rounding error for correlations. Nevertheless, we doubt that the absolute value of biases – even in the most extreme cases as observed in Table 2.2 – will affect substantial conclusions in most applications. Biases for both approaches are largest for values of ρ between .50 and .70 and are smallest for $\rho = 0$. Biases also diminish for larger values of N but do show the same behavior for increasing values of k. Thus, for neither of the approaches does adding more studies to a meta-analysis in a homogeneous situation help in obtaining less biased estimates for the population parameter, as long as study sample sizes are equal for the (added) studies.

Second, the results of the HO/RR approach show an upward bias in all situations whereas the results for the HS approach indicate underestimations of ρ. When comparing the absolute values of biases of the approaches, we note that – except for very small N – biases are equal to the third digit. Contrary

Table 2.2 Results: Estimates of Mean Effect Size

	Number of studies k					
	HO/RS			HS		
N	8	32	128	8	32	128
			$\rho = 0$			
16	.001	.000	.000	.001	.000	.000
64	.000	.000	.000	.000	.000	.000
128	.000	.000	.000	.000	.000	.000
256	.000	.000	.000	.000	.000	.000
			$\rho = .10$			
16	.102	.103	.104	.097	.097	.097
64	.101	.101	.101	.099	.099	.099
128	.101	.101	.100	.100	.100	.100
256	.100	.100	.100	.100	.100	.100
			$\rho = .30$			
16	.306	.310	.310	.290	.291	.291
64	.302	.302	.302	.299	.298	.298
128	.301	.301	.301	.299	.299	.299
256	.300	.301	.301	.299	.299	.299
			$\rho = .50$			
16	.509	.512	.513	.486	.487	.487
64	.502	.503	.503	.497	.497	.497
128	.500	.501	.501	.498	.499	.499
256	.500	.501	.501	.499	.499	.499
			$\rho = .70$			
16	.710	.712	.712	.688	.688	.687
64	.702	.703	.703	.697	.697	.697
128	.701	.701	.701	.699	.699	.699
256	.701	.701	.701	.699	.699	.699
			$\rho = .90$			
16	.905	.906	.906	.894	.894	.894
64	.901	.901	.901	.899	.899	.899
128	.901	.901	.901	.899	.899	.899
256	.900	.900	.900	.900	.900	.900

Note. N = Sample size per study, HO/RR = Hedges-Olkin and Rosenthal-Rubin approach, HS = Hunter-Schmidt approach.

to theoretical expectations, the method that employs Fisher's z (i.e., HO/RR) actually shows slightly less bias in cases when $N = 16$ as compared to the approach in which untransformed correlations are aggregated.

Im sum, the results for the estimates of mean effect sizes confirm prediction 1 and contradict prediction 2 on page 28. The results reported in Table 2.2 are in accordance with those reported elsewhere (Corey et al., 1998; Field, 2001; Silver & Dunlap, 1987) and show small biases in opposite directions as well as nearly equal absolute biases for both approaches.

The estimates of the coverage rates and interval widths for the two approaches are shown in Table 2.3. The values in Table 2.3 show both coverage rates of the 95% confidence intervals for the population effect size (left of slashes) and estimated widths of the intervals (right of slashes). The interval widths are given for values of r for both approaches. That is, interval limits were estimated for the Fisher's z values in the HO/RR approach, backtransformed by applying Equation 2.2 to the estimated values for the limits, and aggregated over iterations.

For the situations in which $\rho = 0$ and practically no bias was observed (see Table 2.2), a symmetrical (normal) sampling distribution can be assumed for both approaches. Because interval widths for the HO/RR approach only depend on k, N, and the quantiles of the normal distribution corresponding to the desired coverage probability, at least for $\rho = 0$ the widths should correspond very closely to theoretical expectations derived from:

$$IW = 2 \times z_{.975} \left(\sum_{i=1}^{k} N_i - 3k \right)^{-\frac{1}{2}},$$

where IW denotes interval widths and $z_{.975}$ is the .975-quantile from the standard normal distribution. The interval widths of the HO/RR approach indeed correspond exactly to the theoretical expectations, except for one case ($N = 16$, $k = 8$) showing a small difference to the expected width of $.380 - .384 = -.004$. Note, that this is also the situation, for which a very small bias was observed. The observed close correspondence to theoretical expectations lends support to the validity of the general procedure used to determine the estimates for the interval widths.

Overall, interval widths shown in Table 2.3 become smaller both for increases in k and N, as would be expected from statistical theory. The results also indicate coverage rates close to the desired level in nearly all situations for the HO/RR approach. Few exceptions are observed for combinations of large k, small N, and medium to high values of ρ. For the same combinations of values, the coverage rates of the HS approach are less than expected.

In a comparison of the performance of the approaches' procedures, two main aspects are noteworthy. First, apart from few exceptions coverage rates for the HS approach are smaller than those of the HO/RR approach and interval widths are simultaneously smaller for the HS approach. Thus, confidence intervals in the HS approach are too small and this ostensibly higher precision

Table 2.3 Results: Coverage Rates and Estimated 95% Confidence Interval Widths

	Number of studies k					
	HO/RS			HS		
N	8	32	128	8	32	128
			$\rho = 0$			
16	.951/.380	.947/.192	.949/.096	.891/.326	.934/.175	.945/.089
64	.954/.177	.948/.089	.950/.044	.888/.158	.936/.085	.946/.043
128	.949/.124	.949/.062	.949/.031	.893/.111	.938/.060	.947/.031
256	.951/.087	.952/.044	.951/.022	.899/.078	.939/.042	.950/.022
			$\rho = .10$			
16	.948/.376	.948/.190	.949/.095	.886/.322	.934/.173	.947/.088
64	.952/.175	.950/.088	.950/.044	.898/.157	.935/.085	.946/.043
128	.953/.123	.949/.061	.946/.031	.893/.110	.938/.060	.943/.030
256	.949/.086	.951/.043	.950/.022	.892/.077	.940/.042	.947/.021
			$\rho = .30$			
16	.952/.345	.943/.173	.932/.087	.898/.298	.935/.161	.932/.082
64	.952/.161	.946/.081	.945/.040	.894/.145	.936/.078	.942/.040
128	.949/.113	.949/.056	.946/.028	.894/.101	.937/.055	.944/.028
256	.950/.079	.950/.040	.948/.020	.889/.071	.935/.039	.943/.020
			$\rho = .50$			
16	.948/.283	.935/.142	.885/.071	.890/.252	.930/.136	.894/.070
64	.948/.133	.950/.066	.940/.033	.890/.119	.938/.065	.934/.033
128	.949/.093	.950/.046	.947/.023	.891/.084	.938/.045	.945/.023
256	.950/.065	.951/.033	.949/.016	.889/.059	.937/.032	.946/.016
			$\rho = .70$			
16	.944/.190	.920/.095	.828/.047	.891/.176	.930/.097	.845/.050
64	.949/.090	.942/.045	.920/.022	.887/.082	.931/.044	.923/.023
128	.951/.063	.946/.032	.934/.016	.893/.057	.934/.031	.937/.016
256	.951/.044	.947/.022	.942/.011	.895/.040	.935/.022	.944/.011
			$\rho = .90$			
16	.940/.070	.898/.035	.741/.017	.885/.068	.920/.038	.786/.020
64	.949/.033	.938/.017	.909/.008	.894/.031	.930/.017	.912/.009
128	.951/.024	.945/.012	.929/.006	.892/.021	.935/.012	.925/.006
256	.950/.017	.945/.008	.942/.004	.887/.015	.935/.008	.938/.004

Note. Numbers on the left of the slashes indicate coverage rates, numbers on the right are estimates of interval widths. N = Sample size per study, HO/RR = Hedges-Olkin and Rosenthal-Rubin approach, HS = Hunter-Schmidt approach.

comes at the cost of coverage rates being smaller than desired. Second, for $k = 8$ the HS approach always shows coverage rates less than .90. In these situations, the differences in interval widths to the HO/RR approach are also at a maximum. The inferior performance of the HS approach may be due to second order sampling error in the estimated standard errors for the mean effect sizes when the number of studies is small.

It is also possible on the basis of the results shown in Table 2.3 to shed some light on the performance of the approaches for significance testing, as is ordinarily done in meta-analyses. From the results on the estimation of the MES in Table 2.2, it is known that biases are generally small, so that interval widths can be used to deduce the following results: First, because of smaller interval widths for the HS approach, the hypothesis in question will be rejected more often as in the HO/RR approach. Thus, statistical power will be comparatively higher for the HS approach when the hypothesis is false but the HO/RR may better conform to the desired nominal α-level when the hypothesis is true. Second, for small effects ($\rho = .10$) and combinations of small to medium levels of k and N there is remarkably low power for both approaches. This can be seen, for example, by considering the situation in which $k = 8$, $N = 64$, and $\rho = .10$. Here, there is only a very small bias for both approaches (.001 in absolute value, see Table 2.2) and interval widths are .175 for the HO/RR approach and .157 for the HS approach, respectively. Thus, intervals will mostly contain the value of zero so that the hypothesis is (falsely) not rejected. The results we deduced on testing the hypothesis of zero correlation in the population, conform to the conclusions drawn by Field (2001), who explicitly examined the test performance of the approaches but not the performance for confidence intervals as we have done.

In sum, the HO/RR approach shows a better overall performance for the confidence intervals in comparison to the HS approach. The most disturbing aspect of the results for the HS approach is the consistently low coverage rate for a small number of studies in a meta-analysis. Although most meta-analyses in practice will have more than eight studies in total, this result is particularly relevant for subgroup analyses often done in detailed analyses of a meta-analytic database.

The last aspect in the comparative evaluation of the approaches is the performance of the proposed homogeneity tests. The estimates of the Type I error rates for homogeneity tests are shown in Table 2.4. The results indicate that the test in the HO/RR approach approximately retains the nominal α (.05) in all situations. In contrast, the results for the HS approach show deficiencies as the population correlation increases, N is small, and k grows larger. This is most easily noticeable by inspecting the rejection rates of the null hypothesis for $\rho = .90$. Again, it is noted that this situation will not be given often in practice. Nevertheless, the deviance of the results from the expected values for the HS approach and its conspicuously low rejection rates for small N and ρ lead to a preference for the HO/RR approach in the situations of the simulation study.

Table 2.4 Results: Type I Error Rates for Homogeneity Tests ($\alpha = .05$)

	Number of studies k					
	HO/RS			HS		
N	8	32	128	8	32	128
			$\rho = 0$			
16	.057	.053	.054	.044	.035	.034
64	.050	.053	.053	.046	.048	.047
128	.052	.051	.050	.052	.049	.047
256	.053	.051	.051	.052	.050	.050
			$\rho = .10$			
16	.056	.053	.053	.044	.037	.037
64	.050	.051	.048	.048	.047	.043
128	.051	.050	.053	.048	.049	.050
256	.049	.052	.049	.048	.052	.049
			$\rho = .30$			
16	.051	.055	.054	.045	.049	.056
64	.049	.054	.055	.050	.053	.054
128	.049	.052	.049	.048	.049	.051
256	.048	.049	.055	.047	.049	.054
			$\rho = .50$			
16	.053	.050	.048	.062	.079	.124
64	.047	.054	.052	.051	.062	.068
128	.048	.048	.051	.050	.051	.059
256	.050	.053	.050	.051	.054	.053
			$\rho = .70$			
16	.049	.048	.040	.079	.138	.247
64	.049	.046	.048	.057	.064	.088
128	.048	.050	.048	.052	.064	.068
256	.054	.048	.048	.055	.055	.057
			$\rho = .90$			
16	.048	.043	.034	.097	.210	.445
64	.053	.045	.046	.065	.081	.125
128	.050	.048	.051	.055	.066	.085
256	.048	.050	.050	.050	.061	.063

Note. N = Sample size per study, HO/RR = Hedges-Olkin and Rosenthal-Rubin approach, HS = Hunter-Schmidt approach.

2.4 DISCUSSION AND CONCLUSIONS

From the findings in our simulation study we conclude that the HO/RR approach leads to more reliable results in a meta-analysis of correlation coefficients in a homogeneous situation. Most of our predictions were supported by the results. For the accurate estimation of a mean effect size it does not seem to be critical which of the two approaches is chosen because estimates were mostly accurate within rounding error for both approaches. However, for the construction of confidence intervals or testing of the hypothesis that the population correlation is zero, the HO/RR approach leads to more appropriate results. The same is true for testing the hypothesis of a constant effect size in the population with the homogeneity test (but see Hartung et al., Chapter 1 in this volume).

There are two aspects of the simulation study that might limit its relevance for practical purposes. First, only the homogeneous case was investigated. It has been argued by several authors (e.g., Erez et al., 1996; Hunter & Schmidt, 2000; Osburn & Callender, 1992) that the assumption of homogeneous effect sizes is not realistic for most practical applications. However, despite several calls for an increased use of random effects procedures, methods of meta-analysis as presented in this chapter are still dominant in the literature in the social sciences. For an evaluation of the methods most often used in practice, it seems reasonable to compare them in simulations of situations for which the procedures were designed, as was done in the present study. Our results showed that in this case the HO/RR approach is preferable to the HS approach.

It might well be the case, however, that the procedures perform differently in heterogeneous situations. Whereas the HS approach is supposed to be applicable both in homogeneous and heterogeneous situations, there have been developed different procedures for the two situations by Hedges and Olkin (1985). That is, for the application of the procedures presented by Hedges and Olkin one has to make a decision between procedures *before* their application and this decision depends on assumptions about the true situation in the population of effect sizes. Field (2001) presented a simulation study in which the random effects procedures by Hedges and Olkin were compared to the HS approach in heterogeneous situations. He reported mixed results insofar as there are advantages in using the HS approach for estimation of the mean effect size but also a better performance of the random effects procedures by Hedges and Olkin for inferential purposes, though none of the approaches performed satisfyingly for all simulated conditions. In a comprehensive effort to compare a large set of approaches and accompanying refinements in a number of situations, Schulze (in press) analyzed the approaches as presented here also in heterogeneous situations. In this study, a series of serious deficiencies of both approaches in heterogeneous situations was found and better alternatives were pointed out.

As an alternative to the above mentioned a priori decision between fixed and random effects procedures, Hedges and Vevea (1998) also proposed a so-called conditionally random effects procedure in which the choice of proce-

dures is conditional on the results of the homogeneity test. They showed analytically that their fixed and random effects procedures perform best with respect to inferential goals when applied to situations for which they were designed. The conditionally random effects procedure showed performance better than the random effects procedure in homogeneous situations but performed not as good as the fixed effects procedures here. The pattern of performance was reversed for heterogeneous situations. Several simulation studies (e.g., Field, 2001; Hardy & Thompson, 1998; Harwell, 1997; Schulze, in press) showed, however, that there are serious problems with the homogeneity test and that it should be used with caution as a device to decide between models.

In sum, at least for inferential purposes the decision between fixed and random effects procedures is critical because they perform best when the decision is correct. Thus, such a decision is important when a comparison of approaches is of interest. Unfortunately, authoritative statistical tests for an empirically based decision do not yet seem to be available and compromise procedures like the conditional random effects model or the HS-approach are not without problems. Additionally, a choice between models also crucially depends on the inferential purposes as Hedges and Vevea (1998) have argued and we share their view that an abandonment of fixed effects procedures – even if heterogeneous situations are assumed by default – would be unnecessary.

A second potential limitation for the conclusions based on the results presented in this chapter is that the sample sizes were held constant in the simulations within levels of k and ρ. Indeed, constant sample sizes have never been observed in published meta-analyses and are therefore not realistic. We argue, however, that if differences in sample sizes would exhibit an influence on the results of a meta-analysis, such an influence would depend on the distribution of the sample sizes, its parameters, and also the covariation of sample sizes with effect sizes. Furthermore, this influence might differentially affect the results for the approaches to be compared. If this would be the case, specific assumptions about the distributional properties of sample sizes would obscure a comparative evaluation of approaches. Only in a situation in which firm knowledge about the distribution and its properties were available there would be a gain in making the comparison of approaches more realistic. However, there seems to be no consensus about which distribution to assume in simulation studies. In some Monte-Carlo studies on meta-analytical methods, the sample sizes across studies have been assumed to be normally distributed with varied parameters (e.g., Field, 2001; Osburn & Callender, 1992), uniformly distributed (e. g., Erez et al., 1996), or have been held constant (e. g., Corey et al., 1998; Overton, 1998) as in the present study. Unfortunately, neither of these distributions seems to mirror the distribution observed in practice. The results of a content analysis of 81 meta-analyses in industrial/organizational psychology reported by Cornwell (1988), clearly show a distribution of sample sizes far from normal or uniform. Instead, at least in this field of research the distribution is characterized by strong positive skewness and kurtosis. We therefore think that holding sample sizes constant across studies is a more sensible choice over assuming a distribution not observed in practice.

Notwithstanding the outlined potential limitations, it is clear from the results presented in this chapter that many of the conclusions concerning the HS approach drawn by Johnson et al. (1995) were based on erroneous results. However, more extensive simulation studies are needed to reach final conclusions about the usefulness of existing approaches.

REFERENCES

Bangert-Drowns, R. L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin*, *99*, 388–399.

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

Corey, D. M., Dunlap, W. P., & Burke, M. J. (1998). Averaging correlations: Expected values and bias in combined Pearson *r*s and Fisher's *z* transformations. *The Journal of General Psychology*, *125*, 245–261.

Cornwell, J. M. (1988, August). *Content analysis of meta-analytic studies from I/O psychology.* Paper presented at the 96th annual meeting of the American Psychological Association, Atlanta, GA. (ERIC Document Reproduction Service No. ED 304469).

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

Feinstein, A. R. (1995). Meta-analysis: Statistical alchemy for the 21st century. *Journal of Clinical Epidemiology*, *48*, 71–79.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte-Carlo comparison of fixed- and random effects-methods. *Psychological Methods*, *6*, 161–180.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*, 507–521.

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

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

Harwell, M. (1997). An empirical study of Hedges's homogeneity test. *Psychological Methods*, *2*, 219–231.

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

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

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society, Series B*, *15*, 193–232.

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

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

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

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

Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology*, *80*, 94–106.

Olkin, I. (1990). History and goals. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 3–10). New York: Russell Sage Foundation.

Osburn, H. G., & Callender, J. C. (1992). A note on the sampling variance of the mean uncorrected correlation in meta-analysis and validity generalization. *Journal of Applied Psychology*, *77*, 115–122.

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

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

Rosenthal, R., & Rubin, D. B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin*, *86*, 1165–1168.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, *92*, 500–504.

Schmidt, F. L., & Hunter, J. E. (1995). The impact of data-analysis methods on cumulative research knowledge. *Evaluation & The Health Professions*, *18*, 408–427.

Schmidt, F. L., & Hunter, J. E. (1999). Comparison of three meta-analysis methods revisited: An analysis of Johnson, Mullen, and Salas (1995). *Journal of Applied Psychology*, *84*, 144–148.

Schulze, R. (in press). *Meta-analysis: A comparison of approaches.* Seattle, WA: Hogrefe & Huber.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology*, *72*, 146–148.

Sutton, A. J., Abrams, K. R., Jones, D. R., Sheldon, T. A., & Song, F. (2000). *Methods for meta-analysis in medical research.* Chichester: Wiley.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.