This is chapter 5: Statistical approaches to meta-analysis (pp. 55-86) from

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Hogrefe & Huber.

5

Statistical Approaches to Meta-Analysis

After having outlined the more general characteristics and procedures in the analysis of effect sizes, the present chapter will provide an overview of more specific procedures and formulae proposed in the literature. As introduced in Section 2.2, comprehensive treatments of meta-analysis associated with different author names and at least partially comprising different sets of procedures and formulae are labeled approaches. The approaches of interest in the present context are widespread predominantly in the social sciences and especially in the psychological literature. Furthermore, the focus of this chapter is narrowed down to the statistical details of the approaches. Whereas the models presented in the previous chapter are also well-known in other areas of research, there are some distinctive features of the following approaches that have to be explicated in detail before an empirical evaluation is undertaken.

After considering the presentation of the models, the question arises why sets of procedures and techniques are subject to a comparative evaluation at all. Why not always choose the most proper model and corresponding estimators, considered as optimal from a statistical point of view for a specific research problem? First, the introduction of meta-analysis as a new statistical tool for the social sciences has been associated with proponents from the beginning of its history. This lead to idiosyncrasies of approaches and preferences of authors becoming entrenched in research practice. For example, the correction of correlation coefficients before their aggregation has become almost mandatory in the field of I/O psychology, whereas in the field of educational psychology, these corrections are only considered optional (see also Section 2.2).

Besides historical reasons, properties of the correlation coefficient as an effect size also require specialized techniques. Transformations of the correlation coefficient as presented in Section 3.3 represent such specialized techniques that are not of relevance when a different effect size, like the odds ratio for example, is given in the studies under investigation.

In this book, major approaches for correlations as effect sizes in the field of psychology are evaluated. Further, a series of refinements are introduced for a more comprehensive evaluation of the available procedures. The following sections are structured in correspondence with this classification and all necessary formulae for computation are given. This entails some redundancies in the presentation of formulae but they are nevertheless completely given for reference and to document the procedures as employed in the Monte Carlo study. In addition, the approaches will also be presented in the same order as the meta-analytical steps in the presentation of the FE and RE model, with estimation first and inference thereafter.

As a final remark with respect to the approaches, the reader may wonder why there is also one section that provides computational formulae for the aggregation of d as an effect size when the focus should be on correlations. These procedures are given because one of the aims of the Monte Carlo study is also to evaluate the results of procedures that are based on transformed effect sizes (see Section 3.3). Thus, the common assumption that the transformation of effect sizes — specifically from r to d — is, in essence, inconsequential for the meta-analytical results will be tested. To do this, procedures for the aggregation of transformed effect sizes (i.e., d in the present context) have to be specified. Of course, the prominent set of procedures proposed by Glass et al. (1981) could have been added as another approach. This was not done in order to keep the number of approaches at a manageable level, keep the focus on approaches for correlation coefficients as effect sizes[1], and to maintain comparability to similar examinations of approaches in the literature (e.g., Johnson, Mullen, & Salas, 1995). The question remains which procedures should be used to aggregate the effect sizes d_i when there are many procedures available. Of the major approaches under examination, any could have been chosen for this task. The approach proposed by Hedges and Olkin (1985) presented in the following section was chosen to provide the procedures for this aggregation. The reason for this choice was that it seemed the statistically best founded set. All other major approaches also provide details on the aggregation of d as an effect size so that the choice may also be considered as somewhat arbitrary.

5.1 HEDGES AND OLKIN

The first approach is most comprehensively explicated in Hedges and Olkin (1985) and will be labeled HO in what follows. Technical details of the approach and further procedures proposed by the authors are scattered across a series of articles that may also be consulted for reference (Hedges, 1982a, 1982b, 1982c, 1983a, 1983b, 1991).

The presentation is divided into two subsections. The first one will give details on the aggregation of correlation coefficients and the second one on ag-

[1]The approach by Glass et al. (1981) does not provide procedures specifically designed for a meta-analysis of correlations.

gregation procedures for d. It should be noted at the outset that the authors of this approach do *not* explicitly advocate the transformation of r to d when the database only consists of correlation coefficients, as will be the case in the Monte Carlo study in Part III of this book. However, they do provide transformation formulae for effect sizes, so that it is possible to apply their procedures as presented. To distinguish between the r-based and d-based procedures, the symbol HOr represents the r-based and HOd the d-based variant, respectively.

In addition to the d-based approach there is also one refinement in procedures introduced that goes back to the work of Hotelling (1953). To differentiate HOr from this refinement, the latter will conveniently be denoted by HOT.

5.1.1 Procedures for r as Effect Size

In the HOr approach, the observed correlation coefficients are first transformed by using the Fisher-z transformation (see also Section 3.1)

$$z_i = \frac{1}{2} \ln \frac{1 + r_i}{1 - r_i}$$

(Hedges & Olkin, 1985, p. 120, Equation 19; p. 227, Equation 4).

Next, the variances of the transformed effect sizes are given by

$$\hat{\sigma}^2_{z_i} = \frac{1}{n_i - 3} \tag{5.1}$$

(Hedges & Olkin, 1985, p. 227). Note that there is no uncertainty in determining these variances since n_i of every study is given and the parameter estimate does not influence the weights as is the case in approaches not using the Fisher-z transformation.

Estimation of Mean Effect Size. The mean effect size estimate in z-space is computed by using

$$\bar{z} = \frac{\sum_{i=1}^{k} (n_i - 3)\, z_i}{\sum_{i=1}^{k} (n_i - 3)}$$

(Hedges & Olkin, 1985, p. 231, Equation 12). This exactly corresponds to the general procedure outlined for the FE model, where the reciprocals of the (estimated) variances of the estimates are used as weights.

Due to the fact that the aim of estimation presumably is never a mean effect size in z-space in practice, the estimate is transformed to an $\bar{r}$ by the inverse Fisher-z transformation

$$\bar{r} = \frac{\exp(2\bar{z}) - 1}{\exp(2\bar{z}) + 1}$$

(Hedges & Olkin, 1985, p. 227, Equation 8). This results in the estimate of the mean effect size in the HOr approach.

Significance of Mean Effect Size. The next step of testing the mean effect size begins by determining the standard errors for the mean effect size with

$$\hat{\sigma}_{\overline{z}} = \frac{1}{\sqrt{N - 3k}} \tag{5.2}$$

(Hedges & Olkin, 1985, p. 231). In Equation 5.2 and in what follows, N denotes the total number of participants in all studies, that is, $N = \sum_{i=1}^{k} n_i$.

Using the standard error, one can test the null hypothesis of zero mean universe effect sizes by using

$$g = \overline{z}\sqrt{N - 3k} \tag{5.3}$$

(Hedges & Olkin, 1985, p. 231), where g is ordinarily assumed to approximately follow a standard normal distribution.[2]

Approximate lower and upper limits of the confidence interval are constructed by

$$\begin{aligned} z_L &= \overline{z} - g_\alpha \hat{\sigma}_{\overline{z}} \\ z_U &= \overline{z} + g_\alpha \hat{\sigma}_{\overline{z}} \end{aligned} \tag{5.4}$$

and are customarily transformed by the inverse Fisher-z transformation when reporting results.

Homogeneity Test Q. The test statistic is provided — as described in the context of the FE model — with Fisher-z transformed effect sizes as

$$Q = \sum_{i=1}^{k} (n_i - 3)\,(z_i - \overline{z})^2 \tag{5.5}$$

(Hedges & Olkin, 1985, p. 235, Equation 16). It is noted that if $\rho_1 = \ldots = \rho_k$ and $N \to \infty$, Q asymptotically follows a χ^2_{k-1}-distribution.

Hotelling's (1953) Adjustment. In his seminal paper Hotelling (1953) proposed several improvements of Fisher-z with the aim to correct the bias in Z and also to stabilize its variance (see also Section 3.1). Of these, the following correction proposed to be applied to an average z seems to be especially attractive

$$\overline{z}_{\text{Hot}} = \overline{z} - \frac{\tanh \overline{z}}{(2n - 9/2)} \tag{5.6}$$

(Hotelling, 1953, p. 219). In Equation 5.6, n denotes a constant sample size across studies. In practical meta-analyses this will rarely be the case, so that the mean of the sample sizes across studies might be used instead.

One reason why the correction given in Equation 5.6 is used instead of others proposed by Hotelling is the fact that it was constructed to be applied to an average z and therefore perfectly fits into procedures of meta-analysis.

[2]To reiterate, the somewhat unusual symbol g is used throughout the text to avoid confusion of standard normal deviates with values of Fisher-z.

Another reason is that with this correction, a reported mean z (or transforms thereof) can be corrected to yield an improved estimate of the mean effect size. An evaluation of this procedure is therefore of relevance for the conduct of meta-analyses as well as their reception. Previous results of a Monte Carlo study conducted by Donner and Rosner (1980) who used the HOT approach as outlined here and compared its performance to HOr, a maximum likelihood estimator and an estimator similar to the one proposed by Hunter and Schmidt (see Section 5.3), suggest a good performance of HOT. They recommended the use of HOT (and the Hunter and Schmidt procedures) for the estimation of μ_ρ in $\mathfrak{S}_1$ in comparison to the other approaches they have evaluated, especially when n is small. For a Monte Carlo study on a different modification of the Fisher-z transformation proposed by Hotelling (1953, p. 223), see Paul (1988) (see also Section 3.1).

A significance test can be performed by using the standard error formula given in Equation 5.2 and applying the procedures outlined in Equations 5.3 and 5.4 for the significance test and the construction of confidence limits, respectively.

5.1.2 Procedures for d as Effect Size

As was outlined in Section 3.3, correlation coefficients may also be transformed to d by the following transformation

$$d_i = \frac{2r_i}{\sqrt{(1 - r_i^2)}}.$$

Of course, there would be no need to apply this transformation if all effect sizes were given as correlation coefficients because procedures to aggregate this type of effect size have just been outlined. In practice, however, it is scarcely the case that all retrieved studies are of the same design and some may be experimental studies, so that only d may be available for some studies. Conventionally, effect sizes are then converted to r or d, depending on convenience. The result is a database that is a mix of converted and non-converted effect sizes r or d. Though not explicitly stated, the usual assumption is that the conversion does not have any influence on the results of the meta-analysis. If this was true, then the application of the following procedures to d values that result from a conversion from r should lead to the same results as the application of the procedures outlined in the previous subsection to the original correlation coefficients r.

Finally, it should be noted that the conversion formula given in the equation above is not the form of effect size that Hedges and Olkin (1985) advocated. Instead, they proposed an unbiased estimator of δ that was already introduced in Section 3.2 and that is considered preferable from a statistical point of view. The conversion formula that represents the d statistic according to Cohen (1988) given here was nevertheless used because of its much more widespread use in the literature and therefore relevance for actual research.

Estimation of Mean Effect Size. The first step is estimation of the estimate's variance. This variance was already given on page 29 in Equations 3.9 and 3.10 for equal n in all studies, respectively.

The reciprocals of these variance estimates can be taken as weights to yield a mean effect size estimate by

$$\overline{d} = \frac{\sum_{i=1}^{k} d_i / \hat{\sigma}^2_{d_i}}{\sum_{i=1}^{k} 1 / \hat{\sigma}^2_{d_i}}$$

(Hedges & Olkin, 1985, p. 111, Equation 6).

Significance of Mean Effect Size. The null hypothesis test follows the general logic outlined for the FE model and can be accomplished by using

$$g = \frac{\overline{d}}{\hat{\sigma}_{\overline{d}}}$$

as the test statistic with $\hat{\sigma}_{\overline{d}}$ given by

$$\hat{\sigma}_{\overline{d}} = \left(\sum_{i=1}^{k} \frac{1}{\hat{\sigma}^2_{d_i}} \right)^{-\frac{1}{2}}$$

(Hedges & Olkin, 1985, p. 112, Equation 9).

Approximate lower and upper limits of the confidence interval are constructed by the following equations

$$d_L = \overline{d} - g_\alpha \hat{\sigma}_{\overline{d}}$$
$$d_U = \overline{d} + g_\alpha \hat{\sigma}_{\overline{d}}$$

The results for the confidence interval limits are transformed to r by Equation 3.11 when results are reported in Chapter 8 to make them comparable to the estimated limits of the other approaches.

Homogeneity Test Q. As for the HOr approach, it is also possible to conduct a homogeneity test by using the Q-statistic

$$Q = \sum_{i=1}^{k} \frac{\left(d_i - \overline{d} \right)^2}{\hat{\sigma}^2_{d_i}}$$

(Hedges & Olkin, 1985, p. 123, Equation 25). Again, Q is supposed to asymptotically follow a χ^2_{k-1}-distribution when the null hypothesis is true. It will be particularly interesting to evaluate the performance of this test in comparison to the HOr approach in the Monte Carlo study to be presented. Differences

between these tests will reflect potential problems concerning the conversion of effect sizes.

5.2 ROSENTHAL AND RUBIN

The methods proposed by Rosenthal and Rubin (RR) are described in Rosenthal (1978, 1991, 1993) as well as Rosenthal and Rubin (1979, 1982). As will become evident from the following presentation, the procedures are very similar or almost identical to those given for HO*r*.

Estimation of Mean Effect Size. In the RR approach, correlations are also transformed via Fisher-*z* prior to further processing

$$z_i = \frac{1}{2} \ln \frac{1 + r_i}{1 - r_i}$$

(Rosenthal, 1991, p. 21, Equation 2.22).

For aggregation, it is not entirely clear what form of weights should be used. With reference to Snedecor and Cochran (1967), Rosenthal (1993, p. 534) proposes for the weighted aggregation of Fisher-*z* values to use the degrees of freedom as weights "or any other desired weight". For the current case this would be $n_i - 3$, so that the mean effect size estimate for RR would be *identical* to the one presented for HO*r*. As an alternative to the degrees of freedom, the sample sizes n_i were chosen as weights but it is noted that these weights are not explicitly recommended by Rosenthal and Rubin. The following computational procedure is given for the mean effect size estimate

$$\bar{z} = \frac{\sum_{i=1}^{k} n_i z_i}{\sum_{i=1}^{k} n_i}$$

(Rosenthal, 1991, p. 74, Equation 4.16; p. 87, Equation 4.32 and 4.33).

In the same way as for the HO*r* approach, the resulting estimates have to be transformed back to $\bar{r}$ by

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

Significance of Mean Effect Size. For significance testing of the mean effect size, the following statistic is proposed

$$g_i = r_i \sqrt{n_i}$$

(Rosenthal, 1991, p. 19, Equation 2.18; see also p. 29). That is, correlations are transformed to standard normal deviates which are aggregated subsequently

by applying the weights as proposed in the context of estimating the mean effect size

$$g = \frac{\sum_{i=1}^{k} n_i g_i}{\sqrt{\sum_{i=1}^{k} n_i^2}}$$

(Rosenthal, 1991, p. 86, Equation 4.31). Recall again that the authors originally proposed to use the degrees of freedom as weights (i.e., $n_i - 3$).

After having computed the standard normal deviates, approximate lower and upper limits of the confidence interval are constructed by

$$z_L = \overline{z} - g_\alpha \hat{\sigma}_{\overline{z}}$$
$$z_U = \overline{z} + g_\alpha \hat{\sigma}_{\overline{z}}$$

Again, such confidence interval limits are transformed by the inverse Fisher-z transformation when results are reported.

Homogeneity Test Q. The homogeneity test Q is the same as proposed in the HOr approach and given by

$$Q = \sum_{i=1}^{k} (n_i - 3) (z_i - \overline{z})^2$$

(Rosenthal, 1991, p. 74, Equation 4.15).

5.3 HUNTER AND SCHMIDT

In contrast to the RR approach, the procedures introduced by Hunter and Schmidt (1990) as well as Hunter et al. (1982) offer a series of new features in comparison to HOr. The approach, labeled HS in what follows, is detailed in a very large series of articles of which only a few are referenced (e.g., Burke, 1984; Hunter, Schmidt, & Pearlman, 1982; Schmidt & Hunter, 1977; Schmidt, Hunter, & Pearlman, 1982; Schmidt, Hunter, Pearlman, & Hirsh, 1985). There have also been a series of refinements that are not dealt with in the present context, so the reader is referred to the relevant literature (e.g., Callender & Osburn, 1980; Callender, Osburn, Greener, & Ashworth, 1982; Raju, Burke, Normand, & Langlois, 1991; Schmidt et al., 1993) and also to a recent assessment of the impact of the methods on research and practice in personnel selection (Murphy, 2000), as well as a discussion of the quality of these so-called validity generalization methods from various perspectives (see Murphy, 2003).

The latter two references signify the close connection of this approach with the field of I/O psychology and personnel selection in particular. Though not limited to this field, the main developments and applications have been done

in the field of personnel selection. The approach is also often called *validity generalization* which expresses its main characteristics.

First, the preoccupation of applications using the approach with correlation coefficients that represent (predictive) validities of personnel selection methods is indicated. Hence, most of the procedures and their refinements proposed are concerned with correlation coefficients as an effect size measure, but procedures for coefficients from the *d* family have also been proposed (see, e.g., Hunter & Schmidt, 1990). The approach is therefore not limited to correlation coefficients.

Second, one major question in personnel selection is whether validities can be generalized. The designation of *generalizable* is done in a binary fashion, that is, either test validity generalizes or not. Hence, the term *validity generalization* denotes a classification of tests in two groups. This seems to be a quite specific use of the word "generalization" in comparison to more popular ones (see, e.g., Shadish et al., 2002) and might be understood only by considering the legal circumstances in the United States of America (for a review, see Landy, 2003). A common misinterpretation of the term is that it is used to characterize the variability in (predictive) validity coefficients for a certain test across situations. If the validity coefficients are not stable across situations, one might easily use a phrase like "test validity does not generalizes (across situations)" to describe this fact. However, in the HS terminology, a different term is used in this case, namely *situational specificity*. It is considered as quite an important question for practice whether a personnel selection method has to be validated in every new situation of application on the one hand. On the other hand, validities might have been demonstrated to be stable across a series of situations so that it can reasonably be assumed that they hold in a new situation without the need for collecting new evidence. The former case describes a test which is situationally specific, and in the latter case validities are not specific for situations.

Whereas the proponents of this approach have always strongly argued in favor of generalizability and situational non-specificity, and also presented evidence to support these claims in the field of personnel selection (e.g., Schmidt & Hunter, 1998), the approach and its procedures has also been severely criticized (e.g., James et al., 1986; James, Demaree, Mulaik, & Ladd, 1992). Because such issues are not of utmost importance for the statistical quality of the approach, the reader is referred to the book edited by Murphy (2003) for a comprehensive overview.

One further important and distinctive feature of the HS approach is the authors' strong recommendation to correct correlations for various so-called *artifacts* before they are aggregated (for the pros and cons of applying the corrections, see, e.g., Schmidt & Hunter, 1999b). A series of research scenarios to illustrate the relevance of correcting artifacts is given by Schmidt and Hunter (1996). It might be noted, however, that proponents of other approaches have provided similar corrections of effect sizes (e.g., Hedges & Olkin, 1985), though not as elaborate as has been developed within the HS approach. Nevertheless,

this feature of the HS approach is not of relevance in the Monte Carlo study of Part III, thus only the basic idea is given here.

One of the potential so-called artifacts which influence the correlation between two variates X and Y is measurement error, another potential artifact is restriction of range[3]. If both artifacts apply in a certain situation, the correlation in the population is attenuated. Let ρ_a be the attenuated correlation and ρ its unattenuated counterpart. Then

$$\rho_a = \rho \times A^{-1}$$

describes the relationship between these two, where A denotes a so-called *artifact multiplier*. The artifact multiplier is considered as a constant which results from one or multiple artifacts operating in a specific situation. For example, if one of the correlated variables has a reliability of $r_{tt} = .81$, then — drawing on results from classical test theory (Lord & Novick, 1968, p. 69) — the artifact multiplier for the correction of unreliability in the predictor is $A = \sqrt{r_{tt}} = .90$. Thus, ρ_a is attenuated by a factor of .90 in this example.

Meta-analysis based on artifact corrected correlations are certainly useful — at least as an addendum to analyses based on uncorrected correlations — to shed light on "what effect size we might expect to find in the best of all possible worlds" (Rosenthal, 1994, p. 240). What the implications and interpretations of meta-analytic results including artifact corrections are, is, however, debatable. In the literature on validity generalization it has been repeatedly argued that analyses based on such a corrected database can lead to estimates of the relationship between *constructs* (e.g., Schmidt & Hunter, 1999b). Unfortunately, this is not the case as Boorsbom and Mellenbergh (2002) have convincingly argued.

In sum, artifact corrections are an important feature of a "full-blown" HS approach but they are neither necessary to evaluate the core of the HS procedures as outlined in the following paragraphs nor do they unequivocally lead to refined interpretations of meta-analytic results as proposed by Hunter and Schmidt. For more details on corrections for artifacts, the reader is again referred to the pertinent literature (Hunter & Schmidt, 1990, 1994a) and also previous Monte Carlo studies that incorporated and partly also evaluated these corrections (e.g., Aguinis & Whitehead, 1997; Callender et al., 1982; Cornwell & Ladd, 1993; Duan & Dunlap, 1997; Law, Schmidt, & Hunter, 1994; Raju, Anselmi, Goodman, & Thomas, 1998).

After these preliminaries, the focus of the following outline of the HS approach will be on the proposed statistical procedures for aggregating the available research database. In the HS terminology, this would be called *bare-bones* meta-analysis.

[3]For a more complete list of potential artifacts, see Hunter and Schmidt (1990).

Estimation of Mean Effect Size. The aggregation of correlation coefficients in the HS approach is done by applying

$$\bar{r} = \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i}$$

(Hunter & Schmidt, 1990, p. 100). It can be seen by inspecting this equation that in contrast to the previous approaches HO*r* and RR, the correlation coefficients are not transformed before the coefficients are aggregated. A negative bias is therefore expected in contrast to the (uncorrected) Fisher-*z* based approaches which exhibit a positive bias (see Section 3.1). Furthermore, the coefficients are weighted by n_i and not by the optimal weights represented by the reciprocals of the squared standard errors of the estimates. From a statistical point of view, this leads to larger standard errors of the mean effect size estimate and therefore less power in testing.

Significance of Mean Effect Size. The estimate for the standard error of the mean effect size is a hotly debated issue in the HS approach (cf. Callender & Osburn, 1988; Duan & Dunlap, 1997; Hunter & Schmidt, 1994b; Osburn & Callender, 1992) and has also lead to some confusion when evaluating the HS approach (Johnson, Mullen, & Salas, 1995; Schmidt & Hunter, 1999a). Indeed, confusion may stem from the various forms of computational formulae that have been proposed in the HS approach. This issue is taken up by evaluating the four most prominent versions for the standard error presented in the following formulae.

The formula recommended for estimation of the sampling variance of the mean effect size estimate by (Schmidt et al., 1988; see also Osburn & Callender, 1992; Whitener, 1990) is

$$\hat{\sigma}^2_{\bar{r}1} = \frac{\left(1 - \bar{r}^2\right)^2}{(N - k)} \tag{5.7}$$

(Osburn & Callender, 1992, p. 115, Equation 3). The index 1 in $\hat{\sigma}^2_{\bar{r}1}$ signifies that it is the first version presented here. When this version of the sampling variance is used in what follows, it will be labeled HS1. This version is supposed to yield the best results when a homogeneous situation like $\mathfrak{S}_1$ is given (Osburn & Callender, 1992).

The second version HS2 is given by

$$\hat{\sigma}^2_{\bar{r}2} = \frac{\sum_{i=1}^{k} \left(1 - r_i^2\right)^2 / (n_i - 1)}{k^2} \tag{5.8}$$

(Osburn & Callender, 1992, p. 116, Equation 4). Except for very small and divergent sample sizes, HS1 and HS2 are expected to yield similar results (Os-

burn & Callender, 1992). The terms summed in the numerator of equation 5.8 are essentially the estimated variances of the individual correlations. These estimates in the numerator have attracted considerable attention in the literature on validity generalization (e.g., Callender & Osburn, 1988; Fuller & Hester, 1999; Hunter & Schmidt, 1994b; Osburn & Callender, 1992) and it has been shown that they depend on several characteristics of the research situation like range restriction, for example (Aguinis & Whitehead, 1997), for which it may also be corrected (e.g., Duan & Dunlap, 1997).

The third version for the sampling variance HS3 is given by

$$\hat{\sigma}^2_{\bar{r}3} = \frac{1}{k}\left(\frac{\left[\sum_{i=1}^{k} n_i \left(r_i - \bar{r}\right)^2\right]}{\sum_{i=1}^{k} n_i}\right) \tag{5.9}$$

(Osburn & Callender, 1992, p. 116, Equation 5; see also Hunter & Schmidt, 1990, p. 100). This version of the sampling variance is supposed to "hold" for the heterogeneous case and should also perform well for the homogeneous case (Osburn & Callender, 1992, p. 116).

The fourth and last form HS4 proposed to estimate the sampling variance is given by

$$\begin{aligned}\hat{\sigma}^2_{\bar{r}4} &= \frac{\left(1-\bar{r}^2\right)^2}{(N-k)} + \frac{1}{k}\left(\frac{\left[\sum_{i=1}^{k} n_i \left(r_i - \bar{r}\right)^2\right]}{\sum_{i=1}^{k} n_i}\right) - \frac{\sum_{i=1}^{k}\left(1-r_i^2\right)^2 / (n_i - 1)}{k^2} \\ &= \hat{\sigma}^2_{\bar{r}1} + \hat{\sigma}^2_{\bar{r}3} - \hat{\sigma}^2_{\bar{r}2}\end{aligned}$$

(Osburn & Callender, 1992, p. 116, Equation 7). It is explicitly recommended for the heterogeneous case (Whitener, 1990).

In principle, each of these formulae discussed in the cited literature can be used for tests and to construct confidence intervals. In the Monte Carlo study presented in Part III all four versions will be evaluated with respect to their performance in the various situations described in Section 4.5 (for an evaluation with real data, see Fuller & Hester, 1999).

The formula to compute a standard normal deviate to test the mean effect size estimates is given for all versions by

$$g = \frac{\bar{r}}{\hat{\sigma}_{\bar{r}}}.$$

As in the previous approaches, the approximate lower and upper limits of a confidence interval are constructed by

$$\begin{aligned} r_L &= \bar{r} - g_\alpha \hat{\sigma}_{\bar{r}} \\ r_U &= \bar{r} + g_\alpha \hat{\sigma}_{\bar{r}} \end{aligned}$$

(Hunter & Schmidt, 1990, p. 121). Both for the test as well as for the construction of the confidence interval $\hat{\sigma}_{\bar{r}}$ stands for one of the four versions of the sampling variance. Although it is well-known that the correlation coefficient is not normally distributed unless n is very large (see Section 3.1), the tests may nevertheless perform as suggested by the formulae, which is due to the central limit theorem. In Chapter 8 the corresponding results on the performance of the four versions will be reported.

Homogeneity Test Q. A homogeneity test is conducted in the HS approach by using

$$Q = \frac{\sum_{i=1}^{k} (n_i - 1)\,(r_i - \bar{r})^2}{\left(1 - \bar{r}^2\right)^2}$$

(Hunter & Schmidt, 1990, p. 111). Though not labeled as such in the cited source, in essence, the above equation enables a Q-test as included in the other approaches. The tendency of the proponents of the HS approach to deny situational specificity is expressed by their suggested interpretation of the test results. They state that "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 & Schmidt, 1990, p. 112). Thus, the result of the test is taken as informative when in favor of "no true variation", that is, situational specificity, and devalued when indicating heterogeneity.

Estimation of Heterogeneity Variance. The estimation of heterogeneity variance only makes sense within the framework of a random effects model, hence it might be considered as obvious that the HS approach assumes a RE model. However, the procedures outlined above suggest the HS approach to assume a FE model because estimated heterogeneity variance is not incorporated in estimation and tests. Thus, a somewhat ambiguous case is given here, as is also evidenced by an inconsistent classification of the HS approach with respect to the FE-RE model distinction in the literature (cf. Erez et al., 1996; Field, 2001; Hedges & Olkin, 1985). The ambiguity may result for several reasons. First, the procedures outlined do not fit clearly in one of the model schemes introduced in Chapter 4. Second, the assumption of differences in universe effect sizes and therefore nonzero σ_ρ^2 is an integral part of the HS approach (Hunter & Schmidt, 1990). At the same time the authors of the approach provide procedures and many arguments to reduce observed variability in effect sizes. They do this up to a point where they conclude that universe variance is negligible and generalization of effects (across situations) is therefore possible. Additionally, they have stated that "...applications of our methods have usually used the fixed effects model described in Hedges and Olkin (1985)" (Hunter & Schmidt, 1990, p. 405) on the one hand, and also "The methods described in Hunter et al. (1982), Hunter and Schmidt (1990) [...] are RE models" (Hunter & Schmidt, 2000, p. 275) on the other hand. Such statements have certainly

contributed to the ambiguity. As a result, it is not entirely clear how the HS approach is to be classified with respect to models in meta-analysis because it is a "hybrid" type in procedures. By taking the answer to the question of whether population correlations are considered as random variables in an approach as an anchor to make the classification, the HS approach qualifies as an RE approach. Thus, it does make sense to estimate heterogeneity variance.

The procedure to estimate heterogeneity variance σ_ρ^2 as proposed in the HS approach is drawing on simply taking the following difference between variance estimators

$$\hat{\sigma}_\rho^2 = \hat{\sigma}_r^2 - \hat{\sigma}_e^2 \tag{5.10}$$

(Hunter & Schmidt, 1990, p. 106), where $\hat{\sigma}_r^2$ is used to estimate the variance of r and $\hat{\sigma}_e^2$ denotes an estimator for the sampling error variance. The reasoning to arrive at this relationship includes the assumptions that r is an unbiased (and consistent) estimator of ρ, and that an error component e and ρ in the relationship $r = \rho + e$ are independent.[4] None of these assumptions is correct in a strict sense. Nevertheless, violations of the assumptions are ordinarily not considered to be reasons for concern in practical applications of meta-analysis (see, e.g., Hedges, 1988).

A little rearrangement of Equation 5.10 shows that the variance of r is decomposed into two parts. One is the heterogeneity variance σ_ρ^2 and the other is the sampling error variance σ_e^2, where estimators are represented in Equation 5.10. Estimation of heterogeneity variance is done by computation of the following terms

$$S_r^2 = \frac{1}{N} \sum_{i=1}^{k} n_i \left(r_i - \bar{r}\right)^2$$

and

$$\hat{\sigma}_{e1}^2 = \frac{\left(1 - \bar{r}^2\right)^2 k}{N},$$

where the observed variance of correlations S_r^2 is used as an estimator for the variance of r. Again, there have been several estimators proposed for σ_e^2 in the literature, so that $\hat{\sigma}_{e1}^2$ is indexed by 1 to signify that this is a first estimator of σ_e^2. According to Hunter and Schmidt (1990, p. 107), this represents an "almost perfect first approximation". Note that this is the formula used by Johnson, Mullen, and Salas (1995), who conducted one of the first comparison between approaches. They used this estimator in the context of significance testing when conducting their comparative evaluation of the HOr, RR, and HS approaches. Of course, it is the wrong estimator of the variance of $\bar{r}$ as it estimates the expected variance in observed effect sizes due to sampling error. If k had been placed in the denominator as in Equation 5.7, it would have

[4]Note the close similarity of this basic equation to those in HLM models, which shows again that many standard meta-analytic models can be considered as special cases of the more general HLM.

been an appropriate estimator of $\sigma^2_{\bar{r}}$ in the HS approach. Thus, the resulting estimated variances were much too large in the Johnson, Mullen, and Salas study. The negative results reported by Johnson, Mullen, and Salas (1995) for the HS approach and the corresponding conclusions are therefore useless (see also Schmidt & Hunter, 1999a).

A second estimator is given as

$$\hat{\sigma}^2_{e2} = \frac{\left(1 - \bar{r}^2\right)^2}{(N/k) - 1}.$$

According to Hunter and Schmidt (1990, p. 108; see also Hunter & Schmidt, 1994b, p. 171) this is supposed to be "an even better estimate of the sampling error variance", that is, for the estimation of σ^2_e. Hence, only the second estimate was actually used in the Monte Carlo study presented in Part III. For a previous Monte Carlo study on the robustness, bias, and stability of σ^2_ρ, see Oswald and Johnson (1998) who report a negative bias of the estimators presented here under various distributional conditions.

There have been presented further estimators within the framework of the HS approach that claim to be applicable also for databases with dependent correlations and to correct for a potential underestimation in the methods presented above. However, they are not presented here (see Martinussen & Bjørnstad, 1999).

Equation 5.10 can also be regarded as the basic equation of the HS approach since many arguments pertaining to developments of the model rest on this equation. As with many procedures in the HS approach, Equation 5.10 has stimulated much criticism in the literature but arguments will not be repeated here. The interested reader is referred to the pertinent literature (e.g., Osburn & Callender, 1990; Thomas, 1989a, 1990a).

75%-Rule. A procedure unique to the HS approach is the so-called 75%-rule originally proposed by Schmidt and Hunter (1977). The reasoning behind this rule is as follows. Recall that the development of the HS approach was done with validity coefficients as the main effect size of interest and personnel selection as the most important field of application in mind. Validity coefficients are supposed to be influenced by a series of mainly methodological factors of which many can in principle be corrected for (see Hunter & Schmidt, 1994a). However, in most applications of meta-analysis all the information necessary to correct for the artifactual factors is not available so that variance in observed effect sizes due to uncorrected artifactual influences is always presumed to remain. The component supposed to account for the largest amount of observed variance (S^2_r) is sampling error. If observed variance is larger than expected by sampling error, then there may be variance in effect sizes left to be explained (i.e., $\sigma^2_\rho \neq 0$). This would represent a challenge to the hypothesis of validity coefficients not being specific to situations, where generalization across situations is a desirable state of affairs for most researchers. Consider in this context the following fraction

$$x = \frac{\hat{\sigma}^2_{e2}}{\hat{\sigma}^2_{\bar{r}3} k}.$$

An estimator of the sampling error variance in observed effect sizes is given in the numerator and the observed variance in the denominator ($S^2_r = \hat{\sigma}^2_{\bar{r}3} k$). Clearly, if there is no artifactual variance in the observed effect sizes left and no explanatory variables exist, observed or unobserved, to explain variability in effect sizes, then this fraction should lead to a value of one because observed variance is totally accounted for by sampling error. As already mentioned, not all artifactual influences can be corrected for, so the following rule of thumb has been proposed

- Homogeneity, if $x \geq 0.75$
- Heterogeneity, if $x < 0.75$

(see e.g., Hunter & Schmidt, 1990, p. 68). That is, effect sizes are considered to be homogeneous, if sampling error accounts for at least 75% of the observed variance in effect sizes, hence the name *75%-rule*.

As examples for previous Monte Carlo studies on this rule, consider Spector and Levine (1987) who found that with small k the ratio as given above is biased (i.e., larger than 1) in homogeneous situations. The ratio quickly increases as the number of k decreases, irrespective of n. In a critique of this article, Callender and Osburn (1988) showed that this result was an artifact stemming from the extremely skewed distribution of the ratio so that the expected value of the distribution of ratios, on which Spector and Levine focused, has an expected value larger than 1 although the individual comparison of estimated error variance and observed variance resulted in no bias.

Like the homogeneity test based on the Q-statistic, the 75%-rule is also taken as indicant in the HS approach of whether there are unsuspected moderators (i.e., explanatory variables) (Hunter & Schmidt, 1990, p. 440). A Monte Carlo investigation on the comparative evaluation of these tests for the detection of heterogeneity will be presented in Part III of this book (see also Cornwell & Ladd, 1993; Koslowsky & Sagie, 1993; Sackett, Harris, & Orr, 1986; Sánchez-Meca & Marín-Martínez, 1997). For a critical appraisal of the rationale of the 75%-rule, the reader is referred to James et al. (1986).

As an addition to the 75%-rule, there has also been proposed a 90%-rule with the same rationale as outlined above, but with a cut-off value of .90 for x that is supposed to be more suitable for Monte Carlo studies in which no artifactual variance exists (Sackett et al., 1986). This rule is also considered in the results to be reported in Chapter 8.

5.4 REFINED APPROACHES

Up to this point, the three main approaches to meta-analysis in the field of psychology have been presented. In the present section, two further sets of

procedures will be introduced, one approach for RE models and another that is suitable both in the FE as well as RE model.

5.4.1 DerSimonian-Laird

The most prominent RE approach in psychology draws on the derivations as given by DerSimonian and Laird (1983, 1986) and will be labeled DSL in the present context. Although it is almost identical with the procedures outlined in Section 4.2, computational procedures are given in this section for completeness and reference.

Estimation of Heterogeneity Variance. The heterogeneity variance is presented first for this approach. This is due to the fact that it is used in the estimator of the mean effect size and significance testing, both of which are presented subsequently. Note that the Fisher-z transformation is used in this approach, so that the variance σ_ζ^2 is of interest, that is, the variance of the universe parameters in z-space. The heterogeneity variance is estimated for correlations as effect size data by the moment estimator

$$\hat{\sigma}_\zeta^2 = \frac{Q - (k-1)}{a},$$

where

$$a = \sum_{i=1}^{k} w_i - \left[\sum_{i=1}^{k} w_i^2 \Big/ \sum_{i=1}^{k} w_i \right].$$

This estimator is unbiased by construction. Ordinarily, $\hat{\sigma}_{\zeta+}^2 = \max\{0, \hat{\sigma}_\zeta^2\}$ is used in applications because $\hat{\sigma}_\zeta^2$ may be negative. $\hat{\sigma}_{\zeta+}^2$ can be called a *truncated estimator* which is no longer unbiased (see also Böhning et al., 2002). There have been published several tests of the quality of this estimator and also alternative estimators have been proposed. They will not be dealt with here and the reader is therefore referred to the relevant literature (e.g., Böhning, 2000; Biggerstaff & Tweedie, 1997; Friedman, 2000; Malzahn, 2003; Malzahn, Böhning, & Holling, 2000).

Estimation of Mean Effect Size. The mean effect size is estimated in the DSL approach by a weighted estimator as follows

$$\bar{z} = \frac{\sum_{i=1}^{k} w_i^* z_i}{\sum_{i=1}^{k} w_i^*},$$

where

$$w_i^* = \left(\frac{1}{n_i - 3} + \hat{\sigma}_\zeta^2 \right)^{-1}.$$

Estimation of the mean effect size follows the procedures as outlined within the general framework of the RE model in Section 4.2. As was shown, the procedures of the RE and FE model differ mainly with respect to the weight used in computations. For the present case, note that there is a special case for which the mean effect size as given above for the DSL approach would be *identical* to the one resulting from the application of FE model procedures as specified for the HO*r* approach. This would be the case if the number of persons per study were constant across studies because both parts of the sum to compute the weights (i.e., $(n_i - 3)^{-1}$ and $\hat{\sigma}_\zeta^2$) are the same for all studies to be aggregated. In other words, in situations of equal n for all studies the estimate of the mean effect size of DSL will not differ from HO*r*. This is due to the fact that the variances of the Fisher-z transformed estimators *only* depend on n. When n is equal for all studies, the weights do not differ. However, when n is different for the studies under investigation the weights will mostly differ between HO*r* and DSL estimators and different estimates may result in practical applications. This should be borne in mind since the design of the Monte Carlo study in Part III will be characterized by a constant n for all studies.

Significance of Mean Effect Size. Significance tests are performed in a usual form by using the test statistic

$$g = \frac{\overline{z}}{\hat{\sigma}_{\overline{z}}}$$

with

$$\hat{\sigma}_{\overline{z}} = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i^*}},$$

so that g can be compared with the critical value from the standard normal distribution for a desired level α.

Approximate lower and upper limits of the confidence interval are constructed by

$$z_L = \overline{z} - g_\alpha \hat{\sigma}_{\overline{z}}$$
$$z_U = \overline{z} + g_\alpha \hat{\sigma}_{\overline{z}}\,.$$

Again, the confidence limits are customarily transformed into r-space subsequently by the inverse Fisher-z transformation.

5.4.2 Olkin and Pratt

The last approaches to be presented are based on an early publication by Olkin and Pratt (1958) on the unbiased estimation of the correlation coefficient (see also Section 3.1), which was applied to the problem posed in meta-analysis by Hedges (1988, 1989; see also Hedges & Olkin, 1985).

Estimation of Mean Effect Size. The estimation of the mean effect size draws on the UMVU estimator G proposed by Olkin and Pratt (1958) (already given in Equation 3.5 on page 26). The following formula repeats the approximation of G also given in Section 3.1.

$$G_i = r_i \left(1 + \frac{1 - r_i^2}{2\,(n_i - 1 - 3)} \right).$$

As a first version of an estimator for the mean effect size, consider

$$\overline{G} = \frac{\sum_{i=1}^{k} n_i G_i}{\sum_{i=1}^{k} n_i}.$$

The estimator and further computational procedures using this estimator will be labeled as OP approach.

A second version of the estimator is established in analogy to the procedures in the FE model. To compute the weights for aggregation according to the FE model the variance of this estimator is needed. The variance of G is given by Equation 3.7 on page 27. Defining the weights $w_{i(\mathrm{FE})}$ as usual in the FE approach as $\hat{\sigma}_G^{-2}$, the weighted estimator is given by

$$\overline{G}_{\mathrm{FE}} = \frac{\sum_{i=1}^{k} w_{i(\mathrm{FE})} G_i}{\sum_{i=1}^{k} w_{i(\mathrm{FE})}}.$$

Since the weights are constructed as is common in the FE model, this will be labeled the OP-FE approach. Recall that in contrast to z-based approaches and HS, the variance strongly changes across values of ρ. This may have a profound influence on the results when applying this approach. Especially when n is small and estimates thus vary strongly, biased results may emerge. This is due to the facts that, first, the variances are smaller for larger absolute values of ρ (see Figure 3.4) and, second, the (strongly varying) r_i are plugged into Equation 3.7 to obtain estimates of the variance of G. Hence, in applying this procedure high correlations emerging by chance will receive a high weight and an upward bias may result in mean effect size estimation.

A third estimator that draws on the general procedures for the RE model is presented next. It uses weights that incorporate an estimate of heterogeneity variance that is given in the last paragraph for this approach. The weights in the random effects version are designated as $w_{i(\mathrm{RE})}$ and are given by $(\hat{\sigma}_\rho^2 + \hat{\sigma}_G^2)^{-1}$. They are used to estimate $\overline{G}_{\mathrm{RE}}$ as follows

$$\overline{G}_{\text{RE}} = \frac{\sum_{i=1}^{k} w_{i(\text{RE})} G_i}{\sum_{i=1}^{k} w_{i(\text{RE})}}.$$

This estimator as well as related computational procedures employing it will be labeled the OP-RE approach.

Significance of Mean Effect Size. The test for the OP approach draws on the fact that G has the same asymptotic distribution as r (Olkin & Pratt, 1958; Hedges & Olkin, 1985). As a result, approximately the same standard error is assumed which is estimated by

$$\hat{\sigma}_{\overline{G}} = \frac{1 - \overline{G}^2}{\sqrt{N-k}}. \tag{5.11}$$

The authors also state that G has larger variance than r so that the proposed estimator can be considered to be only an approximation. Interestingly, this approximation has already been used in a Monte Carlo study on combined estimators for the universe correlation by Viana (1982).

For the OP-FE approach, the standard error is computed by

$$\hat{\sigma}_{\overline{G}_{\text{FE}}} = \left(\sum_{i=1}^{k} w_{i(\text{FE})} \right)^{-\frac{1}{2}}$$

and correspondingly for the OP-RE approach by

$$\sigma_{\overline{G}_{\text{RE}}} = \left(\sum_{i=1}^{k} w_{i(\text{RE})} \right)^{-\frac{1}{2}}.$$

Therefore,

$$g = \frac{\overline{G})}{\sigma_{\overline{G}}}, \qquad g = \frac{\overline{G}_{\text{FE}}}{\sigma_{\overline{G}_{\text{FE}}}}, \qquad g = \frac{\overline{G}_{\text{RE}}}{\sigma_{\overline{G}_{\text{RE}}}}$$

are g-values to be compared with a critical value from the standard normal distribution for the OP, OP-FE and OP-RE approach, respectively.

The confidence limits are constructed by

$$r_L = \overline{G} - g_\alpha \hat{\sigma}_{\overline{G}}$$
$$r_U = \overline{G} + g_\alpha \hat{\sigma}_{\overline{G}}$$

for the OP approach, for the other approaches they are constructed analogously.

Homogeneity Test Q. For the homogeneity test, only OP-FE is considered. For this approach, the test statistic is computed as

$$Q = \sum_{i=1}^{k} w_i \left(G_i - \overline{G}_{\text{FE}}\right)^2.$$

Estimation of Heterogeneity Variance. The estimated variance of G is used to estimate the heterogeneity variance by

$$\hat{\sigma}_\rho^2 = S_G^2 - \frac{1}{k}\sum_{i=1}^{k} \hat{\sigma}^2_{\overline{G}_{\text{FE}}},$$

where S_G^2 is the observed variance of the Olkin-Pratt estimator

$$S_G^2 = \frac{1}{k}\sum_{i=1}^{k} \left(G_i - \overline{G}_{\text{FE}}\right)^2$$

(Hedges, 1988, p. 198; see also Hedges, 1989, pp. 473–474). Again, estimation is restricted to usage of the estimated variance of the OP-FE approach.

5.5 CONSEQUENCES OF CHOOSING AN APPROACH: DIFFERENT ESTIMATED PARAMETERS

After having outlined statistical details of several approaches, some consequences of choosing between approaches will be examined in this section. The common assumption that the choice of an approach is largely inconsequential for the results is thereby scrutinized and challenged. The treatment will be restricted to a theoretical examination. An empirical Monte Carlo study will be presented in the subsequent Part III of the book to validate some predictions derived from theoretical results presented in the present section and to comparatively evaluate the performance of the procedures as proposed in the approaches.

In the present section, the focus will be kept on the expected value and variance of the mixing distribution as parameters of interest in meta-analysis. It will become evident that one of the main differences between the approaches as outlined in this chapter are differences in the use of effect sizes. That is, whether correlation coefficients are used without any transformations or transformed to Fisher-z or d, respectively. Also, the focus will be laid on $\mathfrak{S}_2$ because, on the one hand, there are no relevant modifications of universe parameters in a homogeneous case ($\mathfrak{S}_1$), and, on the other hand, the general problems outlined in the current section readily generalize to $\mathfrak{S}_3$.

There are two different values ρ_1 and ρ_2 in the universe of studies in $\mathfrak{S}_2$. As specified in Section 4.5, both values have equal probability so that for estimators of the expected value of the mixing distribution based on r values it would

be natural to consider the mean of the two universe correlations given by

$$\mu_\rho = \frac{\rho_1 + \rho_2}{2} \tag{5.12}$$

as the parameter to be estimated. This is simply the mean of the two different universe correlations. Of course, it would be reasonable in such a situation not only to estimate a single parameter of the effect size distribution, but — if possible — to identify the classes and estimate ρ_1 and ρ_2 separately in an analysis with HLM, for example. As already stated, an evaluation of such procedures is not the aim of the present context. Instead, the focus will be on an evaluation of the weighted mean effect size as an estimator of the expected value of the mixing distribution.

With regards to the expected value of the mixing distribution, it would intuitively be equally natural to expect the estimators of all approaches to estimate the parameter μ_ρ. To the best of the author's knowledge, all applied meta-analyses on issues of substantive interest which used any of the approaches applying a transformation of the correlation coefficient, seem to presume this. That is, mean effect size estimates are interpreted as if they estimated a mean universe correlation. What this exactly means in applications of meta-analysis is rarely explicated but it seems as if in every case a mean correlation as given in Equation 5.12 was implied. The question to be dealt with here is whether such an interpretation is valid. This is not the case because in contrast to estimators based on r, estimators based on the Fisher-z transformed correlation coefficients (HOr, HOT, RR, DSL) do not estimate a "mean ρ" in the universe of studies but

$$\begin{aligned} \mu_{\rho z} &= \tanh \mu_\zeta \\ &= \tanh \left(\frac{\zeta_1 + \zeta_2}{2} \right) \\ &= \tanh \left(\frac{\tanh^{-1}(\rho_1) + \tanh^{-1}(\rho_2)}{2} \right). \end{aligned} \tag{5.13}$$

It is important to note that $\mu_{\rho z}$ is the expected value *in the space of* ρ that results from the inverse Fisher-z transformation of the expected value μ_ζ of ζ. Hence, the computation of the expected value is carried out in z-space and the result is transformed via the inverse Fisher-z transformation to arrive at an expected value of ρ. To distinguish the expected value of ρ for which computations are carried out in r-space (i.e., μ_ρ) from the one for which computations are done in z-space, a double index is used in $\mu_{\rho z}$ to indicate the origin from another space.

As shown in Equation 5.13, for the given case $\mathfrak{S}_2$ the mean of ζ_1 and ζ_2 transformed to a mean ρ using the inverse Fisher-z transformation is $\mu_{\rho z}$. The focal question is: Is it true for all combinations of ρ in $\mathfrak{S}_2$ that $\mu_\rho = \mu_{\rho z}$? If it were true, then a differentiation of μ_ρ and $\mu_{\rho z}$ would not be necessary and the

aforementioned interpretation of mean effect size estimates based on Fisher-z transformed correlations would be correct.

As already stated, this is not the case and it is quite important to make this distinction since an inverse Fisher-z transformation of μ_ζ does *not* lead to μ_ρ in general. Only when $\rho_1 = \rho_2$, that is in the homogeneous case $\mathfrak{S}_1$, does $\mu_\rho = \mu_{\rho z}$ hold. For the case of only two different ρs, Equation 5.13 can equivalently be expressed as

$$\mu_{\rho z} = \frac{\sqrt{1+\rho_1+\rho_2+\rho_1\rho_2} - \sqrt{1-\rho_1-\rho_2+\rho_1\rho_2}}{\sqrt{1+\rho_1+\rho_2+\rho_1\rho_2} + \sqrt{1-\rho_1-\rho_2+\rho_1\rho_2}} \tag{5.14}$$

in terms of the original ρs. This equation makes it clearer that μ_ρ equals $\mu_{\rho z}$ *only* when ρ_1 and ρ_2 are the same. It may be noted that Olkin (1967, p. 116) has already provided an expression similar to the one given above when considering the weighted average of correlation coefficients from two independent populations with a *common* ρ, a problem not exactly the same as in the present context.

It is important for meta-analysis in general that Equation 5.14 is not restricted to $\mathfrak{S}_1$ and can be generalized beyond this restricted situation. In fact, it can be generalized to an arbitrary number of different values ρ. The following result provides such a general expression for which Equation 5.14 can be regarded as a special case.

By induction we have the following

Lemma. *For all c and $\rho = (\rho_1, \ldots, \rho_c)$ we have*

(i) $\prod_{j=1}^{c}(1+\rho_j) = \sum_\alpha \rho^\alpha$

(ii) $\prod_{j=1}^{c}(1-\rho_j) = \sum_\alpha (-1)^{|\alpha|}\rho^\alpha$

where summation extends over all $\alpha \in \{0,1\}^c$ satisfying $|\alpha| \le c$.

Note that $\alpha = (\alpha_1, \ldots, \alpha_c)$, $|\alpha| = \sum \alpha_j$, and $\rho^\alpha = \rho_1^{\alpha_1} \times \cdots \times \rho_c^{\alpha_c}$. Now, let $\rho = (\rho_1, \ldots, \rho_c)$ and $z = (z_1, \ldots, z_c)$ be the vector of corresponding Fisher-z values. Define $h(\rho) = \tanh(\bar{z})$. Then

Theorem.

$$h(\rho) = \frac{(\sum_\alpha \rho^\alpha)^{1/c} - (\sum_\alpha (-1)^{|\alpha|}\rho^\alpha)^{1/c}}{(\sum_\alpha \rho^\alpha)^{1/c} + (\sum_\alpha (-1)^{|\alpha|}\rho^\alpha)^{1/c}}$$

Proof.

$$h(\rho) = \frac{\left(\prod_{j=1}^{c}\frac{1+\rho_j}{1-\rho_j}\right)^{1/c} - 1}{\left(\prod_{i=j}^{c}\frac{1+\rho_j}{1-\rho_j}\right)^{1/c} + 1} = \frac{\left(\frac{\prod_{j=1}^{c}(1+\rho_j)}{\prod_{j=1}^{c}(1-\rho_j)}\right)^{1/c} - 1}{\left(\frac{\prod_{j=1}^{c}(1+\rho_j)}{\prod_{j=1}^{c}(1-\rho_j)}\right)^{1/c} + 1}$$

The result then is a consequence of the above given Lemma. □

As an example to see how the generic form of the expression $h(\rho)$ works for cases other than the two-point distribution of focal interest, consider the case of three different ρ. Let $\rho = (\rho_1, \rho_2, \rho_3) = (.10, .50, .90)$. Then $(\sum_\alpha \rho^\alpha)^{1/c}$ and $(\sum_\alpha (-1)^{|\alpha|}\rho^\alpha)^{1/c}$ expand to

$$\left(\sum_\alpha \rho^\alpha\right)^{1/c} = \sqrt[3]{1 + \rho_1 + \rho_2 + \rho_3 + \rho_1\rho_2 + \rho_1\rho_3 + \rho_2\rho_3 + \rho_1\rho_2\rho_3} = \sqrt[3]{a}$$

$$\left(\sum_\alpha (-1)^{|\alpha|}\rho^\alpha\right)^{1/c} = \sqrt[3]{1 - \rho_1 - \rho_2 - \rho_3 + \rho_1\rho_2 + \rho_1\rho_3 + \rho_2\rho_3 - \rho_1\rho_2\rho_3} = \sqrt[3]{b},$$

so that

$$\begin{aligned} h(\rho) &= \frac{\sqrt[3]{a} - \sqrt[3]{b}}{\sqrt[3]{a} + \sqrt[3]{b}} \\ &= \frac{\sqrt[3]{3.135} - \sqrt[3]{0.045}}{\sqrt[3]{3.135} + \sqrt[3]{0.045}} = .61. \end{aligned}$$

The resulting value of $h(\rho) = .61$ shows that the use of Fisher-z yields an overestimation in comparison to $\mu_\rho = .50$, but now for the case of a three-point mixing distribution. As can also be easily recognized, the task to explicitly specify the expression for cases with more than three different ρ becomes rather laborious, though widely available computing resources make it accomplishable.

To give a more comprehensive impression of how large the differences can get in $\mathfrak{S}_2$, a series of differences $\mu_\rho - \mu_{\rho z}$ for varying positive ρ_1 and ρ_2 were computed and are depicted in Figure 5.1. The differences in μ_ρ and $\mu_{\rho z}$ for varying ρ_1 and ρ_2 are portrayed with a surface to enhance visibility of the trends.

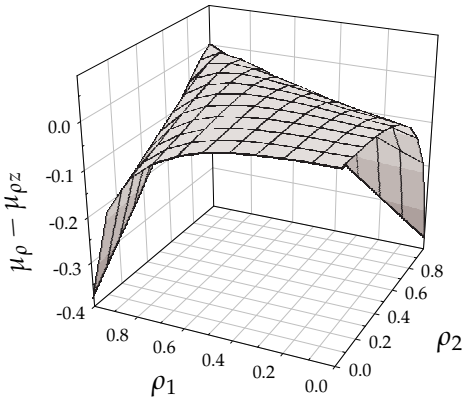

Figure 5.1 Differences between μ_ρ and $\mu_{\rho z}$ by different ρ_1 and ρ_2.

As is evident, μ_ρ is always smaller than $\mu_{\rho z}$ when $\rho_1 \neq \rho_2$. For the homogeneous case there is a ridge from the lower corner of the graph to the upper at a height of zero, indicating the equality of μ_ρ and $\mu_{\rho z}$ for this parameter

constellation. With growing differences between the two ρs, differences in expected values become increasingly larger up to values of approximately $-.35$ in extreme cases.

There are several implications of this observation. Most importantly, the estimated parameters in the universe are indeed *different* for the estimators. That is, the Fisher-z transformation introduces a different estimated parameter through its nonlinear transformation of the correlation coefficients in heterogeneous situations. In general, approaches that employ the Fisher-z transformation will always result in higher absolute values for the estimate of the mean effect size in such situations. This may not be entirely clear to every research consumer of meta-analyses when interpreting the results. Second, as a result Fisher-z based estimators may be regarded as inappropriate as estimators of μ_ρ because estimates will necessarily differ from this parameter as illustrated in Figure 5.1. Although the differences as large as the extreme cases depicted in the figure will probably be easily identified in applications of meta-analyses for a two-point mixing distribution in the universe, by simple inspection of the effect size distribution, smaller differences may remain undetected. Furthermore, simple detection of such cases may become quite difficult with discrete mixing distributions with more support points than two, especially when these are fairly close to each other. The application of the Fisher-z transformation will in these cases inflate the mean effect sizes in relation to μ_ρ, a fact that underscores the importance of homogeneity tests.

Another implication of the fact that Fisher-z based procedures estimate $\mu_{\rho z}$ and not μ_ρ in heterogeneous cases is that it would be somewhat unfair to judge the quality of z-based estimators by comparison with μ_ρ, a parameter they are not supposed to estimate. Rather than discarding Fisher-z based estimators from analyses in $\mathfrak{S}_2$ to be reported for the Monte Carlo study in Part III, the parameters for comparisons of the estimators correspond to the value they actually estimate, with μ_ρ as the value for estimators based on r and $\mu_{\rho z}$ for estimators based on Fisher-z transformed values. The parameters thus will be chosen to match the parameter to be estimated when reporting results of the Monte Carlo study in Chapter 8.

Basically the same is true for estimators based on d. A similar terminology is used to examine this issue. Hence, the expected value in the space of r that results from transforming an expected value computed in d-space and subsequently transformed into r-space by way of Equation 3.11 will be denoted by $\mu_{\rho d}$. As was shown in Section 3.3, the transformation of r to d has a similar functional form in comparison to the Fisher-z transformation. Accordingly, also a similar form for the difference between μ_ρ and $\mu_{\rho d}$ is expected and indeed will be given as shown in Figure 5.2.

The graph depicted is slightly steeper in the tails of large ρ differences as the transformation suggests. The values for $\mu_{\rho d}$ were computed in analogy to Equation 5.13 with the r to d transformation applied to ρ_1 and ρ_2 and the inverse transformation from d to r applied to the mean δ resulting in $\mu_{\rho d}$. As was the case for the Fisher-z transformation, a ridge for equal values of ρ_1 and ρ_2 indicates the equality of μ_ρ and $\mu_{\rho d}$ in the homogeneous case.

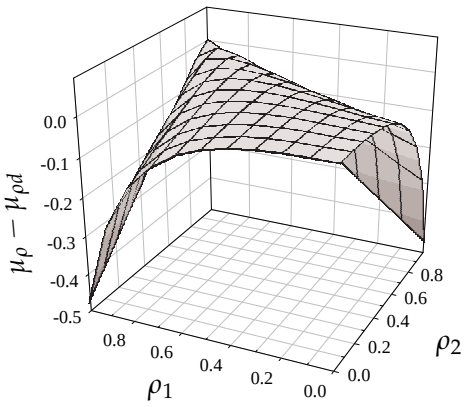

Figure 5.2 Differences between μ_ρ and $\mu_{\rho d}$ by different ρ_1 and ρ_2.

However, values for the mean effect size using HOd are not very close to $\mu_{\rho d}$ as the general logic outlined here would suggest. Actually, they are much closer to μ_ρ. How can this be the case if HOd is assumed to be an estimator of $\mu_{\rho d}$? The reason for this effect lies in the confoundation of the employed weights with δ when aggregating the d values. Holding n constant, the weights are dependent on the parameter δ or estimates thereof, respectively (see Section 3.2). The effect of using these weights is to downweight higher d. Assume equal n in two groups and recall that the weights are the reciprocals of σ_d^2, then by holding n constant, σ_d^2 increases with d, and the weights, being reciprocals of σ_d^2, decrease. The form of the relationship between σ_d^2 and d is illustrated in Figure 5.3.

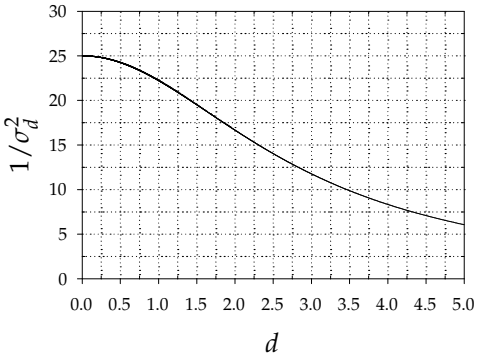

Figure 5.3 Reciprocals of σ_d^2 by d.

In this figure, an n of 100 was assumed for computing the weights and values are depicted up to $d = 5$, which corresponds to $r \approx .93$. There is a clear trend for decreasing weights with increasing d. This leads to mean d values being much closer to, but not exactly at, μ_ρ in comparison to $\mu_{\rho d}$. A selection of varying values for $\mathfrak{S}_2$ is presented in Table 5.1 along with values for μ_ρ, $\mu_{\rho d}$ and a weighted version of $\mu_{\rho d}$. The latter was computed by applying the weights to the population parameters when aggregating.

Table 5.1 Comparison of Theoretical Values of μ_ρ and $\mu_{\rho d}$ in $\mathfrak{S}_2$

ρ_1	ρ_2	μ_ρ	$\mu_{\rho d}$	$\mu_{\rho d}(w)$
.00	.10	.05	.0502	.0501
.00	.20	.10	.1015	.1005
.00	.30	.15	.1553	.1517
.00	.40	.20	.2132	.2039
.00	.50	.25	.2774	.2575
.00	.60	.30	.3511	.3123
.00	.70	.35	.4401	.3675
.00	.80	.40	.5547	.4191
.00	.90	.45	.7183	.4470

Note. The n was fixed at 100 for all values of w. $\mu_{\rho d}(w)$ is the weighted version of $\mu_{\rho d}$.

It is evident by comparison of columns three to five that the weighted version of $\mu_{\rho d}$ leads to results much closer to μ_ρ for larger differences between ρ_1 and ρ_2. Since the weights are not chosen to produce this effect it can be described as somehow incidental. However, recognizing this effect, it would not be reasonable to compare mean effect sizes based on d with an unweighted version of δ, at least not for larger differences. Hence, the results for the bias of the estimators, for example, to be presented in Chapter 8 are based on comparisons between μ_ρ and mean effect size estimates based on d.

Although the estimated parameter for r-based approaches is μ_ρ, there may also arise problems for some approaches in estimating this parameter when the variances of r or G are used in computing the weights for aggregation. The approaches for which this problem may be relevant are OP-FE and to a smaller degree also for OP-RE. The latter also employs estimates of the heterogeneity variance that are equal for all studies to be aggregated so that weights depend on the variance of the estimate to a lesser degree. As already mentioned, this homogenizes the weights.

The problem of this dependency is exacerbated when n is low. In such situations, observed correlation coefficients are highly variable. Theoretically, the variances of estimates are the same in this situation. However, due to the fact that the (highly variable) estimates of the universe parameter (r or G) are used in estimating their variances, the variances also vary strongly and therefore so do the weights. Because there is a relationship of high or low weights occuring along with high or low estimates, a bias in the pooled estimate may be introduced by plugging in the estimates of the variances in the computation of the weights. In cases with (nearly) equal n, it would thus be sensible to estimate the variance of the estimates based on an n-weighted pooled estimate of all effect sizes available. Such a procedure is employed, for example, in the HS approach as outlined in this chapter. Moreover, the problem does not pertain to HS at all for the reciprocals of the variances are not used as weights for the pooled estimate.

In sum, the choice of an approach is often associated with a choice of effect size measures for computations. As shown here, this may have profound effects in heterogeneous situations. The generally applied interpretation of mean effect sizes based on correlation coefficients as estimates of the expected value μ_ρ of the mixing distribution only holds for r-based procedures but not in general for procedures based on transformations of r, since transformations have also to be applied at the level of parameter values (universe of studies).

With regard to the commonly applied Fisher-z transformation this places some remarkable constraints on its usefulness in the context of meta-analysis. It is essential when this transformation is applied for aggregating effect sizes to guarantee the homogeneous case on theoretical or empirical grounds. Otherwise, the mean effect size does not in general estimate what is mostly intended to be estimated, namely μ_ρ. Of course, it may not be an easy task to interpret mean effect sizes in the heterogeneous case without explicitly modeling the situation adequately by application of HLM or mixture modeling, for example. But cases are not uncommon at all in which explanatory variables are not available and the effect size database remains heterogeneous. When using r-based approaches, interpretation of mean effect sizes as estimates of μ_ρ is theoretically founded, whereas for approaches that apply transformations it is not. In the case of HOd, the situation is much more complicated in comparison to HOr because weights also have to be taken into account. To be sure, the expected value μ_ρ is more adequate for most cases in $\mathfrak{S}_2$ treated here, but it is not the parameter to be estimated by HOd from a theoretical point of view. In the Monte Carlo study in Part III an evaluation of the precision of estimates will be reported with respect to the parameters to be estimated as reported in this section.

5.6 COMPARISONS OF APPROACHES: STATISTICAL PROCEDURES

The approaches presented in previous sections are a set of procedures and techniques that has become very common in the application of meta-analysis in psychology and other social science disciplines. The procedures outlined are not the only available. There are even more *statistical* refinements and procedures to be found in the literature (e.g., Kraemer, 1983; Viana, 1980, 1982) than have been presented and referenced up to this point. However, the focus of the following paragraphs will be laid on the more common procedures and their properties. Furthermore, the comparison largely implies correlation coefficients as effect sizes. Some of the following statements might have to be altered when comparing proposed procedures for other effects sizes.

The first characteristic used to distinguish the approaches is the assumed model. Among the approaches considered, the majority can be classified as FE approaches. This also mirrors current research practice, in that procedures based on the FE model are still the most often applied. Approaches based on RE models have been repeatedly called for (Hunter & Schmidt, 2000; National

Research Council, 1992) but this does not yet seem to have had a profound effect on research practice. The FE approaches are HO*r*, HOT, RR, HO*d*, OP, and OP-FE whereas the RE approaches are DSL and OP-RE. Being a hybrid type, the HS approach is not easy to classify for various reasons (see Section 5.3) but it seems to be more of an RE approach in nature. Yet, it is noteworthy that others have classified it as an FE approach (e.g., Erez et al., 1996; Overton, 1998).

The HS approach also stands out somewhat for its peculiar procedures, like the 75%-rule, which is not included in other approaches. A feature of this approach that is very much emphasized by Hunter and Schmidt (1990) are the various techniques to correct for artifacts. These are not of concern in the present context but it should be recognized that an important research problem is addressed with such corrections. Although distinctive in emphasis and elaboration, corrections of effect sizes are not unique to the HS approach (see also Hedges & Olkin, 1985, pp. 131).

With the FE and RE model as outlined in Chapter 4, it is easy to recognize the common structure of the approaches as far as estimation of the mean effect size and inferential procedures are concerned. The commonalities go so far that, in fact, HO*r* and RR are largely indistinguishable and may not count as different approaches at all. Again, it is recognized that they have been classified as such in previous comparisons (e.g., Johnson, Mullen, & Salas, 1995).

A second characteristic for comparing the statistical procedures of the approaches is the effect size measure used in synthesizing correlation coefficients. As has been outlined in the previous section, important differences exist when transformations of the correlation coefficient are applied. This makes the aggregated effect size measure a quite important characteristic, at least in heterogeneous situations. The *r*-based approaches are HS, OP, OP-FE, and OP-RE. The Fisher-*z*-based approaches are HO*r*, HOT, RR, and DSL, whereas HO*d* uses another transformation that also leads to a different estimated parameter in the universe of studies in heterogeneous situations. Regarding bias of the estimators it is expected that the approaches may lead to quite accurate results only with respect to the corresponding estimated parameters.

The third characteristic to compare or classify approaches is the weighting scheme. Whereas some approaches use so-called optimal weights (i.e., reciprocals of squared standard errors), others simply use the individual study sample size as weights in their procedures. To classify the approaches with respect to this attribute, recall that the optimal weights for the approaches using the Fisher-*z* transformation are in essence determined by the sample sizes. This can be seen by inspecting Equation 5.1 on page 57 for HO*r*, for example. Hence, in these cases, approaches can as well be classified as using n_i as weights because the differences are minuscule in general. As a consequence, almost all of the presented approaches use the sample size as weights, except for OP-FE, OP-RE, and HO*d*.

Now, does the weighting scheme really make a difference? At least some expectations retrievable in the literature suggest that this is not the case. For example, Huffcutt (2002) clearly states that "it is unlikely that the choice of

weighting method has any real influence on the mean effect size [...] estimates" (p. 209). Furthermore, Sánchez-Meca and Marín-Martínez (1998b) have not reported any striking differences between weighting methods for d as an effect size measure on the basis of their Monte Carlo study results. It is nevertheless argued that the weighting does make difference.

The reasons for this are, first, that the empirical evidence available is based on the effect size d, and this is a special case as has already been scrutinized. The theoretical analysis in Section 5.5 has revealed an effect of the weights which might have obscured a profound effect of weighting in the Monte Carlo study by Sánchez-Meca and Marín-Martínez (1998b). Their results may therefore not generalize to the present case of interest, correlation coefficients. Second, recall the dependency of the weights for the UMVU estimator on ρ and also bear in mind the potential variability of effect sizes due to sampling error. Taking further into account that the observed effect sizes have to be plugged into the estimator for the standard error reveals that using such weights will lead to an upward bias in mean effect size estimators. This is exactly what can be expected for the estimators in OP-FE and OP-RE.

Hence, the weighting scheme *is* an important classification aspect for approaches, at least in cases for which a similar plug-in procedure is used as in OP-FE and OP-RE. How, then, can these be the statistically optimal weights? The reason is simply that to prove the optimum properties of this weighting scheme, one has to assume that the weights are *known*. Because this is almost never the case, one has to use the plug-in procedure which causes the problem and makes the weighting scheme suboptimal. For a theoretical analysis and empirical demonstration of the considerable effect of using plug-in estimates in the context of estimating heterogeneity variance, see Böhning et al. (2002).

Of the approaches introduced, HOd is somewhat special. It is hardly comparable to the other approaches because in the way it is used in the present examination it would almost never be used in practice (i.e., a database consisting only of r would ordinarily not be converted to d). Remember that the approach was introduced to show how correlation coefficients converted to d would be aggregated. It is intended to enable a test of the common assumption that the well-known conversion of r to d does not have an effect on the results of meta-analysis.

There are some empirical comparisons of meta-analytical approaches in the literature available to date. One quite influential early comparison that has raised serious doubts on the quality of the HS approach was conducted by Johnson, Mullen, and Salas (1995). They compared the approaches HOd, RR, and HS by analyzing a small database which they also transformed by adding constant values, for example. Hence, they have not conducted a Monte Carlo study but analyzed a specific dataset and its transformations to examine the quality of the approaches. Unfortunately, there are several problems with this comparison. First, they stated with reference to the techniques proposed by Hedges and Olkin (1985) that "...study outcomes usually are converted into standard deviation units..." (Johnson, Mullen, & Salas, 1995, p. 95). Hedges and Olkin actually do not advocate transformations of r to d as a standard

technique applied to correlation coefficients. Instead, they provide elaborate procedures for the analysis of correlation coefficients, as can be seen in Section 5.1. Although sometimes approaches as presented in this book are associated with certain effect sizes, it is not true for any of the approaches outlined that they can *only* be applied to a certain kind of effect size family. Admittedly, the HS approach has a main focus on correlations, but is not limited to the analysis of this effect size measure. It is therefore important to recognize that the present examination evaluates the procedures of the approaches that are proposed for correlational data and may not generalize to other procedures proposed. Second, the formula for standard error in the HS approach as used by the authors (Johnson, Mullen, & Salas, 1995, Formula 12, p. 97) is wrong and leads to strong overestimates of the standard error of the mean effect size (see also Schmidt & Hunter, 1999a). Third, Johnson et al. tried to vary certain "parameters of the databases [...] while attempting to hold all other variables constant..." (Johnson, Mullen, & Salas, 1995, p. 99). Unfortunately, there was a (linear) relationship in the database between r and n ($r = .158$) that influenced the results of their comparisons between the approaches. In sum, their comparison is only of limited value for a comparative evaluation of the approaches under consideration.

Despite these problems, the Johnson et al. study may have had a profound effect on other researchers and may have led them to abstain from using the HS approach. Others even tried to "explain" the divergence from conventional statistical expectations that was reported in the Johnson et al. study for the HS approach (e.g., Erez et al., 1996, p. 283). Nevertheless, the Johnson et al. study had at least the beneficial effect of drawing the attention of researchers to the potentially diverging approaches in psychology.

Another more recent comparison of approaches focusing on correlation coefficients as effect sizes was done by Field (2001). This study was not plagued with the problems of the Johnson et al. study and developed this work by conducting a Monte Carlo study. Field (2001) reported a series of results on the estimation of the mean effect size, significance test for the mean effect size, and homogeneity test performance. In separate Monte Carlo studies the performance of the approaches in homogeneous as well as heterogeneous situations was examined. Interestingly, his results indicated a bias in estimating the mean effect size in heterogeneous situations being very much larger for the approach using the Fisher-z transformation (DSL) in comparison to the HS approach. In contrast, such effects were not observed in homogeneous situations (here, HOr was compared to HS). A clear theoretical rationale for this effect was, however, lacking.[5] For more detailed results the reader might wish to consult the original article. Overall, the reported results seemed to favor the

[5]It might be noted that Hunter et al. (1982, p. 42) already pointed to the excess bias resulting from the Fisher-z transformation. In later work (Hunter & Schmidt, 1990, pp. 216–217), they repeated this observation but still without providing an elaborate statistical argument. They only pointed to an (still) unpublished paper which was referenced by Field (2001) to support his prediction. Hence, an elaborate theoretical argumentation as given in the previous chapter has not yet been available.

HS approach over other approaches mainly on the grounds of larger bias in heterogeneous situations for Fisher-z based approaches.

A further comparison, namely of the DSL and HS approach, was conducted by Hall and Brannick (2002). Although large parts of the study focused on artifact corrections, which are not of interest here, they reported some results worth noting in the present context. In a comparison between the approaches based on Monte Carlo study data, a similar pattern of results with respect to the bias of the mean effect size estimators was observed as in the Field (2001) study. That is, in homogeneous situations both approaches lead to approximately equal results and in heterogeneous situations the results differed. Differences grew bigger the larger the variance of universe parameters was, with DSL leading to overestimates. This result is perfectly compatible with expectations on the basis of the theoretical analyses of the estimators' properties outlined in this chapter. Hall and Brannick (2002), however, attributed this observation to some peculiarity of their Monte Carlo procedure. Most interesting for the present examination of approaches are reanalyses of four published meta-analyses. The authors reported higher estimates of the DSL approach in comparison to the HS approach on the basis of the four real datasets, though in one case the estimates were virtually identical. The maximum difference was between a value of .237 (HS) and .286 (DSL) for one study. This difference is remarkable and might have been even bigger if the mean effect size level and the variance of universe correlations would have been larger. These additional results in the Hall and Brannick (2002) study point to the fact that the theoretical analyses of this chapter are not only statistical gimmicks but can have a real impact. For further theoretical and empirical comparisons of approaches with different models and effect size measures, the reader is referred to the pertinent literature (e.g., Overton, 1998; Brockwell & Gordon, 2001).

To summarize, the approaches under examination have many attributes in common as can be recognized from the perspective of the general frameworks of meta-analysis. HOr and RR, for example, are different in a minor detail at best. Nevertheless, important differences between approaches lie in the underlying model (FE vs. RE), the effect size measures used in aggregation (r vs. Fisher-z) and also in the weights employed in aggregation. Previous comparisons of approaches — most of which were based on Monte Carlo study results — show convergence as well as differences in results, where differences can at least partly be attributed to properties of the estimators as outlined in this chapter.