This is chapter 7: Aims, Design, and implementation (pp. 93-114) from

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

# 7

# Aims, Design, and Implementation

In this part of the book, a comprehensive Monte Carlo study for the comparative evaluation of the statistical approaches will be presented. First, the aims and general procedure will be outlined. Procedural details will be given to enable an assessment of the precision of the study and justify the validity of the results to be presented in Chapter 8. Next, the parameters characterizing the universe from which the effect sizes are drawn will be presented and related to the situations of fixed and random effects as outlined in Chapter 4. This will define the scope of interpretation of the results and shed light on viable generalizations of the results. Finally, technical details on the generation of correlation coefficients in Monte Carlo studies in general are discussed and some specifications for software programming to conduct the Monte Carlo study are given.

Monte Carlo studies are designed to investigate the properties of statistical procedures, techniques, or estimators in particular by conducting a specified number of replications of a statistical procedure when an analytical treatment of the problem is not feasible. In a sense, they can be regarded as experiments conducted to study the behavior of statistics of interests subject to the variation of a set of parameters within the framework of a prespecified model. Accordingly, the design of a Monte Carlo study delimits the scope of interpretation of the results (see Skrondal, 2000). If interest lies, for example, in the performance or robustness of a parameter's estimator, it can only be evaluated with respect to the specific other parameters of the model that have been varied or held constant in a Monte Carlo study. Hence, in the following sections the design of the Monte Carlo study conducted to compare the computational approaches of meta-analysis as outlined in the previous chapters will be described in detail.

## 7.1 GENERAL AIMS AND PROCEDURE

The main aim of the Monte Carlo study is to *compare* as well as evaluate the various statistical approaches of meta-analysis as presented in Chapter 5. One of the most important questions to be answered based on the results is whether and when the choice of an approach of meta-analysis makes a difference. In the present Monte Carlo study, the effect sizes under scrutiny will be confined to correlations. Only the $d$-statistic will be of concern insofar as correlations can be transformed to $d$ and the meta-analysis be based on these transformed effect sizes. The correlation coefficient was chosen as an effect size measure to compare the meta-analytical approaches for several reasons. First, it is one of the most often reported effect sizes indices in the empirical literature in the social sciences and psychology in particular. It therefore represents one of the most representative effect size measures in these scientific areas. Second, all the approaches presented in Chapter 5 explicitly propose procedures to aggregate this effect size measure. Third, its various forms can be easily accommodated to express the size of an effect in a wide variety of research situations and also for results from focused hypothesis tests, a fact that lead several researchers to strongly advocate its use (e.g., Rosenthal & DiMatteo, 2001; Rosenthal et al., 2000). The empirical comparison of meta-analytic approaches is thus limited to a research database consisting of correlations.

If present, differences between the results of these approaches will be highlighted and compared to expectations from an analytical point of view. Comparisons of empirical results with the latter type of expectations are also of interest insofar as many of the theoretical results presented and referenced in Part II hold only asymptotically. Thus, it will be investigated whether the application of the proposed procedures yields sufficiently accurate results so that their use is justified under restricted conditions (see Hedges, 1994b). The restriction of conditions predominantly refers to the number of studies to be aggregated and the number of persons in the studies. It is important to recognize and differentiate these two types of asymptotics. On the one hand, holding everything equal, results may be expected to converge asymptotically to some parameter when $n$ grows larger. On the other hand, this might be the case — ceteris paribus — when the number of studies in a meta-analysis grows. It may well be the case that for some estimator of interest, only one of these types is relevant and a growing $n$ or $k$ does not have an effect on the results.

The approaches will be compared with respect to the statistical properties of the proposed procedures for the various meta-analytical tasks. The tasks are estimation, testing, and confidence intervals for the mean effect size, homogeneity tests, and estimating heterogeneity variance. The presentation of the results is structured correspondingly. Special attention will be payed to indices that were developed in individual approaches, like the 75%-rule of the HS-approach), and their usefulness as meta-analytical tools will be assessed in a separate subsection.

Some computational details and specifications on the indices used to make comparisons (e.g., for the mean squared error) will be reported when they are

needed, that is, in the relevant sections in Chapter 8. It will then be pointed out which of the respective indices performs best with regards to conventional statistical criteria, like the mean squared error of the point estimators, for example. A possible and hardly surprising result might be that there is no single approach to meta-analysis performing best under all conditions defined by the design. Instead, a newly assembled collection of procedures from various approaches might emerge as a set of meta-analytical techniques performing best under the examined conditions. If the performance of the indices varies strongly in dependence on parameters varied in the Monte Carlo design, the parameter configurations under which single indices perform best will be highlighted. This can be useful information for future meta-analyses to condition their choice of an index on the specific circumstances (e.g., mean $n$ of the studies to be integrated and number of studies $k$).

The comparison of approaches will be conducted under various parameter configurations that correspond to different models. As mentioned in Section 4.1, the most often applied approaches in meta-analyses on research topics in psychology assume a fixed effects model. This has been severely criticized for various reasons and calls have been made for an increased use of the random effects model (e.g., National Research Council, 1992). Nevertheless, the research practice in meta-analysis has not yet followed this call (for examples, see Hunter & Schmidt, 2000). Hence, a comparative evaluation of the effects of applying the fixed effects procedures of the approaches in heterogeneous cases is of vital interest. This is the case for at least two reasons. First, it will be possible to point out situations in which flaws in the conclusions of such meta-analyses are likely to prevail. Second, it will be possible to assess the tenability of conclusions of such meta-analyses and the potential need for reevaluations. Comparisons of meta-analytical approaches that pursue a similar goal have already been conducted (e.g., Johnson, Mullen, & Salas, 1995), but there are some shortcomings in procedure and design associated with these comparisons that make it reasonable to reinvestigate this topic (see also Section 5.6). Furthermore, most comparisons of procedures referenced in Section 5 have focused on single indices and used different procedures in their simulation studies that complicates and exacerbates the comparison of approaches. The present effort therefore also aims at comparing the approaches within a single simulation framework and to evaluate the approaches *comprehensively* in procedures and design.

## 7.2 GENERAL EXPECTATIONS AND PREDICTIONS FOR THE RESULTS

On the basis of the many properties of estimators and procedures highlighted in Part II, some more specific predictions for the results can be made. These will be highlighted in the following paragraphs.

***Estimation of the Mean Effect Size.*** The bias of the correlation coefficient and Fisher-$z$ transformed correlations is a very well investigated topic in the statistical literature. At least since Hotelling's seminal paper in the 1950s, the biases are well-known and can count as *theoretically* well understood. Nevertheless, there is a plethora of articles investigating the comparative biases of the $r$ and Fisher-$z$ transformed $r$ in simulation studies (e.g., Corey et al., 1998; Donner & Rosner, 1980; Field, 2001; Silver & Dunlap, 1987). Adding yet one more Monte Carlo study to demonstrate the biases seems like flogging a dead horse.

Hence, it is expected on the basis of theoretical results outlined in the previous part that in a *homogeneous* situation ($\mathfrak{S}_1$) the approaches, as categorized by type of effect size measure and using $n$ as weights, show slightly different biases in opposite directions, especially when $n$ is small. Corresponding results will only add to the credibility of the simulation procedure and represent nothing new. Yet, additional estimators are included in this Monte Carlo study which have not been investigated as thoroughly as HO$r$ and HS, for example. It is expected that OP will be unbiased and HOT will show a similar behavior. These predictions are expected to hold across all values of $n$ and $k$ in the Monte Carlo design. OP-FE and OP-RE are expected to show positive biases due to the weighting scheme used (see Section 5.6). In comparison, OP-FE will show a larger bias than OP-RE because incorporation of (estimated) heterogeneity variance will level differences in weights in the latter approach. The size of the biases is not easy to predict and will emerge as a result of the Monte Carlo study. Biases for these two approaches will also diminish when $n$ grows larger because of decreasing variability of observed effect sizes for larger $n$. The bias will stay unchanged across values of $k$, which will be true for all approaches since biases are not expected to vanish or be exacerbated because more (biased) data points are added. Predictions for HO$d$ can hardly be made due to its strange behavior (see Section 5.5). As a consequence, there are also no good reasons to expect HO$d$ to show similar results as any of the other approaches.

In a *heterogeneous* situation (i.e., $\mathfrak{S}_2$ and $\mathfrak{S}_3$) predictions are quite different. Here, Fisher-$z$ based approaches estimate $\mu_{\rho z}$ and not $\mu_\rho$. Only with respect to the parameter which is estimated, approaches are expected to perform well. It should nevertheless be borne in mind that Fisher-$z$ based approaches have a *positive* bias with respect to $\mu_\rho$ the larger the heterogeneity variance. The reasons for this are expounded in Section 5.5. Again, OP is expected to perform uniformly best in the heterogeneous situations because of its UMVU qualities. HOT is now expected to estimate a different parameter in comparison to OP due to its Fisher-$z$ basis. OP-FE and OP-RE are expected to retain their bias in general, but it is again predicted that OP-RE will show smaller bias the larger the heterogeneity variance. This interesting effect is expected because with growing heterogeneity variance the weights will be dominated by the estimates of the heterogeneity variance. Again, it is unclear how HO$d$ will perform.

Overall, OP is expected to perform uniformly best. In some of the situations under investigation other approaches might nevertheless show acceptable performance with the standards of precision in the social sciences in mind.

With respect to another aspect of estimating the mean effect size, namely the estimators' mean squared error, predictions are not easy to make. Of course, there will be a tendency of estimators with large bias to also show large mean squared errors, but it is not necessarily the case that estimators with small bias will perform well with respect to this criterion. These facts notwithstanding, OP is again expected to perform best in relation to all other estimators.

***Significance Tests.*** With respect to tests of the mean effect size, it is important to discriminate between two cases: when $\mu_\rho = 0$ and when it does not. Of course, the null hypothesis need not be $H_0$: $\mu_\rho = 0$, any other value of interest might be inserted instead of 0, but this traditional "nil hypothesis" will be focused on here. For the purpose of testing, there is a large set of candidates included in this Monte Carlo study.

Predictions concerning Type I error rates will first be explicated, that is, the performance of the approaches when the null hypothesis is true will be examined. In a *homogeneous* situation $\mathfrak{S}_1$, all approaches are expected to retain an a priori chosen $\alpha$ level to an acceptable degree, except for cases in which small $n$ is coupled with small $k$ and a disadvantageous weighting scheme is used, as for OP-FE, OP-RE, and HO$d$. This is due to the facts that all testing procedures follow the same basic rationale on the one hand, and deleterious effects of some weighting schemes as already outlined on the other. When the null hypothesis is not true, the power of the approaches' procedures is concerned. Power will be higher for all approaches the larger the effect, that is, the higher the absolute value of $\mu_\rho$. With regards to power it is important to recognize both $n$ and $k$ being relevant for the performance. One of the reasons to apply meta-analysis at all is because of its suspected high power due to aggregating study results (i.e., increasing $k$ and total $n$). This is indeed a valid suspicion as Cohn and Becker (2003) as well as Hedges and Pigott (2001), for example, have demonstrated. However, in these papers it was also demonstrated that power is not always high in meta-analysis. For example, adding studies with small sample sizes may decrease power (Hedges & Pigott, 2001) for RE approaches. Hence, this effect is expected to occur in the results of the simulation study. Otherwise, FE approaches will tend to reject the null hypothesis more often than the more conservative RE approaches (see Hedges & Vevea, 1998, for example). Because OP is expected to be most precise in estimation, this is expected to generally translate to better performance in testing as compared to other FE approaches.

In *heterogeneous* situations the RE approaches are expected to perform better overall as compared to FE approaches, because the basic model assumption (heterogeneity) is correct and the approaches account for this in their procedures. This will also generally lead to more conservative decisions in comparison to FE approaches. Hence, higher power of FE approaches is expected to come at the cost of excessive Type I error rates. More specifically, the null hypothesis will always be false in $\mathfrak{S}_2$ in the Monte Carlo study due to the design which does not include negative universe parameters. In essence, basically the same results as in $\mathfrak{S}_1$ are expected to emerge with acceptable performance of

most approaches in most situations, except for some combinations of the design variables (low $n$ coupled with high $k$). In $\mathfrak{S}_3$, RE approaches are expected to perform best when the null hypothesis is true (i.e., $\mu_\rho = 0$) because FE approaches do not incorporate heterogeneity variance in the standard error. More specifically, this prediction applies to DSL and OP-RE. However, DSL will retain the prescribed $\alpha$-level best because it is suspected that the weighting scheme of OP-RE will impair its performance when $n$ is small. With reference to Osburn and Callender (1992), it was pointed out in the context of presenting the HS approach that using the variants HS3 and HS4 may result in good performance in heterogeneous situations (see also Whitener, 1990). Hence, these two approaches are also expected to perform well. If the null hypothesis is false, then the power can again be examined. In these cases DSL is again predicted to show more conservative behavior in comparison to FE approaches. However, it is again predicted that the higher precision of OP will have a beneficial effect on its performance, though at the cost of an excessive Type I error.

***Confidence Intervals.*** In evaluating confidence intervals of the approaches two aspects have to be accounted for: coverage rates and interval widths. Coverage rates refer to the proportion of intervals covering the universe parameter in a series of replications. High coverage rates may come at the cost of large intervals, so they are not unequivocal indicators of the quality of the procedures. Thus, such rates have to be qualified by simultaneously considering the interval widths.

The most important property of estimators to attain high coverage rates — disregarding interval widths — is low bias. Since OP is expected to show the smallest bias in all situations, it is again predicted to show the best performance. This is anticipated to be true in all situations $\mathfrak{S}_1$, $\mathfrak{S}_2$, and $\mathfrak{S}_3$. RE approaches will show high coverage rates in all situations but these approaches will also have the largest interval widths notwithstanding which situation is examined. This is caused by incorporating estimates of heterogeneity variance in standard errors, which will almost always be positive, even in the homogeneous situation. Coverage rates are expected to become better for all approaches for larger $n$ and $k$ as conventional statistical results would suggest.

***Homogeneity Tests.*** Two different types of homogeneity tests were introduced in Chapter 5, those based on the $Q$-statistic and the 75% or 90% rule, respectively. First, focus will be on the $Q$-statistic, which is available to conduct a test in the approaches HO$r$, HO$d$, HS, and OP-FE. For the predictions it is important to recall one of the most important assumptions of this test, namely the normal distribution of the deviates. These are squared, weighted and summed over $k$ studies to arrive at the $Q$-statistic. Under this assumption and if the null hypothesis is true, the $Q$-statistic has an asymptotic central $\chi^2_{k-1}$-distribution, that is, when study sample sizes are (very) large.

As a consequence, it can be predicted that HO$r$ will perform best in comparison to the other approaches. The basis for this prediction is the reasonableness of the assumption of normally distributed deviates for Fisher-$z$ transformed

correlations. In contrast, it is not wise to assume a normal distribution either in the case of $r$-based approaches (HS and OP-FE) or for HO$d$. At least when $\rho$ is moderate to large and/or sample sizes are not huge, the assumption is not tenable. With regards to HO$d$, the assumption might be sensible if $d$ was not a transformed $r$ as in the present case. In addition to the distributional assumption, the deleterious effects of the weighting scheme are expected to operate again for HO$d$ and especially OP-FE. In sum, HO$r$ is expected to perform best amongst the approaches under examination. However, on the basis of theoretical analyses (see Hedges & Pigott, 2001) and previous evidence from attempts to evaluate this test (e.g., Harwell, 1997; Sánchez-Meca & Marín-Martínez, 1997), it can be expected that at least for some combinations of the design variables Type II errors will occur. More specifically, it is predicted that particularly for situations of small $n$ and large $k$ — which operates to exacerbate the small $n$ problem — low power of the homogeneity test based on the $Q$-statistic will be observed.

The 75% and 90% rules as homogeneity tests are not expected to represent viable alternatives to the homogeneity test mentioned in the above paragraph. Due to their crude rationale and previous evaluations of these rules (for an overview of results, see Cornwell & Ladd, 1993), a high Type I error rate and relatively low power for combinations of moderate to low $n$ and $k$ is to be expected.

***Estimation of Heterogeneity Variance.*** A total of three estimators for the heterogeneity variance is available in the Monte Carlo study: DSL, HS, and OP-RE. Of these, DSL is Fisher-$z$ based and therefore in $z$-space and not directly comparable to the other two estimators based on $r$. Unfortunately, there is no transformation formula available to date to make these three estimators directly comparable. Nevertheless, some predictions for their comparative performance are possible.

In the *homogeneous* situation $\mathfrak{S}_1$, all estimators are expected to show positive but small bias. This is due to their construction which prescribes negative estimates to be set to zero. If either non-truncated (i.e., negative estimates are not set to zero) or truncated estimators are compared, then DSL and HS are predicted to perform better than OP-RE. Despite the fact that OP-RE is unbiased (see, e.g., Hedges, 1989), the deleterious effects of the weighting scheme are expected to hamper good performance, at least when $n$ is small. DSL is not expected to suffer from any weighting scheme problem and is unbiased by construction, though the weighting scheme might cause problems with the estimator when other effect sizes are used as in the present case (see Böhning et al., 2002). Hence, the DSL estimator is expected to perform best.

Basically the same predictions for relative performance of the approaches can be made for *heterogeneous* situations. DSL is expected to perform best, HS might show negative bias (see Hall & Brannick, 2002), OP-RE will perform worst.

## 7.3 DISTRIBUTIONS IN THE UNIVERSE OF STUDIES

In line with the differentiation between fixed and random effects models drawn in Chapter 4 and with regard to the importance of assumptions about the distribution of the universe effect sizes to be modeled, the situations introduced in Section 4.5 will be distinguished in the Monte Carlo study. The choice of the situations is mainly oriented on the assumptions that are ordinarily made in published meta-analyses about the universe of studies.

As a consequence, $\mathfrak{S}_1$ is an important situation to include in the design, mostly because of its prevalence in the literature. It represents the homogeneous case where only a single universe effect size is assumed to be estimated by the studies under investigation. Additionally, $\mathfrak{S}_1$ is also included in the design of the Monte Carlo study to test whether the FE methods work properly when their basic assumptions are met and also to explore how RE methods perform in homogeneous situations.

One of the heterogeneous situations included in the design is $\mathfrak{S}_2$. Here, two different values $\rho_1$ and $\rho_2$ are present in the universe of studies with equal probability of occurrence. The difference between these values is not yet specified, but is a design aspect. The probabilities of .50 associated with the two values of $\rho$ are the weights of the components in mixture distribution parlance and will *not* vary as part of the design. The Monte Carlo study thus also investigates the performance of the approaches in $\mathfrak{S}_2$ and compares estimates of mean effect sizes, for example, with the expected value of the mixing distribution they are intended to estimate. It should finally be kept in mind that the design of the Monte Carlo study will be limited to situations in which the component weights will also always be equal. Of course, this restriction precludes reliable generalizations to situations in which there are more than two classes and where components weights are very different.

The second type of heterogeneous situation included in the Monte Carlo study is $\mathfrak{S}_3$. Here, a univariate continuous distribution is given in the universe of studies which is the beta distribution in the present study. The normal distribution was not used because it is not bounded by the interval $[-1, 1]$ and to avoid discarding invalidly large values. An alternative procedure was realized by Overton (1998), for example, who randomly set invalid values from the normal distribution to values between .90 and .9999 according to a uniform distribution extending over this range. This is certainly an unsatisfactory state of affairs because the density of the normal distribution is distorted by such trimming of values and a determination of its actual parameters and properties is thus impeded. Hence, such procedures are not considered as satisfactory.

In contrast to an earlier work that also used the beta distribution (Hedges, 1989), using parameters for the beta distribution that yield U-shaped or rectangular distributions was refrained from because they did not appear as plausible for the distribution of effect sizes in the universe. The same is true for distributional forms of a J-shape. Discussion of this issue will be resumed in the following section when the specific parameters values for the Monte Carlo design will be introduced.

The question remains how the particular parameters of the beta distribution were calculated in the Monte Carlo study. To elucidate the procedure, consider the following probability density function of the beta distribution with parameters $p$ and $q$

$$p_X(x) = \frac{1}{\mathrm{B}(p,q)} x^{p-1}(1-x)^{q-1}, \quad 0 \leqslant x \leqslant 1,$$

given in standard form (see Johnson, Kotz, & Balakrishnan, 1995). As is evident, this standard form is bounded by the interval $[0, 1]$. Next, the aim is to find values of the parameters $p$ and $q$ that correspond to desired expected values and variances in terms of $\rho$. For example, how should $p$ and $q$ be chosen to yield a beta distribution with an expected value of $\mu_\rho = .10$ and $\sigma_\rho^2 = .15$? To find an expression for the computation of the parameters it is important to note that a random variable $X$ following a beta distribution still continues to be beta-distributed when linearly transformed. Accordingly, the transformation $\mathrm{P} = 2X - 1$ is applied to a standard beta-distributed random variable $X$ to yield a distribution on the interval $[-1, 1]$. Furthermore, the moments of this transformed variable are (see Johnson, Kotz, & Balakrishnan, 1995, p. 219)

$$E(\mathrm{P}) = \frac{2p}{p+q} - 1,$$

and

$$\mathrm{Var}(\mathrm{P}) = \frac{4pq}{(p+q)^2(p+q+1)}.$$

Equating $E(\mathrm{P})$ and $\mathrm{Var}(\mathrm{P})$ with $\mu_\rho$ and $\sigma_\rho^2$, respectively, and solving simultaneously for $p$ and $q$ leads to the following equations

$$\begin{aligned} p &= \frac{1 + \mu_\rho - \mu_\rho^2 - \mu_\rho^3 - \sigma_\rho^2 - \mu_\rho\sigma_\rho^2}{2\sigma_\rho^2} = -\frac{(1+\mu_\rho)(-1+\mu_\rho^2+\sigma_\rho^2)}{2\sigma_\rho^2} \\ q &= \frac{1 - \mu_\rho - \mu_\rho^2 + \mu_\rho^3 - \sigma_\rho^2 + \mu_\rho\sigma_\rho^2}{2\sigma_\rho^2} = \frac{(-1+\mu_\rho)(-1+\mu_\rho^2+\sigma_\rho^2)}{2\sigma_\rho^2} \end{aligned}$$

Now $\mu_\rho$ and $\sigma_\rho^2$ correspond to the desired values for the expected value and variance in terms of the beta-distributed variate on the interval $[-1, 1]$. Applying these equations to the example given above ($\mu_\rho = .10$ and $\sigma_\rho^2 = .15$) yields $p = 23.65$ and $q = 19.35$, respectively. In the Monte Carlo study these equations were applied to compute the parameters of the beta-distribution for a whole set of combinations of $\mu_\rho$ and $\sigma_\rho^2$. The resulting values are reported in Tables A.1 and A.2 in the appendix. The type of continuous distribution of the random variable P in the universe of studies is now specified and characterizes $\mathfrak{S}_3$ of the Monte Carlo study.

In sum, a total of three situations $\mathfrak{S}_1$ to $\mathfrak{S}_3$ is given of which the first represents a homogeneous case and the second and third are heterogeneous cases. The first two situations $\mathfrak{S}_1$ and $\mathfrak{S}_2$ are characterized by discrete distributions

whereas the third is continuous. Of course, one could easily imagine a host of further situations: for example, situations with more than two groups and a discrete distribution, a variety of different component weights in the discrete situations, different parametric continuous distributions in the universe and maybe even mixtures of continuous distributions at the universe level. Thus, although this Monte Carlo study can count as one of the most comprehensive in design on the present research topic up to date, it is necessarily limited. These limits of the design should be borne in mind when examining the results. Having described the types of situations under investigation, the parameter values that were chosen for the various variables of the design will now be specified.

## 7.4 PARAMETERS

The variables of the design to evaluate the approaches of meta-analysis are

- the values of $\mu_\rho$ for all situations $\mathfrak{S}_1$ to $\mathfrak{S}_3$,
- the variance of the beta distribution ($\sigma^2_\rho$) in $\mathfrak{S}_3$,
- the number of studies $k$ to be aggregated in a meta-analysis, and
- the number of persons $n$ to compute the effect sizes in the individual studies.

The values for $\mu_\rho$ represent one single universe effect size common to all $k$ studies in $\mathfrak{S}_1$ and the expected value of the beta distribution in $\mathfrak{S}_3$. For $\mathfrak{S}_2$, two different values $\rho_1$ and $\rho_2$ were chosen. Of course, there is also a corresponding $\mu_\rho$ in $\mathfrak{S}_2$, however, it is ambiguous for the specification of $\rho_1$ and $\rho_2$.

As specified in Section 4.5, the weights for the components in $\mathfrak{S}_2$ were held constant. Additionally, the number of persons for each effect size is considered to be invariant within each simulated meta-analysis. That is, if in $\mathfrak{S}_1$ there is one universe parameter $\mu_\rho$ underlying a number of $k = 32$ studies, for example, then the effect sizes $r_i$ of all 32 studies have some fixed number of persons. Although not representative of published meta-analyses, $n$ is held constant mainly to exclude any interaction effects of $n$ with other aspects of the design. An interaction effect of $n$ and $\mu_\rho$, for example, is indeed a very interesting research topic on its own. The well known publication bias in meta-analysis (Begg, 1994; Rosenthal, 1979) can be regarded as such an interaction and continues to stimulate research efforts to assess and eliminate such influences on the results of a meta-analysis (e.g., Hedges & Vevea, 1996; Iyengar & Greenhouse, 1988; Rust, Lehmann, & Farley, 1990; Schwarzer, Antes, & Schumacher, 2003; Vevea & Hedges, 1995). Thus, the reasons to hold $n$ constant are to ensure exclusion of such interaction effects from the results and to keep the focus on the effects of the design variables as implemented.

The specific values that are used for the design variables listed above are presented in Table 7.1. The first row of Table 7.1 shows that the values for $\mu_\rho$ are positive in all simulated cases.

**Table 7.1 Parameter Values in the Simulation Procedure**

| Parameter | Values | Number of Values |
|---|---|---|
| $\mu_\rho$ | 0; 0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9 | 10 |
| $\sigma_\rho^2$ ($\sigma_\rho$) | 0.0025 (0.05); 0.01 (0.10); 0.0225 (0.15); 0.04 (0.20); 0.0625 (0.25) | 5 |
| $k$ | 4; 8; 16; 32; 64; 128; 256 | 7 |
| $n$ | 8; 16; 32; 64; 128; 256 | 6 |

*Note.* $\mu_\rho$ = expected value of correlation in the universe, $\sigma_\rho^2$ = variance of correlations in the universe, $k$ = number of studies per meta-analysis, $n$ = number of persons per study used to compute the observed correlations ($r_i$).

Since the distributions in the interval below zero would mirror those simulated on the positive side, only the given set of values is of concern. Note that the values provided for $\mu_\rho$ also represent the range of values chosen for $\rho_1$ and $\rho_2$ in $\mathfrak{S}_2$. The second row shows the values for the variances and the respective standard deviations in parenthesis for the beta distributions in situation three. The given values are considered to cover the range of plausible variances for the mixing distribution. The third and fourth row show the range of values for the number of studies in a meta-analysis and the number of persons per study, respectively. The values were chosen to yield a higher resolution for small values but also to extend to relatively large values. This was achieved by calculating powers of two beginning with $2^2$ for $k$ up to $2^8$.

The reader might wonder whether the values used in the Monte Carlo study are representative of published meta-analyses. Unfortunately, investigations of the characteristics of the distributions of the design variables are quite rare, so the main resource to judge adequacy of the values is research experience. At least, there are content analyses of meta-analytic studies of correlations in I/O psychology available (Cornwell, 1988; see also Lent, Aurbach, & Levin, 1971). The results of Cornwell's study on 81 meta-analyses published in seven major journals of I/O psychology provides descriptive statistics for the distributions of $n$ and $k$, respectively. Since there are extreme values for both variables (Maximum $n = 45,222$ and Maximum $k = 2,162$), the author provided the statistics for a truncated distribution for both variables[1]. The mean value for $n$ was 283 (Median = 102; Mode = 73) and a value of 37 (Median = 12; Mode = 6) resulted for $k$. Hence, the choice of including a value of $k = 4$ in the present Monte Carlo study seems warranted. An additional argument in favor of the inclusion of such small values is the fact that subgrouping of studies corresponding to the levels of assumed explanatory variables very often leads to very small $k$ in the subgroups (see, e.g., Farrell & Hakstian, 2001). The largest value for $k$

[1]The distributions were truncated at $n \leq 500$ and $k \leq 120$.

included in the design does not occur very often but there are other research areas, like attitude–behavior research, for which a very large number of research articles is available (see, e.g., Eckes & Six, 1994; for an overview, see Schulze & Wittmann, 2003). The values chosen for $n$ in the present design seem to cover the range of values occurring in practice although very small values are not very often reported in the content analysis (however, Minimum $n = 7$). But again, there are research areas for which very small $n$ is customary as Hunter and Schmidt (1994b), for example, have pointed out. The variances chosen in the present design also match those used by Cornwell and Ladd (1993) and those used in other Monte Carlo studies in the field.

In sum, the levels chosen for the design variables seem to cover customary values of research practice, at least in the I/O psychology area. Nevertheless, the criterion of realism should not be overvalued when considering the levels of the design, since research practice might change and in some fields of study totally different characteristics may prevail. Moreover, against the background of the aim to study the properties of statistics that draw on asymptotic statistical theory, it is important to include also low values of the design variables. The inclusion of values for $\mu_\rho$ that can be judged as very high in comparison to estimates observed in meta-analyses in any field of psychology is also intended not only to mirror research practice but to study the behavior of procedures also under extreme conditions. However, it is not the case that such high values do not occur in I/O psychology (see Hite, 1987) or social psychology (see Schulze & Wittmann, 2003), for example.

To gain an overview of the large number of design variable combinations under study, it is instructive to review their number. In $\mathfrak{S}_1$, only one of the ten universe effect sizes is given and the variance is zero in all cases. These 10 values are combined with all of the $k$ and $n$ values resulting in a total of $10 \times 7 \times 6 = 420$ meta-analyses to be simulated.

The second situation $\mathfrak{S}_2$ differs from the first in that two different values for $\rho$ are given, resulting in a number of non-redundant combinations of $\frac{(10\times 9)}{2} = 45$. The differences between the values range from .10 to a maximum of .90. Again, these 45 universe value combinations are combined with all of the $k$ and $n$ values leading to a total of $45 \times 7 \times 6 = 1890$ meta-analyses.

Finally, the values for $\mu_\rho$ of the beta distribution were combined with the variances of row two in Table 7.1. The full combination of all values unfortunately lead to J-shaped beta distributions in extreme cases. This is illustrated in Figures A.1 to A.5 in the appendix and can be seen by inspecting the parameter values $p$ and especially $q$ in Table A.2 in the appendix. For values of $p$ or $q$ less than one, the beta distribution turns into a J-shape (Johnson, Kotz, & Balakrishnan, 1995). The utilization of such distribution types would lead to sampling values from the beta distribution that are predominantly very large and close to one, with a few extremely low values to attain the prescribed values for $\mu_\rho$ and $\sigma^2_\rho$. Two reasons lead to the omission of such distributions from the Monte Carlo design. First, the described problem only applies to very high values of $\mu_\rho$ in combination with high values of $\sigma^2_\rho$ that are very unlikely to

arise in practice. Second, the utilization of these distributions would presumably have an enormous impact on the results of the study, in particular on the tests of homogeneity and estimates of heterogeneity variance. They would lead to biased assessment of the overall performance of estimators due to extreme values that emerge from J-shaped distributions. The following combinations were omitted: $\mu_\rho = .90$ with variances $\sigma^2_\rho = .0625, .04, .01, .0225$ and $\mu_\rho = .80$ with variances $\sigma^2_\rho = .0625, .04$. The omission of the six combinations of $\mu_\rho$ and $\sigma^2_\rho$ resulted in $10 \times 5 - 6 = 44$ combinations and thus a total of $44 \times 7 \times 6 = 1848$ meta-analyses in $\mathfrak{S}_3$.

The sum of the meta-analyses of all situations amounts to a total of 4158. For all these combinations, the statistics and tests of the several approaches to meta-analysis are computed to facilitate a comparative evaluation of the approaches in a wide range of possible situations given by these combinations. For all of the 4158 combinations, the computations were repeated in 10,000 iterations. Correspondingly, the results to be presented in Chapter 8 are either the means of certain statistics computed over all iterations or statistics derived from the iterations, like the standard deviation of the estimators for the mean effects over 10,000 values. The number of iterations in the present study can be considered to be relatively large in comparison to other Monte Carlo studies in the context of meta-analysis. Most previous Monte Carlo studies have chosen 1000 iterations (e.g., Cornwell, 1993; Law, 1995; Sackett et al., 1986; Sánchez-Meca & Marín-Martínez, 1998a; Spector & Levine, 1987) or 5000 iterations (e.g., Harwell, 1997; Sánchez-Meca & Marín-Martínez, 1998b), only a few studies have used 10,000 iterations (e.g., Alexander et al., 1989; Silver & Dunlap, 1987) and rarely were more iterations used (100,000 by Field, 2001). The number of iterations chosen here is therefore regarded as sufficient.

## 7.5 DRAWING RANDOM CORRELATION COEFFICIENTS

A final important technical aspect of the simulation study will now be discussed in considerable detail because it is one of the most important steps in the Monte Carlo study. Up to this point it has been laid out which variables and values are chosen for the design of the Monte Carlo study. The next task is to generate random correlation coefficients $r_i$ that conform to these prescriptions. For convenience, assume correlations coefficients have to be generated for a single $\rho$, that is, $\mathfrak{S}_1$ is of concern and $\rho = \rho_1 = \ldots = \rho_k$. Note that the problem to be described is fully equivalent for all other situations and the results of this sections do not only pertain to $\mathfrak{S}_1$.

There are $k$ independent studies with a common sample size $n$. The observed correlation coefficients $r_i$ provided by the studies are assumed to be based on pairs $(x_1, y_1), \ldots, (x_n, y_n)$ of two variates $X$ and $Y$ having a joint bivariate normal distribution. Drawing random correlation coefficients means that we want to generate a set of $k$ values of $r_i$ for a given $\rho$. The first procedure that comes to mind is to generate $n$ pairs for the two variates $X$ and $Y$ and compute the sample correlation correlation coefficient. That is, one would use

the following equations to generate the values for $x$ and $y$ on an observational level:

$$\begin{aligned} x &= v \times \sqrt{\rho} + e_1 \times \sqrt{1-\rho} \\ y &= v \times \sqrt{\rho} + e_2 \times \sqrt{1-\rho} \end{aligned} \tag{7.1}$$

where $v$, $e_1$, and $e_2$ are realizations of corresponding variates that follow a standard normal distribution and are mutually independent. This procedure has often been used in Monte Carlo studies to generate correlation coefficients (e.g. Corey et al., 1998) and it is easy to show that $X$ and $Y$ have correlation $\rho$ when generated by this procedure.

Unfortunately, this procedure is computationally rather intensive and takes up a great amount of computation time in a large simulation study. This is the case because to generate a single correlation $r_i$ one has to draw $n \times 3$ times for $v$, $e_1$, and $e_2$ from a standard normal distribution and correlate the resulting values of $x$ and $y$ subsequently. To speed up the whole process of generating correlation coefficients it would be much more efficient to directly draw correlations from the sampling distribution of the correlation coefficient or approximations thereof without generating pairs of values for $X$ and $Y$. Several candidates for a more efficient approach using this strategy are considered now. The reader not interested in technical details may skip the following section and directly go to Section 7.6 without loss of understanding for the subsequent chapter.

### 7.5.1 Approximations to the Sampling Distribution of *r*

Alternatives to the computationally intensive procedure are given by using a series of analytical results on the distribution of the correlation coefficient. First, it is well known that when $\rho = 0$

$$\frac{r\sqrt{\mathrm{df}}}{\sqrt{1-r^2}} \sim t_{\mathrm{df}}, \tag{7.2}$$

where df are the degrees of freedom ($\mathrm{df} = n - 2$) in Equation 7.2. Accordingly, one could draw values from a central $t$ distribution with df degrees of freedom and use the transformation

$$r = \sqrt{\frac{t^2}{t^2 + \mathrm{df}}} \tag{7.3}$$

that results from solving Equation 7.2 for $r$ to simulate a series of $r_i$ values. This procedure would fulfill the need for a more efficient strategy but the question arises how $r$ values can be generated in cases where $\rho \neq 0$.

Ideally, one would draw directly from the sampling distribution of such correlation coefficients but their distribution is unfortunately mathematically rather complex (see Section 3.1). Since it cannot be given in closed form, its usage as a distribution to draw correlations from is obstructed. Nevertheless, it will be considered as a benchmark to judge the quality of the approximations

to the distribution of $r$ we will turn to in the following paragraphs. The PDF of $r$ was already given in Equation 3.1 on page 21.

Three approaches that rely on different distributional approximations to the PDF as given by Equation 3.1 are considered and discussed in some detail in the following paragraphs. A comparison and evaluation of these approaches with respect to the proximity to the PDF of $r$ will follow their presentation.

***The Fisher Approximation.*** A first method to generate $r$ values with a specific $\rho$ in the underlying population is to randomly draw values from a normal distribution $\mathcal{N}(\mu_Z, \sigma_Z^2)$ with the following parameters

$$\mu_Z = \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right) = \tanh^{-1} \rho \tag{7.4}$$

and

$$\sigma_Z^2 = \frac{1}{n-3}. \tag{7.5}$$

In the next step, the resulting values of $z$ are transformed to sample correlation coefficients $r$ via

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

Although refinements of these formulae have been proposed in intensive investigations of the mathematical properties of the distribution function of $r$ (Hotelling, 1953; Ruben, 1966), the approximation by using Equations 7.4 and 7.5 is a very popular one (see Chapter 3.1) and might therefore be considered as a possible and natural procedure for a simulation study.

***The Harley Approximation.*** A second approach is based on the analyses reported by Harley (1957) that dealt with an approximation of the noncentral $t$ distribution by the distribution of a transformed correlation coefficient. She showed that in a population with a given $\rho$ of

$$\rho = \sqrt{\left( \frac{2\tau^2}{2n - 3 + \tau^2} \right)}, \tag{7.6}$$

the function

$$\sqrt{\frac{2\,\mathrm{df}(1-\rho^2)}{2-\rho^2}} \frac{r}{\sqrt{1-r^2}} \sim t_{\mathrm{df},\tau} \tag{7.7}$$

is distributed as noncentral $t$ with noncentrality parameter $\tau$. Using equation 7.6 and solving for $\tau$ leads to

$$\tau = \sqrt{\frac{(2\,\mathrm{df} + 1)\,\rho^2}{2-\rho^2}}.$$

This result can be used in simulation studies to compute $\tau$ of the noncentral $t$ distribution to randomly draw values from. From a rearrangement of Equation 7.7, the resulting values of $t$ can then be transformed back to $r$ using the following equation

$$r = \sqrt{\frac{t^2}{t^2 + \frac{(2\,\mathrm{df}+1)\,\mathrm{df}}{2\,\mathrm{df}+1+\tau^2}}},$$

which is given here in a form remarkably similar to Equation 7.3.

***The Samiuddin-Kraemer Approximation.*** The last approximation to be considered here is based on the work by Samiuddin (1970) that was later refined and extended by Kraemer (1973, 1975; Kraemer & Paik, 1979). Due to the elaborations mainly presented by Kraemer it will be labeled *Kraemer approximation* in what follows. The approximation also draws on the $t$ distribution.

It was shown that

$$\frac{(r-\rho')\sqrt{\mathrm{df}}}{\sqrt{(1-r^2)(1-\rho'^2)}} \sim t_{\mathrm{df}} \tag{7.8}$$

has a central $t$ distribution with df degrees of freedom. In Equation 7.8, $\rho'$ is a function of $\rho$ that has to satisfy a series of requirements not repeated here (see Kraemer, 1973). Although Kraemer (1973) proposed that the median of the distribution of $r$ is a good approximation to $\rho'$ and better than $\rho$ at least for small sample sizes while Mi (1990) was able to show that if $\rho$ is taken as $\rho'$ the distributional result stated above holds. Accordingly, the possibility for a simulation study established by this approximation is to draw $t$ values from a central $t$ distribution with df degrees of freedom and convert the resulting $t$ values to $r$. The conversion can be done by solving Equation 7.8 for $r$, which leads to

$$r = \frac{(n-2)\,\rho - \left(\rho^2-1\right) t\sqrt{n-2+t^2}}{n-2-(\rho^2-1)\,t^2}.$$

Of course, it is claimed that all the approximations presented here are satisfactory as compared to the sampling distribution of $r$. The results reported by the referenced authors seem to support this claim. It is therefore reasonable to consider these approximations when a Monte Carlo study is conducted, in which a large amount of $r$ values has to be randomly generated, as is the case for the present study. It should finally be noted that none of the authors of the approximations advocated their use for the purpose of conducting Monte Carlo studies. The utilization of the approximations can consequently be regarded as an innovative aspect of their usefulness, but their utility has to be scrutinized beforehand. We will now turn to an evaluation of the presented approximations for this purpose.

### 7.5.2 Evaluation of the Approximations

Among the most important questions within the framework of an evaluation are the procedures of evaluation and provision of according criteria. The approximations presented in the previous subsection will be evaluated by determining the distribution and distributional properties of the $r$s they produce. As a first criterion, a visual inspection of the probability density function of the $r$ values of the approximations in comparison to the exact density given in Equation 3.1 will be carried out. Additionally, the expected value and variance of the distributions will be compared as a second set of criteria, again with the exact density as a standard for comparisons. The approximations will be considered as satisfactory if the distributions of the $r$ values generated by the procedures very closely match those of the values as given by the exact distribution. To accomplish this type of evaluation, the distributions in question were determined by numerical methods using MATHEMATICA. For details on the specific procedures applied, the interested reader is referred to Section B of the appendix where an annotated MATHEMATICA notebook can be found. It can be used to understand the genesis of the results reported here and to reproduce and possibly extend them.

The general logic underlying the computations is as follows. The characteristic feature of all the approximations is that there is a transformation $T$ of the correlation coefficient $R$, denoted as $T \circ R$, the distribution of which ($\mathcal{P}_{T\circ R}$) can be approximated by a member of a well-known family of distributions. That is, the values $T \circ R$ are conceived as if they were generated by a variate $X$ with a probability distribution $\mathcal{P}_X$ belonging to that family. Yet in other words, the distributions $\mathcal{P}_{T\circ R}$ and $\mathcal{P}_X$ — or equivalently the random variables $T \circ R$ and $X$ — are equated. For example, in the Fisher approximation the $r$s are transformed to $z$ values that have an approximate normal distribution as described above. In the proposed procedures to generate $r$ values the first step is to draw values from the hypothesized probability distribution of $X$ and convert the resulting values back to $r$ subsequently by applying the transformation formulae presented in the previous subsection. To be clear, it is thereby assumed that $\mathcal{P}_X$ is not only the asymptotic but the exact distribution of $T \circ R$. In the case of the Fisher approximation $\mathcal{P}_X$ is the normal distribution, whereas for Harley's and the Kraemer-Samiuddin approximation it is the central and noncentral $t$ distribution, respectively. The question to be answered by the evaluation is whether the distribution of the $r$s, which are generated by the outlined procedure, does indeed show the same properties as $R$ expected by exact theory.

All transformations $T$ are strictly increasing so that the inverse function exists. For an example, consider again the Fisher approximation where it is the inverse Fisher transformation $\tanh z$. That is, for all transformations there is $T(R) = X$, and we also have $T^{-1}(X) = R$. Under these conditions the aim can be restated as to determine the probability density of $R$ when the density $p_{T\circ R}(x)$ is given and the inverse transformation is applied. To achieve this aim, we first consider

$$\mathcal{P}(R \leq r) = \mathcal{P}(T^{-1}(X) \leq r) = \mathcal{P}(T \circ R \leq T(r)).$$

Again, this will be illustrated for the Fisher approximation by

$$\mathcal{P}(R \leq r) = \mathcal{P}(\tanh(Z) \leq r) = \mathcal{P}(Z \leq z).$$

The probability density distributions can therefore be computed as

$$\mathcal{P}(R \leq r) = \mathcal{P}(T \circ R \leq T(r)) = \int_{-\infty}^{T(r)} p_{T\circ R}(x)dx = \int_{-1}^{r} p_{T\circ R}\left(T(y)\right) T'(y)dy,$$

where the critical step is a change of variables[2] in the equation above. Yet again, this can be illustrated as an example for the Fisher approximation by

$$\mathcal{P}(R \leq r) = \int_{-\infty}^{z} \varphi(x)dx = \int_{-1}^{r} \varphi\left(\tanh^{-1}(y)\right) \tanh^{-1'}(y)dy$$

where $\varphi(x)$ is the normal distribution with expected value $\zeta$ and variance $(n-3)^{-1}$. The change of variables is extraordinarily useful for the present purpose of inspecting a distribution of a random variable ($R$) when it is subject to a transformation, since the result of this procedure is of utmost importance for the evaluation of such transformations. Applying this procedure to the present set of transformations allows a comparative inspection of the distribution of $R$ for different cases. The following Figures 7.1 and 7.2 depict examples of the density distributions for the transformations of interest in the present context. In Figure 7.1 the densities are plotted for $n = 32$ and $\rho = .30$. As can be seen, the densities are virtually indistinguishable by inspection in this case. All approximations coincide with the exact density of $R$ and can therefore count as very satisfactory. In Figure 7.2 a more extreme case is depicted, also for $n = 32$ but $\rho = .90$. Note that this case is of interest for the present study as it is part of the design. Here, the curves do not all coincide. The density for the Harley approximation is obviously most "off" from the others. The Fisher and Kraemer approximation are virtually identical but do not perfectly match the exact density of $R$. Nevertheless, from inspection of the figures they may still count as very good approximations to the exact density. The point to be noted is that graphical comparisons between the densities of the approximations in comparison to the standard distribution allow some of the approximations to appear as quite satisfactory. Of course, up to now only two special cases with a fixed $n$ and two different $\rho$ values were chosen for comparison and for other constellations of the parameters the approximations may be even better or worse

[2]As a reminder, a change of variables is given for two continuous functions $f$ and $g$, where $f$ is continuous and $g$ is continuously differentiable with derivative $g'$, by

$$\int_{g(a)}^{g(b)} f = \int_{a}^{b} (f \circ g) \cdot g'$$

$$\int_{g(a)}^{g(b)} f(u)du = \int_{a}^{b} f\left(g(x)\right) \cdot g'(x)dx.$$

The two equivalent forms for the change of variables are given here for ease of comparison with the equations given in the text (see e.g., Spivak, 1967).

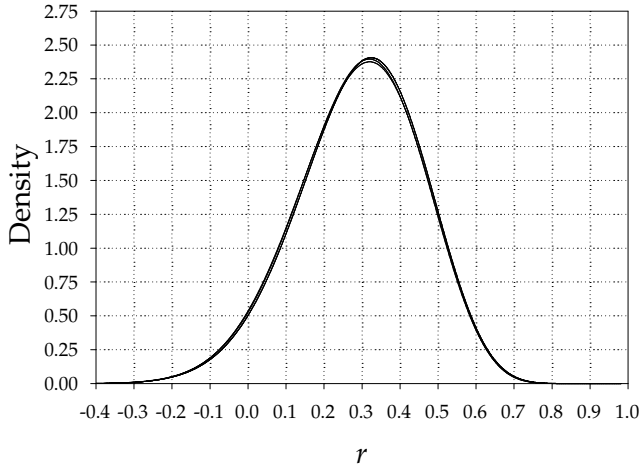


**Figure 7.1** Densities for $R$ of the approximations, $n = 32$, $\rho = .30$.

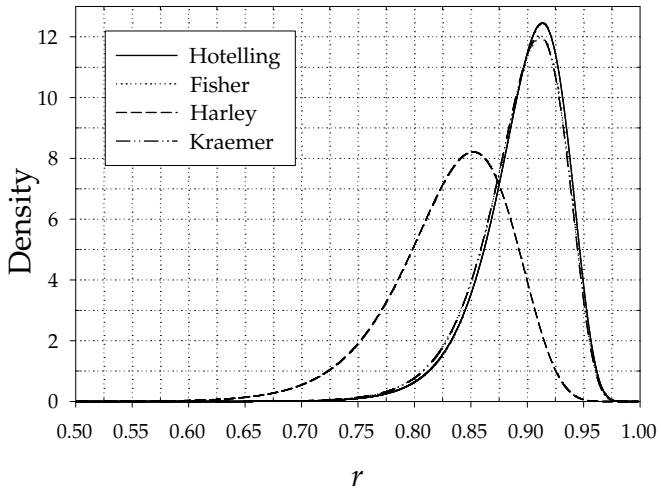


**Figure 7.2** Densities for $R$ of the approximations, $n = 32$, $\rho = .90$.

than indicated in the figures. The parameter constellation was actually deliberately chosen to illustrate a general trend of the value of the approximations. First, all approximations become worse the higher the $\rho$ that is chosen. Second, all aproximations are almost perfect in the region about $\rho = 0$. Third, a point not illustrated by the figures, all approximations perform better the higher the value for $n$ that is is chosen, but they are still visually distinguishable from the exact density for $\rho \gtrapprox .90$ when $n$ is not extremely large. In sum, the approximations perform very well for some constellations of the parameters but not for all. As will become evident by the following inspection of the numerical properties, that is, the expected values and variances of the distribution, the visual inspection of such graphs can be quite deceptive insofar as good-looking approximations may nevertheless not achieve satisfactory values for distributional properties.

**Table 7.2 Expected Values and Variances for the Approximations and the Exact Density of $R$, $n = 8$ and $n = 128$, $\rho = 0, \ldots, .90$**

| | | Hotelling | | Fisher | | Harley | | Kraemer | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $\rho$ | $\mu$ | $\sigma^2$ | $\mu$ | $\sigma^2$ | $\mu$ | $\sigma^2$ | $\mu$ | $\sigma^2$ |
| 8 | .00 | .000 | .1429 | .000 | .1472 | .000 | .1429 | .000 | .1429 |
| | .10 | .093 | .1409 | .085 | .1455 | .093 | .1409 | .086 | .1412 |
| | .20 | .187 | .1349 | .171 | .1403 | .186 | .1349 | .172 | .1363 |
| | .30 | .281 | .1250 | .259 | .1317 | .280 | .1251 | .260 | .1281 |
| | .40 | .376 | .1115 | .348 | .1196 | .374 | .1116 | .349 | .1166 |
| | .50 | .472 | .0945 | .440 | .1040 | .469 | .0947 | .441 | .1017 |
| | .60 | .571 | .0747 | .536 | .0851 | .563 | .0753 | .537 | .0835 |
| | .70 | .672 | .0529 | .637 | .0629 | .654 | .0547 | .638 | .0623 |
| | .80 | .776 | .0305 | .745 | .0386 | .741 | .0350 | .746 | .0388 |
| | .90 | .885 | .0105 | .864 | .0145 | .820 | .0187 | .864 | .0151 |
| 128 | .00 | .000 | .0079 | .000 | .0079 | .000 | .0079 | .000 | .0079 |
| | .10 | .100 | .0077 | .099 | .0077 | .100 | .0077 | .099 | .0077 |
| | .20 | .199 | .0073 | .198 | .0073 | .199 | .0073 | .198 | .0073 |
| | .30 | .299 | .0065 | .298 | .0066 | .299 | .0065 | .298 | .0066 |
| | .40 | .399 | .0056 | .397 | .0056 | .397 | .0056 | .397 | .0056 |
| | .50 | .499 | .0045 | .497 | .0045 | .495 | .0045 | .497 | .0045 |
| | .60 | .598 | .0033 | .597 | .0033 | .589 | .0033 | .597 | .0033 |
| | .70 | .699 | .0021 | .697 | .0021 | .679 | .0022 | .697 | .0021 |
| | .80 | .799 | .0010 | .798 | .0011 | .761 | .0013 | .798 | .0011 |
| | .90 | .899 | .0003 | .899 | .0003 | .834 | .0006 | .899 | .0003 |

*Note.* The columns labeled "Hotelling" correspond to the exact density whereas the other columns are labeled in accordance with the approximations introduced in the previous Subsection 7.5.1.

Table 7.2 presents a series of expected values and variances of the approximations and the exact density for comparison. Similar to the visual inspection of the figure, only a small subset of possible combinations of $n$ and $\rho$ is chosen for comparison and presented in Table 7.2, but these values suffice to illustrate the general points to be highlighted[3]. First, the approximations by Fisher and Kraemer seem to fare equally well in comparison to Harley's, which generally leads to inferior values for higher $\rho$ in terms of differences to the Hotelling standard. Second, all approximations get worse for higher $\rho$ and better for larger $n$. This is what is to be expected from statistical theory, because the distribution of $R$ is central $t$ at zero and all the approximations are expected to be almost perfect in this region. Furthermore, the distributional properties of the approximations hold only asymptotically, so they get better with growing

[3]With the notebook presented in Appendix B it easy to compute any desired values to extend the comparisons.

numbers of $n$. Lastly, none of the approximations seems to provide satisfactorily similar expected values and variances to the standard, except in the case of $\rho$ in the region of zero combined with high $n$. Although the reported differences may appear quite small in value, they are actually too large for the purpose of generating correlation coefficients by these procedures. To understand this judgment, focus on the Fisher approximation as an example. This approximation leads to generally smaller expected values in comparison to the exact density and somewhat larger variances. This means that a simulation study in which $z$ values are drawn from a normal distribution, these are converted to $r$ by the inverse Fisher transformation, and estimates of the mean effect size are based on these $r$ values, may possibly report flawed conclusions for an assessment of the bias. The reasons for a potential flaw lie in the difference between the expected value of the Fisher approximation and the exact distribution. The expected value for a situation of $n = 64$ and $\rho = .50$ is $\mu = 0.49701$ for the exact density and $\mu = 0.49398$ for the Fisher approximation. Now suppose a comparison between the biases of $r$ and Fisher-$z$ is of interest in this situation. For the exact density, the biases to be anticipated by statistical theory[4] are $\text{Bias}_r = -0.002943384$ for $r$ and $\text{Bias}_z = 0.003987731$ (given in the space of $r$) (see Hotelling, 1953, p. 212 for $r$ and p. 216 for $z$), the well known negative bias for $r$ and positive bias for $z$. But the biases are not to be anticipated when the expected value of the probability distribution is shifted downwards by the simulation procedure as is the case with the Fisher approximation. This downward shift will have the effect that the overestimation of Fisher-$z$ will not be as large as expected by theory and the negative bias of $r$ will emerge as larger in absolute value than it effectively is when assessed in relation to the exact density. In short, the positive bias of $z$ is compensated by using the Fisher approximation in the simulation procedure and incorrect conclusions with respect to biases may result.

It is therefore concluded that the considered candidates for a simulation procedure cannot be used because they produce distortions of the sampling distributions for statistics. This happens to an extent that is of relevance for Monte Carlo studies. Hence, none of the candidates will be used and the computationally more costly procedure introduced at the beginning of this section in the Equations 7.1 will be used instead. The results of the evaluation have relevance not only for a decision in this Monte Carlo study, but also for a reevaluation of previous ones. For example, Spector and Levine (1987) have employed the Fisher approximation to generate $r$ values in a Monte Carlo study on the susceptibility of the HS-procedure to Type I and II error rates. In light of the results presented here, at least some doubt is cast on the results and conclusion of the Monte Carlo study by Spector and Levine and others who have used the approximations.

[4]Note that the bias of $r$ given here does not add up with the expected value to $\rho$ exactly. This occurs because the approximation by Hotelling (1953, p. 212) is only given up to the third term of an expansion.

## 7.6 DETAILS OF PROGRAMMING

A computer program for MS-DOS was designed and programmed in Borland C++ Version 5.02. The procedure to generate the correlation coefficients followed Equations 7.1. Since a very large amount of numbers had to be randomly drawn for these correlations, a random number generator with a very long period length was of interest. Remarkably, standard random number generators in common use appear to have serious deficiencies (Hellekalek, 1998). According to the review of random number generators by Hellekalek (1998), the Mersenne Twister (TT800) (Matsumoto & Nishimura, 1998) was the only random number generator with a flawless performance,[5] and was therefore implemented in the program.

To speed up the draws from the standard normal distributions, an array of two million values was filled from which values were drawn. The array was randomly refilled 8 times in the course of the whole computations.

## 7.7 SUMMARY

The current Monte Carlo study is designed for a comparative evaluation of the approaches to meta-analysis in common use in the social sciences and the procedures they propose as valuable tools for meta-analysis. To achieve this aim, the design of the Monte Carlo study includes a wide range of different values for the universe correlations $\rho$ in the universe of studies (from .00 to .90 in increments of .10), the number of studies to be aggregated (from 4 to 256), and number of persons in the studies to be aggregated (from 8 to 256). Additionally, different situations are implemented in the design that correspond to the assumptions of the fixed and random effects models in meta-analysis. The situations were distinguished by the form of the distribution of universe effect sizes that were classified as discrete and continuous. For the discrete distributions homogeneous and heterogeneous situations are included where the heterogeneous situations have two distinct values. In the case of a continuous distribution, six different variances of a beta-distribution were additionally varied.

As a result, the whole procedure can be thought of as a two-stage sampling process. In the first step, universe values are drawn from a distribution in the universe of studies with prespecified properties as described above and in the second step, observed values are drawn from distributions with properties that depend on the universe values drawn in step one. For the second step, different forms to generate the observed correlation coefficients were considered and the possibility to draw correlations directly from approximate distributions was rejected as unsatisfactory.

[5]The source code of the Mersenne Twister TT800, as well as reviews of the quality of random number generators can be found at http://random.mat.sbg.ac.at