This is chapter 8: Results (pp. 115-180) from

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

8 Results

The following sections provide an overview of the results for the Monte Carlo study of meta-analytic approaches. First, a brief introduction to the presentation style will be given. This seems necessary because of the complex structure and multitude of results. The intention is to make the presented results more easily comprehensible and to point out how a maximum of information can be gathered from the graphics found in the subsequent sections. The presentation of results diverges from the structure of Chapter 5 in that the focus is kept on the questions to be answered by the statistical analyses. First, Section 8.2 is devoted to questions pertaining to the estimation of the effect size in the universe of studies, for example, issues regarding the bias and relative efficiency of the proposed estimators. Next, the results on the accuracy of homogeneity tests will be reported in Section 8.5. Finally, estimators of the heterogeneity variance — which are important in random effects approaches — are examined in Section 8.6. The sequence of sections thus resembles the conduct of a meta-analysis, while not exactly mirroring it. The situations $\mathfrak{S}_1$ to $\mathfrak{S}_3$ will be separated in all sections to assess the statistics' performance under different conditions.

8.1 PRELIMINARIES

One of the characteristic features of the present study is the wealth of situations, design variables, and number of different approaches to be compared. Most of the results are subject to levels of the dimensions k, n, μ_ρ, or differences of ρ_1 and ρ_2 in $\mathfrak{S}_2$. Additionally, results are compared for levels of variances (σ_ρ^2) in $\mathfrak{S}_3$. The number of dimensions obviously precludes any simple pictorial or tabular presentation. As a consequence, the report of the results needs to be brought into an easily comprehensible form.

The results presented in the text are always selected to represent and illustrate the primary aspects of the respective results. Mostly, results will be collapsed over at least one design variable. More specifically, collapsed means that the *mean* over all levels of such a variable will be computed. The resulting mean values will then be presented for the levels of all other variables in the design. For example, in the presentation of results for the biases of estimators over several levels of n, the mean values computed across levels of k will be presented. The absolute values of biases are then easily interpretable, if biases do not (greatly) vary over levels of k. A more complicated picture emerges in cases where results differ across *all* levels of *all* design variables. This will be highlighted in the presentation and should be borne in mind when inspecting collapsed results. Nevertheless, even in these more complicated cases, a *comparison* of the approaches is still possible.

In general, much more emphasis will be placed on graphical rather than tabular presentation of the results to facilitate illustration of trends and relationships which often go unrecognized in tables. The figures will prevalently be three-dimensional graphs since they often give a better impression of interactions of the design variables and are also very compact ways of representing a wealth of results and general trends. All three-dimensional graphs will display smoothed data or surfaces using negative exponential smoothing. This is a local smoothing technique using polynomial regression with weights.[1] In short, the weights are chosen in this technique so that the influence of points decreases exponentially with the horizontal distance from certain points of the surface.

The following graphs illustrate the effect of smoothing and how graphs produced by this technique can be interpreted. The upper left and right panel in Figure 8.1 depict a three-dimensional scatterplot of the bias of a statistic for varying μ_ρ and n. Both upper panels show the same results, each from a different angle of view. The lower panels depict the same graphs with smoothed surfaces added resulting from negative exponential smoothing. As can easily be seen, the lower set of graphs gives a much clearer and more easy to grasp picture of the relationships between the variables depicted. These types of graphs also supersede series of two-dimensional line graphs as ordinarily presented in the literature, where it is left to the observer to synthesize the graphs cognitively. When inspecting the graphs it is important to recognize — as can be verified in Figure 8.1 — that the intersections of the meshes on the surfaces correspond to the *data points* plotted. The mesh intersections are *not* to be interpreted as projections of the plots' grid intersections onto the surface. Data point dots will therefore be omitted in the graphs of the results sections. Since the data points are not equally spaced with respect to a linear scale on the design variables n and k, the mesh of the surface will also be more tightly interconnected in some areas when the results are plotted by n and k. Additionally, when different shadings occur on the surfaces they can be interpreted

[1] For the graphical presentation in this chapter, SigmaPlot for Windows Version 8.02 and its smoothing facilities were used to prepare the figures.

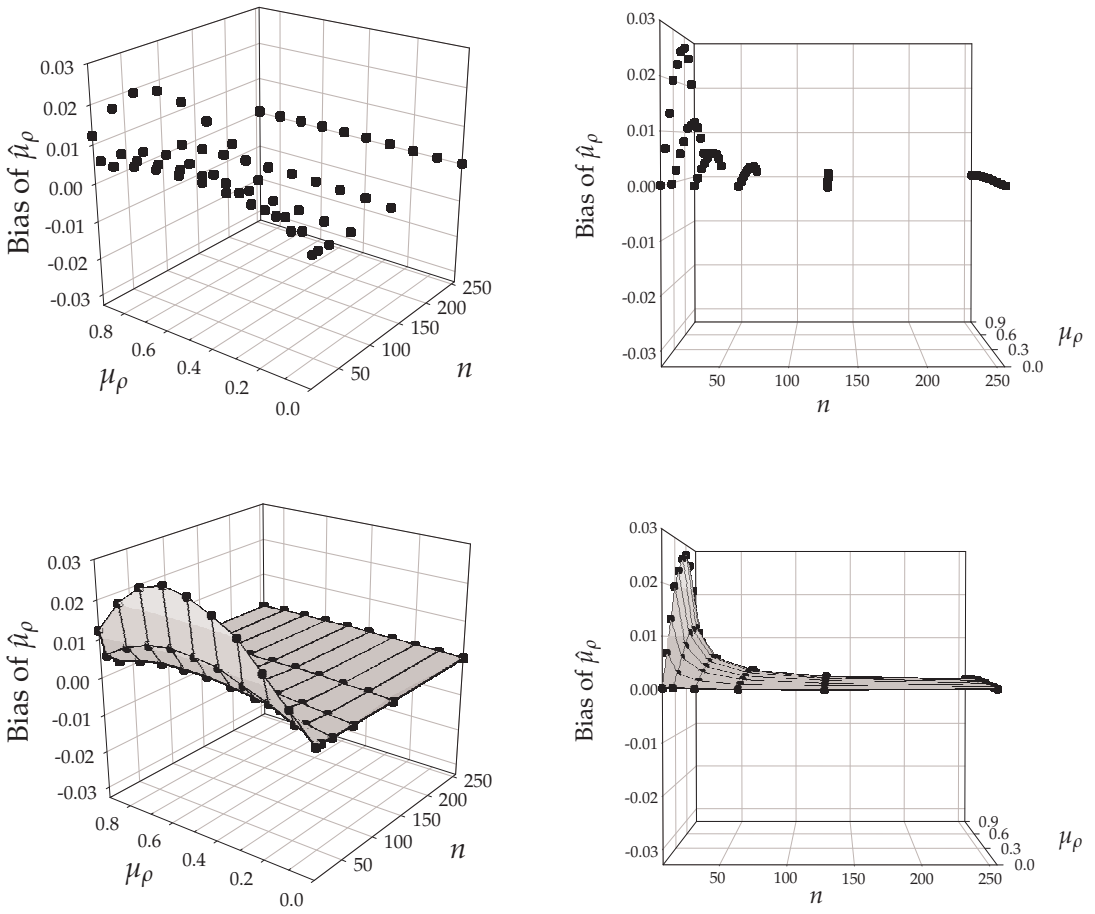

Figure 8.1 Illustration of smoothing in graphical presentations.

as contours with respect to the vertical axis. In the example graphs this is the axis labeled "Bias of $\hat{\mu}_\rho$". This enables the reader to see the height of values on the surfaces even in the middle of a three-dimensional graph. At least rough estimates of the actual values plotted can thereby be gathered from the graphs.

Unfortunately, the virtues of a concise graphical presentation of the results are accompanied by a loss in numerical precision in the report. More precise results not readily read from the figures are provided by the author to the interested reader upon request.[2]

8.2 ESTIMATION OF THE MEAN EFFECT SIZE IN THE UNIVERSE OF STUDIES

At the core of most meta-analyses is the estimation of a mean effect size. Two connotations are usually associated with this phrase. First, a summary of the available effect size data is intended to be given by the meta-analyst that is a

[2]Email: rs@psy.uni-muenster.de

good representation of the data at hand. The weighted mean of the observed effect sizes usually gives such a good summary in a least-squares sense. Second, the phrase also alludes to estimating a parameter of the distribution of effect sizes in the universe of studies. The parameter supposed to be of most concern to meta-analysts is μ_ρ, the expected value of the universe distribution in the space of r. It is this latter sense that will be of concern in the following subsections. The main question to be answered is how well the different r-based estimators of the various approaches are in estimating μ_ρ. Recall from Section 5.5 that z-based approaches do not estimate μ_ρ but $\mu_{\rho z}$. This issue will be elaborated when it is of most concern, namely when presenting the results for heterogeneous situations.

All approaches outlined in Chapter 5 provide procedures that yield estimates either for μ_ρ or $\mu_{\rho z}$. Results on the bias of these estimators and their accuracy are given first, followed by results for the proposed significance tests of the approaches. The two subsections will thus provide an evaluation of the estimators with respect to estimation and inference.

8.2.1 Bias

The bias of an estimator is one important aspect of its statistical quality (see Stuart et al., 1999). The biases were computed for the following presentation so that positive biases indicate estimators for which the mean exceeds the parameter to be estimated. As in most previous studies on the bias of some of the estimators under investigation, the biases will first be examined in a homogeneous situation.

8.2.1.1 Homogeneous Situation $\mathfrak{S}_1$ In $\mathfrak{S}_1$ we have the simple situation of only one effect size in the universe of studies that is estimated by all k studies. Hence, μ_ρ and $\mu_{\rho z}$ are equal and estimators of approaches using r versus its Fisher-z transform need not be differentiated here. The bias of approaches that apply the Fisher-z transformation was computed for the mean effect size transformed into r-space, whereas for approaches that do not apply this transformation the estimators were used directly. For convenience, the value in the universe to be estimated is denoted by μ_ρ in all situations. This notation is used also in describing the results in $\mathfrak{S}_1$ for reasons of consistency. Of course, μ_ρ is a constant ρ in $\mathfrak{S}_1$, and the reader should not be confused by this notation.

The following graphs show the biases of all approaches by the design variables k and n. As mentioned in the introduction of this chapter, statistics have to be combined across levels of other design dimensions (i.e., μ_ρ in the present case) to facilitate the presentation of results. To create the graphs depicted in Figure 8.2, the mean bias of the estimators over the omitted dimension μ_ρ was computed and the data points in the figure represent these mean biases. As will become evident from the subsequent presentation, biases vary substantially over levels of μ_ρ, so that it should be borne in mind for interpretation of the depicted values that the graphs represent aggregates over the omitted dimension.

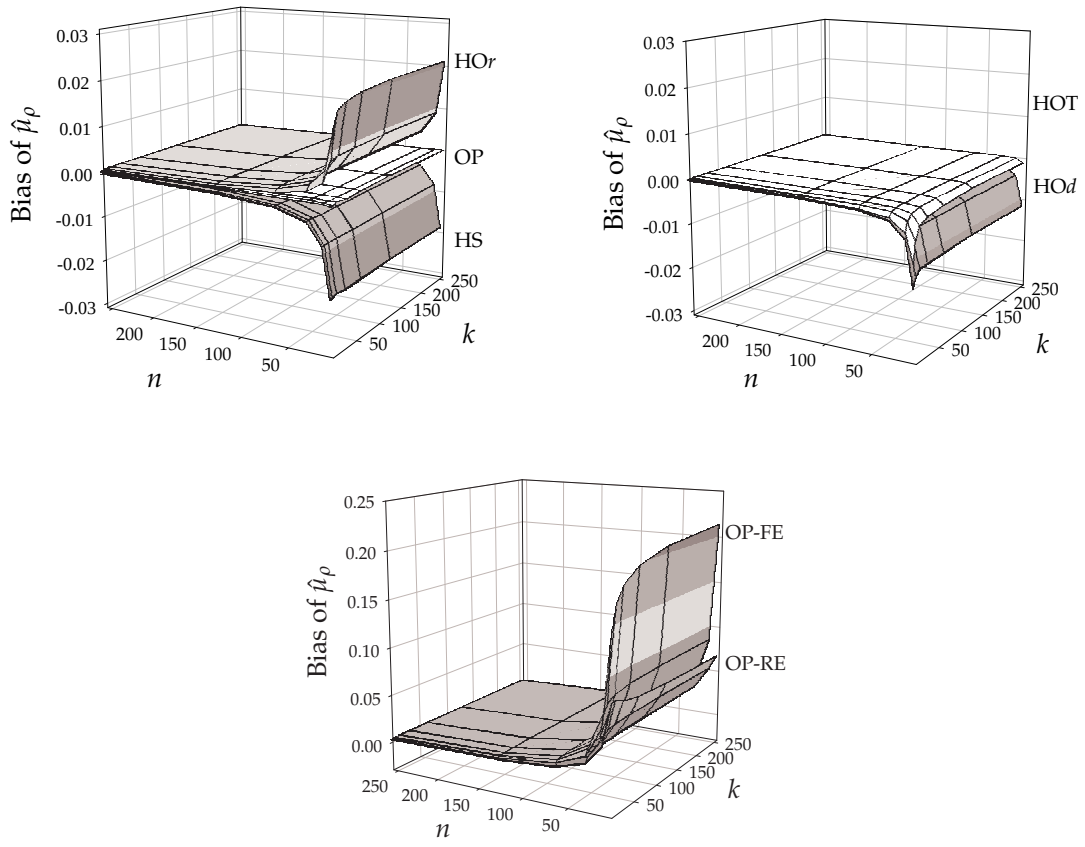

Figure 8.2 Bias of μ_ρ estimators in $\mathfrak{S}_1$ by k and n.

Here and in the following graphs, the three panels show the results of all approaches. The arrangement of approaches is not oriented on theoretical concerns but for clearer representation of the results. The reader may wonder why the results for DSL and RR are omitted. This is due to the fact that the results for both approaches are identical to the results for HOr as far as bias and mean squared errors are concerned. For a recapitulation of the reasons for these identities the reader is referred to Sections 5.2 and 5.4.1. The results for RR are generally omitted from the presentation in the text — except for the subsection on significance testing — because the results for RR and HOr are indistinguishable for theoretical reasons.

The bias of all estimators strongly depends on the sample size whereas biases show practically no variability with respect to k. The strongest change in biases occurs from very small $n = 8$ to approximately $n = 64$. For values larger than 64 the biases for all approaches vanish, as one would expect from consistency of the estimators. Estimators of approaches that use the Fisher-z transformation without corrections (HOr) generally show a positive bias and estimators simply based on r (HS) always show a negative bias. This is to be expected from the theoretical analyses reported in Section 3.1.

Two estimators can be identified in Figure 8.2 that show outstanding performance in biases. The correction of r by Olkin and Pratt (OP) as well as the

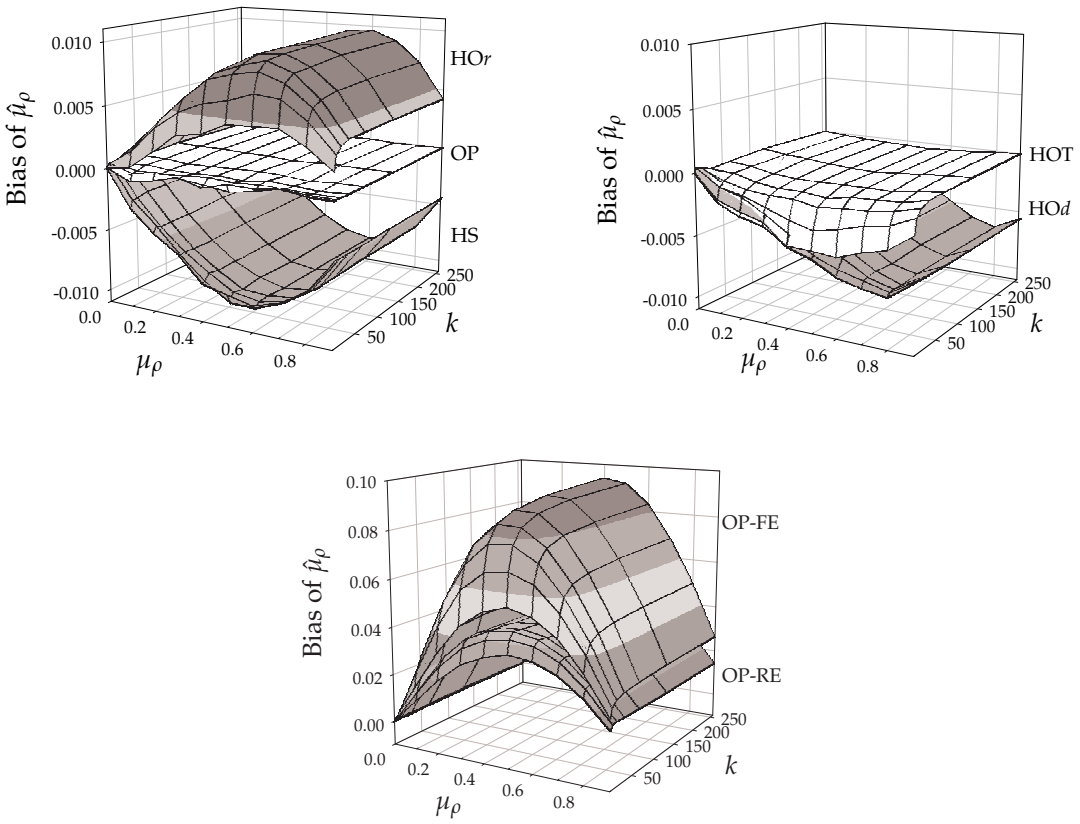

Figure 8.3 Bias of μ_ρ estimators in $\mathfrak{S}_1$ by μ_ρ and k.

correction of z proposed by Hotelling (HOT) show nearly flat planes at a value of zero bias, though OP seems to be slightly better for very small n and k.

In contrast to these extraordinarily good estimators, OP-RE and especially OP-FE stand out with a very poor performance. Whereas both upper panels in Figure 8.2 have a similar scaling on the vertical axis, the scaling of the lower panel had to be strongly extended to show the surfaces for these latter estimators. The surfaces for OP-FE and OP-RE depicted in the figure clearly show the inadequate performance of these estimators of μ_ρ when n is small. The proposed reason for these poor results of the estimators is the weighting scheme they apply. As already mentioned, accidently high values of r receive a very high weight in comparison to lower values and thereby they exert a strong influence on the overall estimate, leading to the high positive bias. OP-RE performs better in $\mathfrak{S}_1$ than OP-FE because it incorporates estimates of heterogeneity variance in its weights that are equal for all aggregated effect sizes. Since these estimates are most frequently non-zero even though the universe variance is zero in $\mathfrak{S}_1$, the deleterious effect of the weights for the biases of OP-FE is somewhat levelled out in OP-RE. The performance of these approaches was expected to be impaired when n is small, however, the magnitude of bias seems surprising. The poor performance of both estimators can also be seen in the graphs shown in Figures 8.3 and 8.4.

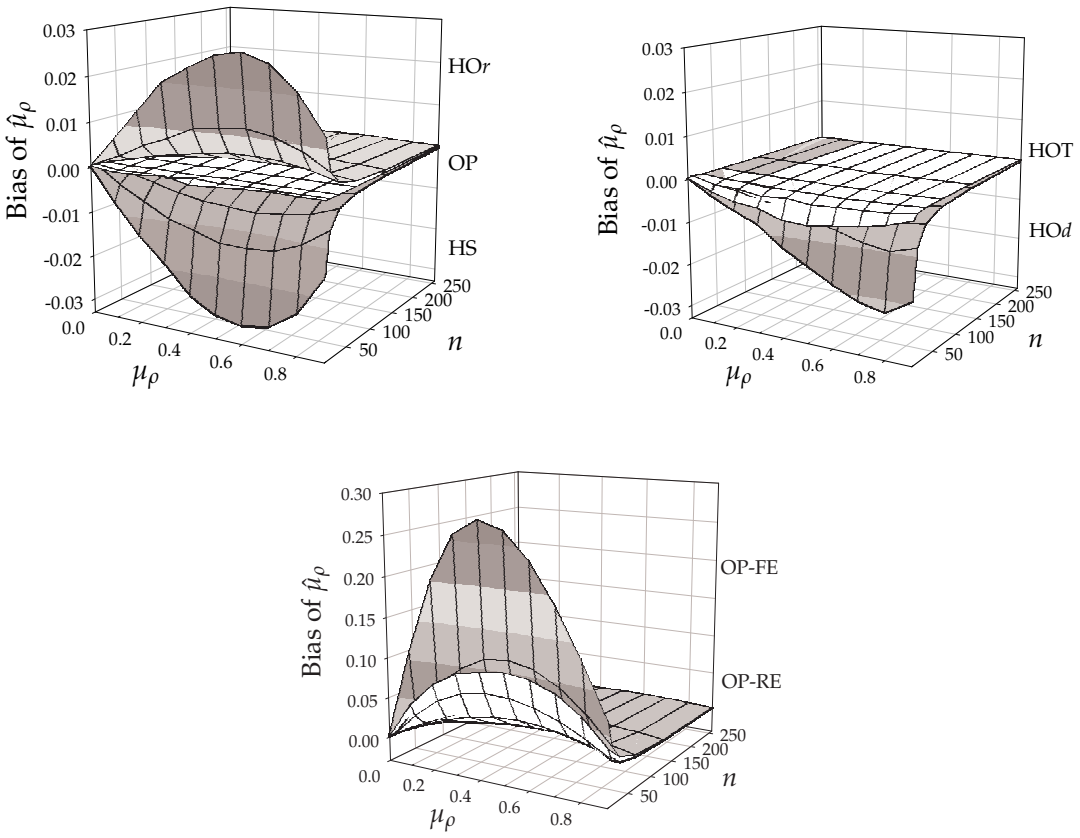

Figure 8.4 Bias of μ_ρ estimators in $\mathfrak{S}_1$ by μ_ρ and n.

Figure 8.3 shows the biases of the estimators for varying μ_ρ and number of studies. Biases are shown not to strongly vary across values of k, only for small values of k below approximately 16 studies do biases show *smaller* values in comparison to higher values of k. This somewhat surprising finding was also reported in a comparison of r and the Fisher-z transformation by Corey et al. (1998) and is not expected from theoretical examinations given in Section 3.1.

The arrangement of estimators in all panels of Figure 8.3 is the same as before and shows the same direction of bias for all estimators. Again, OP and HOT appear as flat planes in the graphs with OP showing slightly better performance for very small values of k. The curvature of the graphs across values of μ_ρ is representative of the general behavior of the estimators. The largest values of bias occur in the region about $\mu_\rho \approx .60$. Scaling of the vertical axis has again to be extended for OP-FE and OP-RE to show the very high values of bias for these estimators.

This has also to be done for the graphs depicted in Figure 8.4, where biases are shown across values of μ_ρ and n. Because the biases do not show substantial variability across values of k and Figure 8.4 shows aggregates over this design dimension, it can be regarded as the best representation of the results on biases in $\mathfrak{S}_1$.

Table 8.1 Descriptive Statistics for the Bias of μ_ρ Estimators in $\mathfrak{S}_1$

	Statistic				
Approach	Max.	Mean	Median	Min.	SD
HO*r*	.0324	.0053	.0018	−.0008	.0075
HOT	.0008	−.0011	−.0002	−.0174	.0023
HO*d*	.0009	−.0036	−.0011	−.0240	.0056
HS	.0008	−.0058	−.0025	−.0301	.0077
OP	.0026	.0000	.0000	−.0033	.0004
OP-FE	.3432	.0514	.0184	−.0005	.0740
OP-RE	.1180	.0280	.0146	−.0010	.0303

Note. The total number of values described by these statistics is 420.

The direction of the estimators' biases and their absolute values closely match the results depicted in Figure 8.2. It becomes evident in Figure 8.4 that in contrast to the estimators based on r or its Fisher-z transform, HOd shows its maximum bias not in the region about .50 but at higher values around .80. However, this slight departure from the behavior of the other estimators does not seem to be of great importance for an overall evaluation of the estimator. Nevertheless, it is remarkable that the application of the transformation from r to d and the meta-analytical aggregation of the resulting effect sizes retains the negative bias of r that becomes a positive bias through the application of the Fisher-z transformation.

OP and HOT again appear as the best estimators in terms of bias and can be designated as the best estimators over the design dimensions n, k, and μ_ρ after inspection of the graphs presented up to this point. The proposed refinements of the estimators in common use show a very satisfying behavior at all levels of the design variables.

The graphs presented here also point to two types of convergence. First, biases converge to zero with larger n, as would be expected. The second type is convergence for larger values of k. Biases do not converge to a value of zero for larger k but instead converge to the bias expected from statistical theory. This is important insofar as it makes clear that adding more studies to a meta-analysis does not lead to vanishing biases in the pooled estimator.

The absolute values of the reported biases may seem very small in magnitude. In fact, most descriptive statistics presented in Table 8.1 show relatively small values of bias for all estimators, except OP-FE and OP-RE. The absolute mean values seem to be of trivial magnitude and not of relevance for interpreting meta-analytical results at all.

If the sole purpose of a meta-analysis would be the estimation of μ_ρ in a homogeneous situation this may indeed be regarded as a valid summary statement for the results presented here. Correspondingly, it has been stated that the cases are very rare in which a correction of bias is worthwhile (Hunter & Schmidt, 1990, p. 71). Although the results also indicate that bias can be of

substantial magnitude when n is very small, such values of n are rarely encountered in practice.

Nevertheless, when evaluating the results one should also take into consideration the importance of the estimates for other analytical steps in a meta-analysis. They play a prominent role, for example, in the computation of the Q-statistic. Although seemingly of inconsequential magnitude, a small bias transfers to and may add up in other statistical analyses based on these estimators. Apart from the small biases of most approaches, the observed biases for OP-FE and OP-RE are of such magnitude that it does not seem sensible to use them as estimators when n is less than approximately 60.

8.2.1.2 Heterogeneous Situation $\mathfrak{S}_2$ The next situation for which performance of the estimators will be evaluated is $\mathfrak{S}_2$. A two-point distribution of effect sizes is given in the universe of studies in $\mathfrak{S}_2$. In analogy to the previous section, mean biases will be computed for several combinations of the design variables.

For better comprehension of the results presented, a reconsideration of the estimated universe parameters seems necessary. In Section 5.5 it was shown that the estimated parameters are different in $\mathfrak{S}_2$ for estimators based on r, Fisher-z transformed r, and d (as resulting from a conversion of r). Recall, however, that in the case of HOd the weights have an effect making it more sensible to use μ_ρ as a universe parameter for comparison. Hence, in the following presentation of results the bias of HOd was not computed with respect to $\mu_{\rho d}$ as the general logic outlined for Fisher-z based approaches would suggest but with respect to μ_ρ. This seemingly inconsistent procedure was applied due to the fact that the values to be presented for HOd are actually much closer to μ_ρ than $\mu_{\rho d}$.

For the approaches using the Fisher-z transformed correlation coefficient, the universe parameters $\mu_{\rho z}$ are higher as compared to μ_ρ. This is illustrated in Figure 5.1 on page 78. To give a more precise impression of this difference consider Table 8.2.

The first two columns in this table provide combinations of the two different parameters in the universe of studies. In the third column the corresponding μ_ρ is given, in the forth column $\mu_{\rho z}$, and the difference between these two parameters can be seen in the fifth column. These differences are actually part of the values depicted in Figure 5.1. It is important to realize that the values in the fifth column are theoretically derived and *not* estimated. As an interpretation of these differences in the context of estimation, one can think of them as providing the biases for Fisher-z based approaches *if* they were *un*biased with respect to $\mu_{\rho z}$ but evaluated with respect to μ_ρ. Hence, it would come as no surprise to observe a "bias" of $-.10$ for the HOr mean effect size estimator in the case as specified in the penultimate row of Table 8.2, for example. Note that this "bias" would be observed (only) if the HOr estimator was indeed *un*biased (with respect to the parameter $\mu_{\rho z}$ it in fact estimates)!

Table 8.2 Comparison of Values of μ_ρ and $\mu_{\rho z}$ in $\mathfrak{S}_2$

ρ_1	ρ_2	μ_ρ	$\mu_{\rho z}$	$\mu_\rho - \mu_{\rho z}$	Est. Bias	$\mu_\rho - (\mu_{\rho z} + \text{Est. Bias})$
.00	.10	.05	.0501	−.0001	.0016	−.0017
.00	.20	.10	.1010	−.0010	.0034	−.0044
.00	.30	.15	.1535	−.0035	.0051	−.0086
.00	.40	.20	.2087	−.0087	.0063	−.0150
.00	.50	.25	.2679	−.0179	.0080	−.0259
.00	.60	.30	.3333	−.0333	.0093	−.0426
.00	.70	.35	.4084	−.0584	.0105	−.0688
.00	.80	.40	.5000	−.1000	.0093	−.1093
.00	.90	.45	.6268	−.1768	.0080	−.1848

Note. The estimated bias is for the HO*r* approach (Est. Bias) and was taken from the results for $k = 16$ and $n = 16$.

As a consequence, the biases for Fisher-*z* based approaches reported in this section are evaluated with respect to $\mu_{\rho z}$. To facilitate comparisons of these biases with others reported for *r*- and *d*-based estimators, they are given in *r*-space, that is, the inverse Fisher-*z* transformation is applied. As an illustration, column six in Table 8.2 provides estimated biases for HO*r* from the Monte Carlo study results for the case of $k = 16$ and $n = 16$. As can be seen, the values are small and round off to approximately −.01 in most cases. Thus, it can be concluded that HO*r* has a small bias with respect to $\mu_{\rho z}$ in $\mathfrak{S}_2$. If interest lies in biases with respect to μ_ρ, they can easily be estimated as well by computing values as given in column seven. By inspecting these values it becomes clear that — at least in this case — the biases of HO*r* with respect to μ_ρ are predominantly composed of the theoretically derived values in column five and the estimated biases in column six only account for a small part.

Amongst the available estimators only those based on *r* provide estimates of μ_ρ in $\mathfrak{S}_2$. This is quite an important theoretical result for the estimation of a mean effect size with correlational data in a heterogeneous situation of the given type. Biases for these approaches are not transformed and will be given as they result in the Monte Carlo study.

In sum, when inspecting the following results, the reader should bear in mind that Fisher-*z* based approaches are evaluated with respect to a different universe parameter as the other approaches. In addition, since μ_ρ was used as the standard of comparison for HO*d*, but μ_ρ can not be considered the estimated parameter when weights are disregarded, its role is somewhat special. To highlight these facts, the following presentation of results is subdivided in accordance with these distinctions.

***r*-Based Estimators in $\mathfrak{S}_2$.** There are four *r*-based estimators under investigation: OP, OP-FE, OP-RE, and HS. Figure 8.5 gives an overview of the results for these estimators for varying *k* and *n* (upper panels).

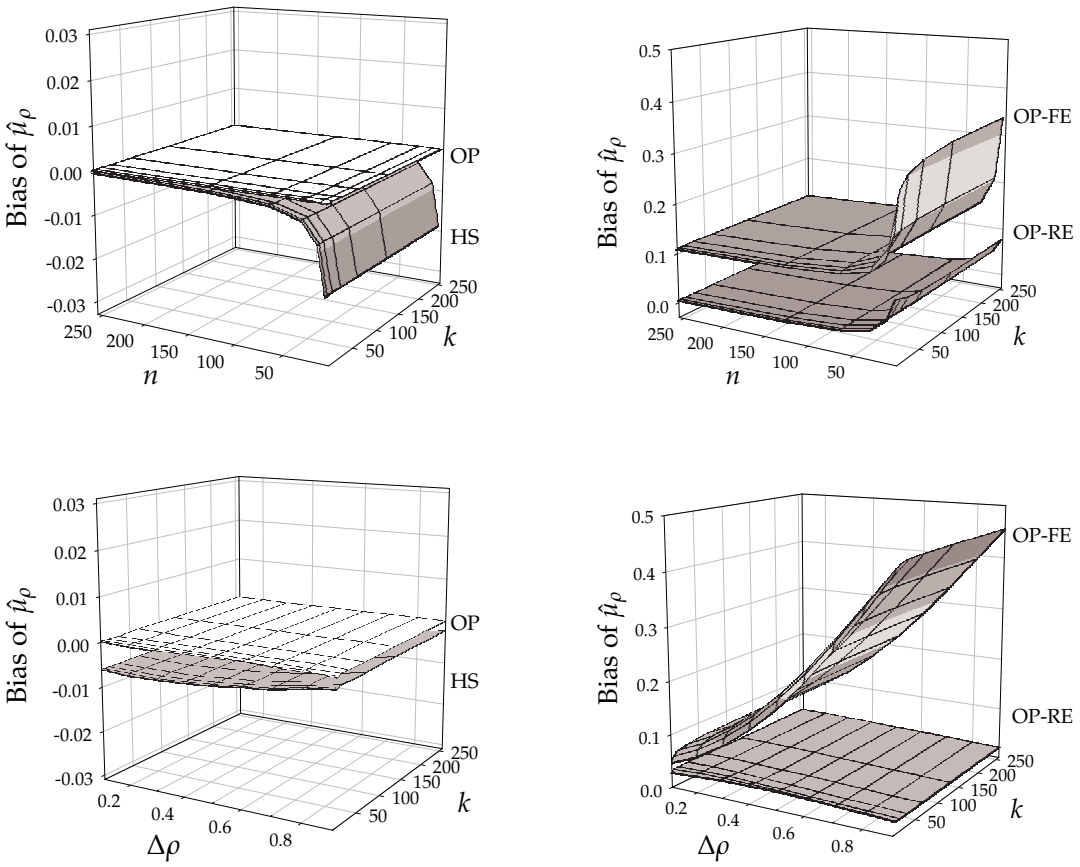

Figure 8.5 Bias of r-based μ_ρ estimators in $\mathfrak{S}_2$ by k and n (upper panels) as well as by $\Delta\rho$ and k (lower panels).

Both upper panels in the figure show similar behavior of the estimators as compared to the results in the previous situation. The only difference is an even worse performance for the OP-FE estimator approximating a value of .10 in bias with growing n. As before, OP is also in $\mathfrak{S}_2$ clearly the best estimator available in this category of estimators, showing almost no bias at all. HS also shows good performance, at least for sample sizes of 32 or larger.

The lower panels in Figure 8.5 depict the biases of the estimators across values of k and differences between ρ_1 and ρ_2, which will henceforth be denoted by $\Delta\rho$, that is, $\Delta\rho = \rho_1 - \rho_2$. The forms of the surfaces differ somewhat more from those in $\mathfrak{S}_1$. The direction of biases is still the same, with the OP-estimator being best across all values of the design. HS is depicted in the same graph and shows small negative biases which are almost invariant across values of $\Delta\rho$. Biases of HS can again be considered as negligible at least when sample sizes are 32 or larger.

The biases of OP-RE are approximately the same as those reported in $\mathfrak{S}_1$. OP-FE shows steadily increasing biases with higher values of $\Delta\rho$ that rapidly reach levels that can be considered to be unacceptable. As is evident from these results, the approximation of an OP-FE bias of .10 in the upper right panel is

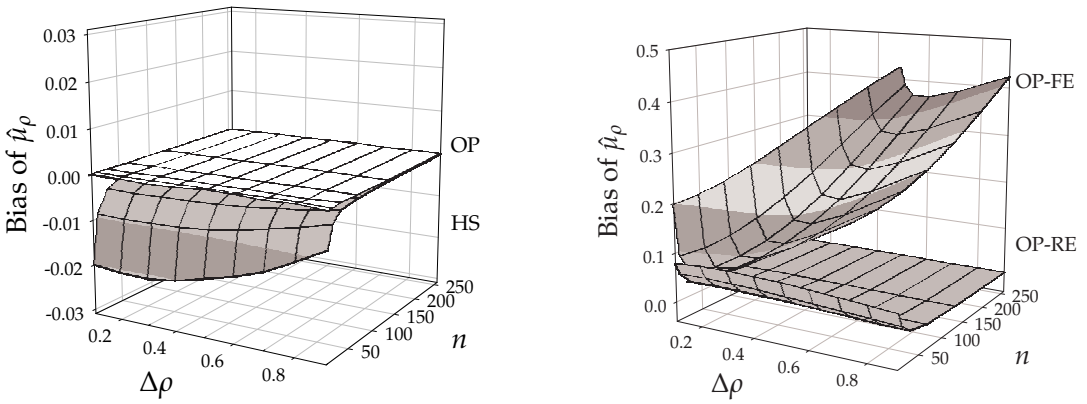

Figure 8.6 Bias of r-based μ_ρ estimators in $\mathfrak{S}_2$ by $\Delta\rho$ and n.

due to very large differences in bias across values for the difference between universe parameters. OP-FE is the only estimator in this class that is strongly affected by $\Delta\rho$. In addition to the strong effect of $\Delta\rho$, small n even amplify the bias depicted in Figure 8.5. This can be seen by inspecting the results shown in Figure 8.6.

The values for biases of r-based estimators across values of n and differences between ρs are depicted in Figure 8.6. The general trends and evaluation of the estimators do not change in comparison to $\mathfrak{S}_1$, as can be seen by inspection of this figure. OP is consistently showing a flat surface of zero bias across all values of n and $\Delta\rho$, hence it is clearly also the best point estimator of μ_ρ in $\mathfrak{S}_2$. HS only shows a small bias for very small n and does not perform as well as OP overall. In marked contrast, OP-RE and especially OP-FE show relatively bad performance, as can be seen in the right panel of Figure 8.6.

The reported biases of OP-FE across the design variables are huge in magnitude. It is remarkable that even with very high n biases do not diminish but actually rise. Such an observation is counterintuitive for at least two reasons. First, the results in the previous figures also show that the k point estimates of OP — on which OP-FE is based — are very accurate and show almost no bias at all in any of the situations and combinations of design variables. Hence, problems with biases of the OP-FE estimator cannot be caused by the point estimates. Second, consistency of estimators suggests that biases do not rise for increasing n but decline (and vanish for very large n). The converse is observed for OP-FE. All this clearly points to an effect of the weighting scheme because point estimates are very accurate on the basis of the UMVU estimator. Since the highest values for the bias of OP-FE are almost as high as the mean effect size in the universe (see the combination of lowest n and highest $\Delta\rho$ in the right panel of Figure 8.6), this shows that the class having higher ρ of the two-point distribution exclusively dominates the estimates. Hence, it must be the case that the correlations arising from the class with a higher ρ receive an excessive weight in comparison to the ones of the lower ρ class. Recall, first, that the weights are the reciprocals of the variances; second, that the variances of the estimator are different across values of ρ (see Figure 3.4); and third, that

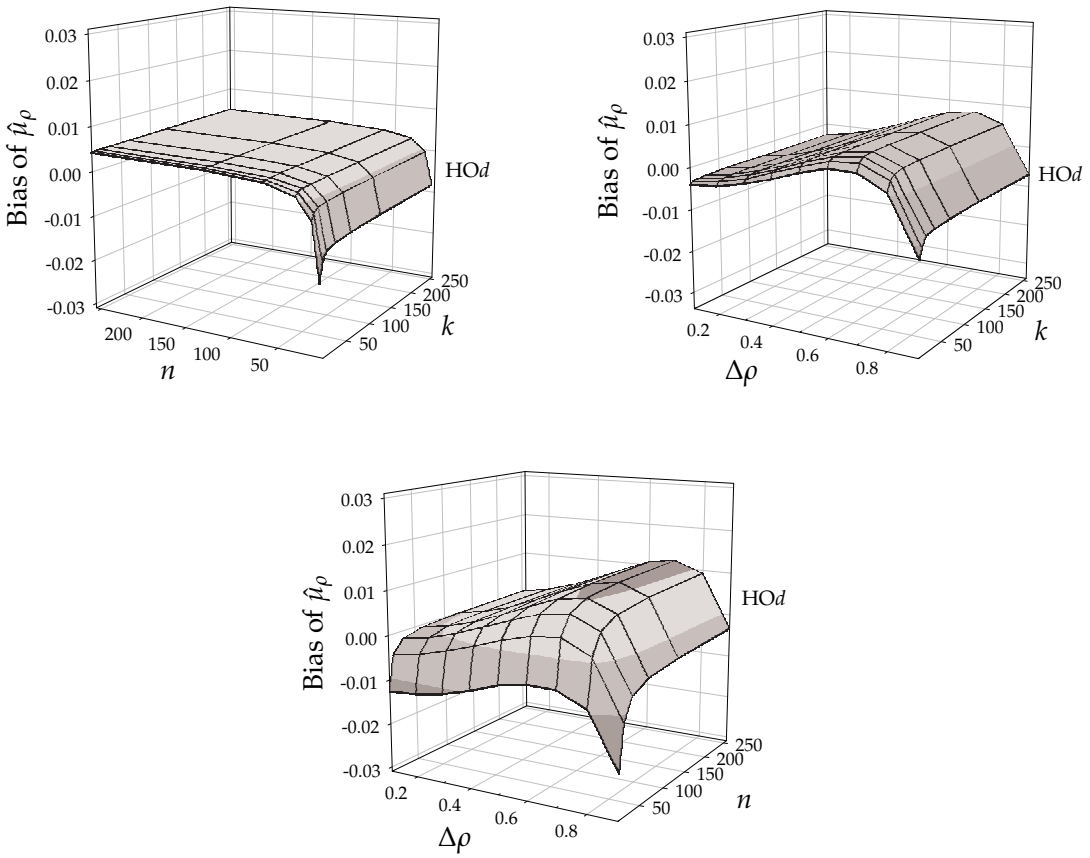

Figure 8.7 Bias of d-based μ_ρ estimators in $\mathfrak{S}_2$.

the estimates are plugged into Equation 3.7 to arrive at the estimates for the variances. Putting these facts together explains the high biases of OP-FE. As an example, consider the case of $\rho_1 = 0$, $n = 256$ and $\rho_2 = .90$, $n = 256$. For simplicity, assume that all r arising from ρ_1 are exactly zero and all r from ρ_2 are .90. As an aside, this is not far from what is actually observed with $n = 256$. In the given case, the weight for the first class is $w_1 = 252.00$, and $w_2 = 7104.54$ for the second. Applying these weights in the given situation and aggregating a total of $k = 256$ studies leads to an estimate of $\hat{\mu}_\rho = .87$. Subtracting $\mu_\rho = .45$ leads to a bias of .42, a value corresponding to the highest biases of OP-FE in the Monte Carlo study as can be observed in Figure 8.6, for example. Hence, large difference in weights lead to the huge biases observed for OP-FE in $\mathfrak{S}_2$.

Overall, the general trends in biases across levels of the design variables are similar to those resulting in $\mathfrak{S}_1$. Biases are fairly stable across values of k, OP clearly shows the best performance, and the weighting scheme emerges as a profound problem for OP-FE making the use of this estimator very unreasonable.

***d*-Based Estimator in $\mathfrak{S}_2$.** The results for the d-based estimator are depicted in Figure 8.7. As can be seen, biases show a somewhat strange behavior that differs from those of r-based estimators.

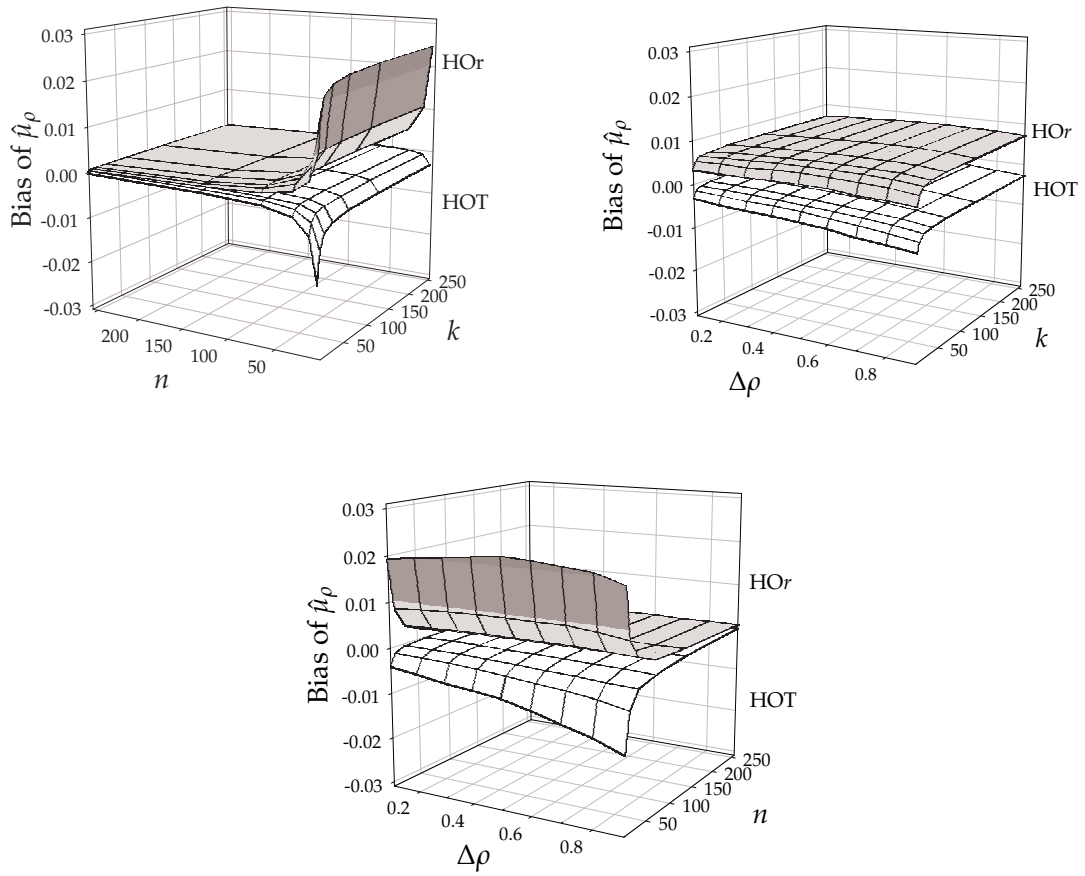

Figure 8.8 Bias of Fisher-z-based $\mu_{\rho z}$ estimators in $\mathfrak{S}_2$.

Although there are regions of the design variables where the d-based estimator shows no bias, it is quite sensitive with respect to differences between universe parameters in comparison to other estimators. The highest absolute values occur for combinations of small k and n. Recall again that the bias is computed with respect to μ_ρ and that varying weights of d also exert an influence on the behavior of the estimator. As a result, HOd shows a different behavior in bias in comparison to the situation $\mathfrak{S}_1$. Over- or underestimation of μ_ρ is harder to predict than for other estimators.

Fisher-z Based Estimators in $\mathfrak{S}_2$. The estimators of this category are HOr and HOT. To reiterate, as the universe parameter for these estimators $\mu_{\rho z}$ was used which differs from μ_ρ the larger the difference between ρ_1 and ρ_2 (see Section 5.5). The upper left panel of Figure 8.8 shows biases for this category of estimators similar to those in $\mathfrak{S}_1$.

Although HOT shows some deficiencies in bias with combinations of very small k and n it can still be considered as a better estimator than HOr. The estimates are therefore also improved in $\mathfrak{S}_2$ by the application of Hotelling's correction. The upper right panel of Figure 8.8 shows the biases by k and differences between ρs. In contrast to the r-based estimators there is a tendency

Table 8.3 Descriptive Statistics for the Bias of μ_ρ and $\mu_{\rho z}$ Estimators in $\mathfrak{S}_2$

	Statistic				
Approach	Max.	Mean	Median	Min.	SD
HO*r*	.0323	.0061	.0026	−.0012	.0078
HOT	.0011	−.0019	−.0006	−.0253	.0032
HO*d*	.0189	.0003	.0003	−.0343	.0078
HS	.0004	−.0058	−.0026	−.0289	.0070
OP	.0041	.0000	.0000	−.0051	.0005
OP-FE	.5175	.1622	.1324	.0011	.1256
OP-RE	.1166	.0297	.0188	.0006	.0283

Note. The total number of values described by these statistics is 1890.

of larger negative biases to occur for small values of k. The biases of estimators by n and differences between ρs are shown in the lower panel of the figure. As can be seen in the upper right and lower panel, biases of HO*r* do not differ very much across levels of $\Delta\rho$ and the same is true for k. The number of persons shows a strong influence on biases only for very small n. In sum, the Fisher-z-based estimators do not show larger biases of concern in comparison to the results reported in $\mathfrak{S}_1$, but estimators are only precise with respect to $\mu_{\rho z}$. Finally, a highly condensed overview of the descriptive statistics for the estimators in $\mathfrak{S}_2$ is presented in Table 8.3. The values in this table underscore the conclusions for $\mathfrak{S}_2$ already drawn.

In sum, values for biases are relatively small for all estimators, except for OP-FE, and may generally not be of concern at all, as was the case in $\mathfrak{S}_1$. Although OP-RE does not show mean biases in Table 8.3 as high as those for OP-RE, the estimates are highly variable in comparison to those of other estimators. This undesirable property points to the existence of cases in which the biases for this estimator are high. Hence, it does not appear attractive as an estimator even though it is designed for heterogeneous situations.

It must again be emphasized that biases have always to be judged against the background of different universe parameters. From a substantive point of view, r-based estimators address the parameter of interest best. The d-based estimator also performs relatively well in estimating μ_ρ but its bias is less predictable in comparison to r-based estimators. The results from meta-analyses for mean effect sizes in heterogeneous situations of type $\mathfrak{S}_2$ can therefore be interpreted as estimating quite accurately the expected value of the mixing distribution. However, for Fisher-z based estimators it has to be taken into account that it is $\mu_{\rho z}$ that is estimated, a parameter usually *not* of interest to the researcher. Hence, concerns are in order regarding the usage of Fisher-z based in heterogeneous situations like $\mathfrak{S}_2$. The interpretation of the mean effect size as a "mean ρ" is not warranted in a strict sense under these circumstances. Whereas the assumption of homogeneity can be regarded as a prerequisite for an interpretation of the mean effect size of z-based approaches in $\mathfrak{S}_2$, the re-

sults from the other approaches can safely be interpreted as estimates of the expected value of the distribution of ρs. Nevertheless, whether such an estimate is of real interest must be decided by the researcher based upon substantive concerns, because a vastly different set of ρ_1 and ρ_2 may have produced the observed mean effect size.

8.2.1.3 Heterogeneous Situation $\mathfrak{S}_3$ The last situation $\mathfrak{S}_3$ for which biases of the estimators will be examined is characterized by a continuous distribution of effect sizes in the universe of studies. Analogous to the introductory remarks made in the previous Subsection 8.2.1.2, the estimated parameters μ_ρ and $\mu_{\rho z}$ used as standards of comparison for the various estimators are considered first. For the case of correlation coefficients not subjected to the Fisher-z transformation, the expected value of the beta distribution is taken as the parameter of interest. For the Fisher-z transformed coefficients, the expected value in z-space (i.e., μ_ζ) given by

$$\mu_\zeta = \int_{-1}^{1} \tanh^{-1}(r) f(r) dr$$

constitutes the standard of evaluation. Here, $f(r)$ is the beta probability density function as described in Section 4.5. The values of μ_ζ are subsequently transformed into the space of r by the inverse Fisher-z transformation $\mu_{\rho z} = \tanh \mu_\zeta$. The resulting values computed for the expected values and variances of the beta distribution are reported in Tables A.1 and A.2 in the appendix. For the same reasons as in $\mathfrak{S}_2$, the expected value μ_ρ was used for the d-based estimator and results are presented separately for the three groups of estimators in the following paragraphs.

***r*-Based Estimators in $\mathfrak{S}_3$.** Biases of the estimators of this category over different combinations of the design variables can be inspected at a glance in Figure 8.9.

Evidently, the biases of the r-based estimators do not differ much from the previous two situations in overall quality. The OP estimator is again characterized by showing practically no biases notwithstanding which parameter constellation is prescribed by the design variables. The behavior of the estimators with respect to n and k is quite the same as before with biases showing practically no variation across values of k and larger biases for smaller n. It can also be seen that except for OP, the biases tend to be slightly smaller for HS as σ_ρ^2 becomes larger. The scaling of the vertical axis, however, shows that overall biases for HS are very small and only grow to a noticeable magnitude for extremely small sample sizes not likely to be encountered in practice. OP-FE again shows unacceptable behavior making it unsuitable as an estimator for situations of type $\mathfrak{S}_3$ as well. Hence, it does not seem reasonable to include OP-FE in all of the following performance evaluations of the various estimators.

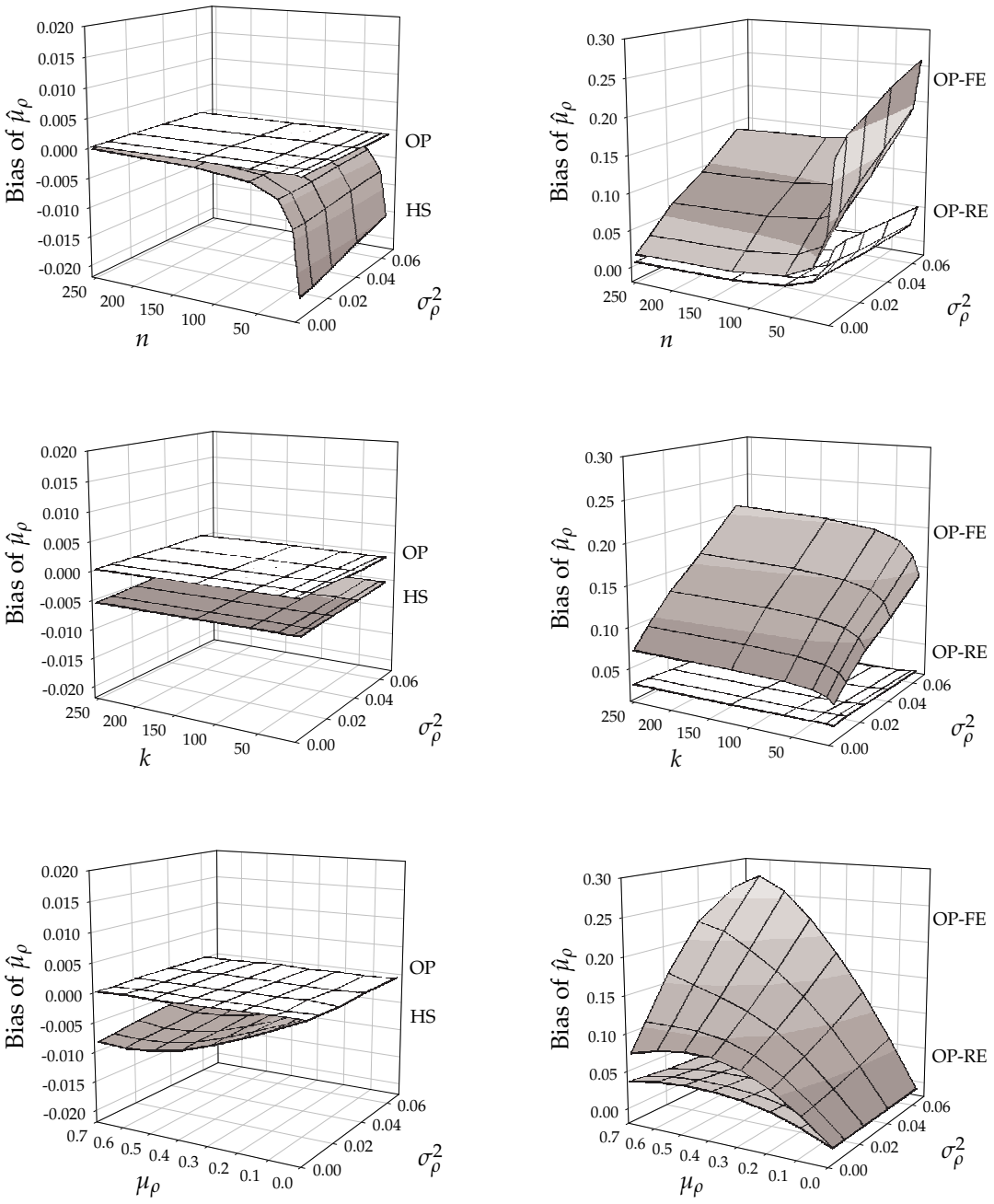

Figure 8.9 Bias of r-based μ_ρ estimators in $\mathfrak{S}_3$.

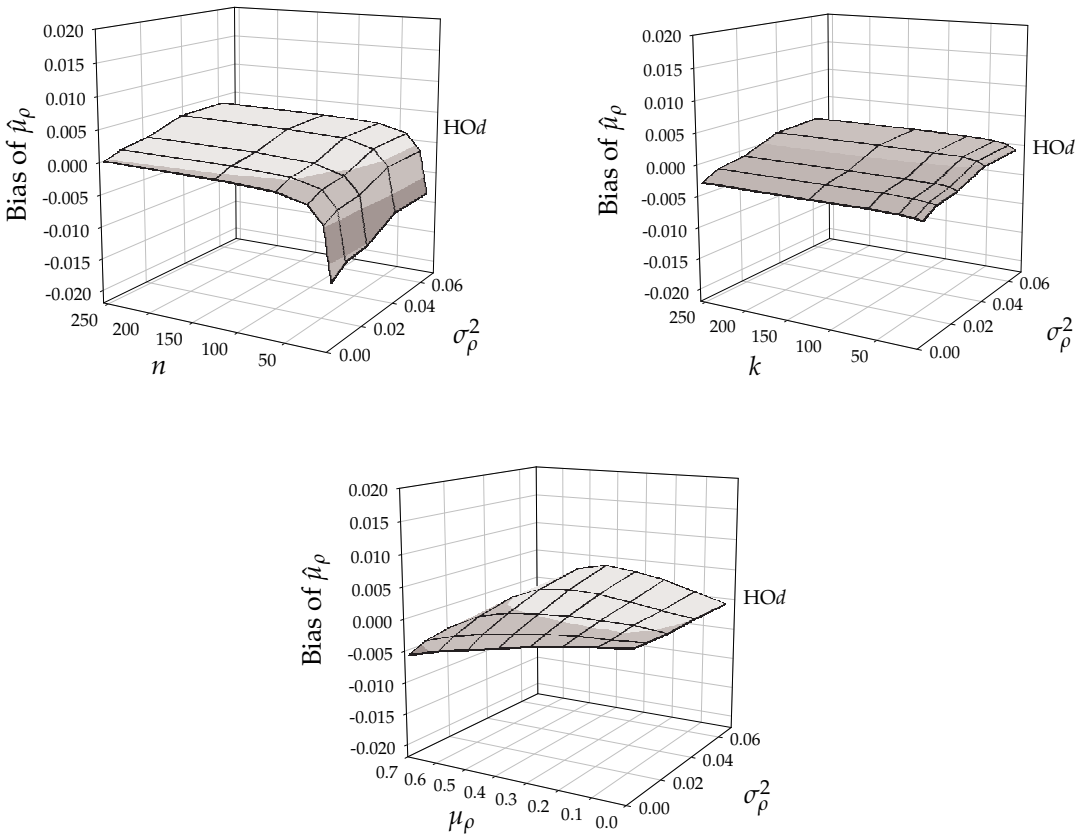

Figure 8.10 Bias of d-based μ_ρ estimators in $\mathfrak{S}_3$.

Note that the present situation is perfectly suitable for random effects approaches like OP-RE, but the performance in estimating μ_ρ is actually best for a fixed effects approach, namely OP. This is proposed to be due to the deficiencies of the weights used for OP-RE as already mentioned. Although the very good performance of OP is remarkable, disadvantages of FE approaches are suspected to lie more in testing, for example, rather than estimation of the universe parameter. The reader is also reminded that DSL, a random effects approach, leads in the present situation to the same results as HOr.

***d*-Based Estimator in $\mathfrak{S}_3$.** The next estimator for which results on biases are presented is the d-based estimator HOd. Again, an ensemble of graphs is given in Figure 8.10 to present results at a glance.

Evidently, biases of this estimator become larger only for very small n. This can be seen in the upper left panel in Figure 8.10. There is a slight tendency for higher values of HOd to occur for larger values of σ^2_ρ but again, the pattern of relationships of biases across design variables is not as clear as for other approaches. Large values of μ_ρ are accompanied by stronger negative biases (see lower panel). Although all the biases depicted in the three panels are very small in absolute terms, the observed effects are supposed to be due to the weights employed in computing the mean effect sizes using this estimator.

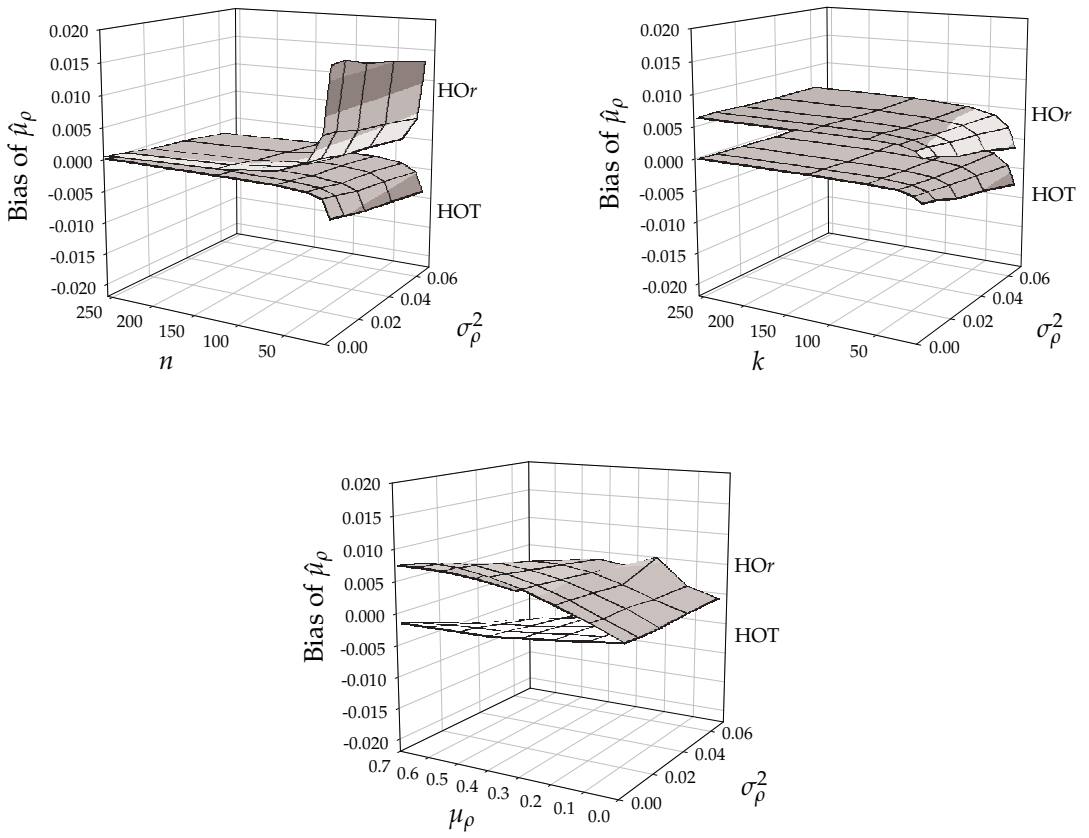

Figure 8.11 Bias of Fisher-z-based $\mu_{\rho z}$ estimators in $\mathfrak{S}_3$.

Because larger values of d are downweighted by using the weights, as already discussed in detail, and such values occur more often with larger variances of the mixing distribution, the negative bias for small σ^2_ρ visible in the upper right panel of Figure 8.10 seems to be compensated. For large values of μ_ρ, a stronger negative bias results but all in all the values of bias are very small and not of practical concern except for cases of very small sample sizes n.

Fisher-z-based estimators in $\mathfrak{S}_3$. The biases of the Fisher-z-based estimators are presented in Figure 8.11. The relevant estimators in this class are HOr and HOT.

As in the situations before, HOT performs better than the non-corrected Fisher-z-based estimator HOr. Very small n influences biases of these estimators in a negative way and can lead to a noticeable bias of the HOr estimator. Nevertheless, biases are not large in general and only become discernible for extreme levels of the design variables, especially n (see upper left panel in Figure 8.11. The variance of effect sizes σ^2_ρ does not have a profound effect on the estimates of $\mu_{\rho z}$. Especially in the lower panel of Figure 8.11 some values for the estimators are hard to inspect precisely. Finally, descriptive statistics for the biases in $\mathfrak{S}_3$ are again presented in Table 8.4 for an overview of the results in $\mathfrak{S}_3$.

Table 8.4 Descriptive Statistics for the Bias of Estimators μ_ρ in $\mathfrak{S}_3$

	Statistic				
Approach	Max.	Mean	Median	Min.	SD
HO*r*	.0325	.0039	.0013	−.0286	.0080
HOT	.0038	−.0025	−.0008	−.0447	.0045
HO*d*	.0090	−.0014	.0000	−.0337	.0060
HS	.0037	−.0053	−.0022	−.0306	.0071
OP	.0037	.0000	.0000	−.0052	.0007
OP-FE	.4545	.1071	.0775	−.0068	.0992
OP-RE	.1161	.0248	.0126	−.0073	.0276

Note. Valid values for all entries are 1848.

Evidently, a similar picture as compared to $\mathfrak{S}_2$ emerges. It can be seen that biases for OP-RE and OP-FE are far more variable in comparison to the other estimators and can produce biases in maximum that are certainly not acceptable. All other estimators fare quite well with respect to the parameters they estimate. OP is clearly the best estimator also in the given situation. It is not only closest on target overall, but also shows the smallest variability in biases. For a comparison of the Fisher-z-based and the other estimators it is quite instructive to also consult Tables A.1 and A.2 in the appendix to gain an impression of how different μ_ρ and $\mu_{\rho z}$ can become.

8.2.2 Relative Efficiency

Besides the bias of an estimator as an expression of how close it is to the estimated parameter, the variance is also an aspect of its closeness to the parameter. However, it is not the variance of an estimator per se that is of concern here but the variance about the parameter of interest. That is, the squared distances from the universe parameter to be estimated are taken and not the ones with respect to the expected value of a potentially biased estimator. The well-known decomposition (see Stuart et al., 1999, p. 24)

$$\mathrm{MSE}(T) = E(T-\theta)^2 = \mathrm{Var}(T) + (E(T)-\theta)^2 \tag{8.1}$$

shows that for unbiased estimators T the mean squared error (MSE) equals the variance of the estimator. In the Monte Carlo study, the values computed are actually

$$\mathrm{MSE} = \frac{1}{\mathrm{Iter}} \sum_{l=1}^{\mathrm{Iter}} (\hat{r}_l - \mu_\rho)^2,$$

where *Iter* signifies the number of iterations (10,000) and $\hat{r}_l$ denotes an estimator based on r. That is, the squared distances from the expected value of the distribution of universe effect sizes are summed over all iterations. In the following comparisons, the MSE-ratios of different estimators will be presented.

In $\mathfrak{S}_1$, μ_ρ is used as the universe parameter in the computation of the MSEs, whereas in $\mathfrak{S}_2$ and $\mathfrak{S}_3$ the question of choosing an appropriate universe parameter for comparing the approaches arises again. For the approaches that use the Fisher-z transformation in estimating the mean effect size, there are two possibilities. First, the MSEs can be computed in z-space. That is,

$$\text{MSE}_z = \frac{1}{\text{Iter}} \sum_{l=1}^{\text{Iter}} (\hat{z}_l - \mu_\zeta)^2,$$

where $\hat{z}_l$ is the estimator and μ_ζ the expected value of the mixing distribution in z-space. The additional problem arises that the MSEs for all situations of z-based estimators are not directly comparable to MSEs for r-based estimators. To make the MSEs of the various estimators in $\mathfrak{S}_2$ comparable, the term

$$h(\rho) = \frac{\sqrt{1+\rho_1+\rho_2+\rho_1\rho_2} - \sqrt{1-\rho_1-\rho_2+\rho_1\rho_2}}{\sqrt{1+\rho_1+\rho_2+\rho_1\rho_2} + \sqrt{1-\rho_1-\rho_2+\rho_1\rho_2}} - \frac{\rho_1+\rho_2}{2},$$

which is the difference between μ_ρ and $\mu_{\rho z}$, is used for correction. The correction factor gives the difference between μ_ρ and $\mu_{\rho z}$ *theoretically* to be expected for the various parameter values in $\mathfrak{S}_2$. Subtracting $h(\rho)$ from the Fisher-z transformed values ($\tanh \hat{z}_l$) offers the opportunity to compare the MSEs of the Fisher-z based and r-based estimators on a common scale via

$$\text{MSE} = \frac{1}{\text{Iter}} \sum_{l=1}^{\text{Iter}} \big((\tanh \hat{z}_l - h(\rho)) - \mu_\rho\big)^2.$$

What should become evident here is that in essence the estimator is actually changed by $h(\rho)$ to estimate a different value, namely μ_ρ. For this corrected estimator, the mean squared distances about μ_ρ are computed as for the r-based estimators. Of course, the above equation can be simplified by eliminating the redundant term μ_ρ, resulting in

$$\text{MSE} = \frac{1}{\text{Iter}} \sum_{l=1}^{\text{Iter}} \big(\tanh \hat{z}_l - \mu_\zeta\big)^2,$$

which may be conceived as the natural conception for computing the MSEs for Fisher-z-based estimators in r-space. The derivation given above has just demonstrated that using this conception of the MSE can be justified on theoretical grounds.

Following this general logic, it would be natural to use the expected values of the beta distribution in the computation of the MSEs in $\mathfrak{S}_3$ correspondingly. Unfortunately, numerical integration necessary to compute these values was considered to be computationally too expensive and a correction factor like $h(\rho)$ is not readily available for the continuous case. Hence, the comparison of all approaches is not possible in $\mathfrak{S}_3$.

Table 8.5 Relative Efficiencies of $\hat{\mu}_\rho$ in $\mathfrak{S}_1$

Approach	HO*r*	HOT	HO*d*	HS	OP	OP-RE
HO*r*	1					
HOT	1.1118	1				
HO*d*	1.1008	.9901	1			
HS	1.0931	.9832	.9930	1		
OP	1.0683	.9609	.9705	.9773	1	
OP-FE	.4741	.4264	.4307	.4337	.4438	1

Note. Table entries are the fraction of the approach found in the column header in the numerator and the row-labeled approach in the denominator. For all approaches mean values were computed over all values of ρ, k, and n.

As in the previous sections, results are presented for $\mathfrak{S}_1$ and $\mathfrak{S}_2$ consecutively, beginning with $\mathfrak{S}_1$. First, it can be noted that all MSEs of the estimators become smaller for larger k, n, and μ_ρ, respectively. Since these general trends apply to all approaches the presentation will be confined to overall results, that is, means of MSEs over all values of k, n, and μ_ρ. Table 8.5 provides a condensed overview of the relative efficiencies of the estimators in $\mathfrak{S}_1$.

Note that in $\mathfrak{S}_1$ the MSEs are comparable for all estimators without any corrections because $\mu_\rho = \mu_\zeta$. The entries in the table can be read as follows. The estimators found in the column are the entry in the numerator and the estimator in the row is the denominator of a fraction of MSEs. Values larger than one therefore represent smaller MSEs for the estimator in the denominator and vice versa. The values in Table 8.5 suggest that HOT is the most efficient estimator in terms of MSE. This is somewhat surprising given that OP has shown remarkably small biases overall, as shown in the previous sections. The reason for this finding lies in the fact that *variances* of the estimators contribute an important part to the corresponding MSEs and also to variation of MSEs across values of the design dimensions. Because the variability across levels of the design variables is very similar for all approaches, no graphical representation is given here.

To facilitate interpretation of the results in Table 8.5, the following remarks seem warranted. As can be seen in Equation 8.1, the MSEs amount to the variances of unbiased estimators and these variances may well be larger for unbiased in comparison to biased estimators. The important distinction between MSEs for these two types of estimators lies in the second term of Equation 8.1 being zero for unbiased and nonzero for biased estimators. Now consider the case of $\mu_\rho = 0$ where the HS (and also OP) estimator actually estimates μ_ρ (see page 121, for example). Here, the variance of the HS estimator is smaller than the variance of the OP estimator and the MSE ratio restricted to this case leads to a value of .9059 in favor of HS. On the other hand, if the comparison is restricted to $\rho = .50$ where the HS estimator performs worst in terms of bias, the same comparison leads to a value for the ratio of 1.0295, now favoring OP in terms of MSE. Of course, this phenomenon pertains to all comparisons in

Table 8.6 Relative Efficiencies of $\hat{\mu}_\rho$ and $\hat{\mu}_{\rho z}$ in $\mathfrak{S}_2$

Approach	HO*r*	HOT	HO*d*	HS	OP	OP-RE
HO*r*	1					
HOT	1.0968	1				
HO*d*	1.0425	.9505	1			
HS	1.0582	.9648	1.0150	1		
OP	1.0290	.9381	.9870	.9724	1	
OP-FE	.4829	.4403	.4632	.4564	.4693	1

Note. Table entries are the fraction of the approach found in the column header in the numerator and the row-labeled approach in the denominator. For all approaches mean values were computed over all values of ρ, k, and n.

Table 8.5 and also what follows. The values reported in the tables for a comparison of estimators regarding their MSEs are therefore to be interpreted with respect to the performance of estimators across the levels of the design variables. Hence, across all values of μ_ρ from zero to .90, HOT is the most efficient estimator. This does not imply that HOT is the most efficient estimator for all possible values of μ_ρ.

Table 8.6 presents the results for $\mathfrak{S}_2$ that closely mirror the results in $\mathfrak{S}_1$. The only notable difference is that HO*d* is slightly less efficient than HS in this situation.

As remarked at the beginning of this subsection, the comparison of estimators with respect to MSE in $\mathfrak{S}_3$ is not possible due to values in different spaces that could not be transformed or corrected. In sum, the surprising result for the MSEs is that OP — which performed uniformly best in all situations with respect to bias — does not also show up as the best estimator in terms of MSE. As a result, it is not more precise in general as compared to HS when all situations under investigation are taken into account (see in this context Hedges & Olkin, 1985, p. 226). Instead, HOT performed best, an estimator that also showed good performance with respect to biases. Taking these two criteria together, it can be recommended for $\mathfrak{S}_1$ to use OP when a relatively small n is given and when μ_ρ is not suspected to be very low. HOT can be considered as an alternative for larger n when μ_ρ is suspected to be small because of its higher efficiency. For $\mathfrak{S}_2$ and $\mathfrak{S}_3$ OP should be considered as first choice for it estimates the parameter usually of interest μ_ρ rather precisely in terms of bias, in contrast to HOT which estimates $\mu_{\rho z}$.

8.3 SIGNIFICANCE TESTS FOR THE MEAN EFFECT SIZE: TYPE I ERRORS AND POWER

Apart from an accurate parameter estimation, significance tests are a common feature of meta-analysis. The significance testing practice in psychology has often been criticized as mentioned in the introductory chapters. Nonetheless,

all approaches offer procedures to test the estimates of the effect sizes, although some authors explicitly deemphasize using such tests (e.g., Hunter & Schmidt, 1990). In this section, the proposed procedures are evaluated with respect to their performance in testing the generally adopted null hypothesis $\mu_\rho = 0$. This will be done by examining the rejection rates of the null hypothesis under various conditions and comparing these rejection rates to the α-level a test is supposed not to exceed when the null hypothesis is true. The second case of interest is the performance of the tests when the null hypothesis is false. Both cases will be separated in the following presentation.

Since this subsection on testing is the only one in which differences between HO*r* and RR can occur, the latter will be included in the following tables for this subsection only. Because of the very poor performance of OP-FE reported in the previous sections of results, it will be omitted from the following presentation.

Rejection Rates in $\mathfrak{S}_1$ The results for $\mathfrak{S}_1$ are considered first. In addition to the estimators under investigation in the previous sections, four variants for testing the mean effect size in the framework of the HS approach are included. The several variants correspond to four possibilities to compute the standard error of the mean effect size. The reader is referred to Section 5.3 for a recapitulation of the several forms of standard errors proposed in this approach. Furthermore, the DSL approach is also added to examine its performance in the homogeneous case. As a criterion for the evaluation of the approaches, whether they show *higher* rates than nominal α will be assessed. Values lower than α are interpreted as indicating good performance since the null hypothesis is true.

Table 8.7 shows the rejection rates of the tests aggregated over all combinations of k and n. Readers interested in the results for specific combinations of the design variables may consult Table C.1 in the appendix where detailed results for $\alpha = .05$ are provided.

The first line for each approach in Table 8.7 provides the results for tests at $\alpha = .05$ and the second line those for $\alpha = .01$. As can be seen, the approaches that perform best are DSL and HOT. Despite both approaches having slightly higher standard deviations in comparison to HO*r*, which more closely attains nominal α, they also show mean rejection rates below α in this situation. At least for DSL — a random effects approach — it is suspected that this more conservative behavior comes at the cost of a loss in power. The downward correction of the mean Fisher-z based effect size by HOT effectively leads to smaller rejection rates in comparison to HO*r*. Note that the same standard errors were applied for HO*r* and HOT. Hence, apparently small differences in bias between these approaches indeed transfer to differences in test results.

For the present purpose of testing the null hypothesis, there are small procedural differences between HO*r* and RR, which were omitted until now. The only difference between RR and HO*r* lies in the weights employed where RR uses degrees of freedom instead of standard error based weights. This ob-

Table 8.7 Rejection Rates for Testing the Mean Effect Size in $\mathfrak{S}_1$, $\mu_\rho = 0$

	Statistic				
Approach	Max.	Mean	Median	Min.	SD
HO*r*	.0547	.0501	.0504	.0437	.0022
	.0119	.0100	.0099	.0077	.0009
HOT	.0535	.0458	.0472	.0313	.0054
	.0110	.0086	.0090	.0044	.0017
HO*d*	.0702	.0554	.0530	.0476	.0059
	.0190	.0119	.0111	.0078	.0025
RR	.0688	.0543	.0523	.0475	.0052
	.0170	.0114	.0110	.0078	.0020
HS1	.0695	.0519	.0511	.0460	.0041
	.0250	.0111	.0104	.0077	.0027
HS2	.0866	.0591	.0547	.0480	.0104
	.0316	.0141	.0120	.0078	.0054
HS3	.1291	.0705	.0577	.0475	.0257
	.0741	.0247	.0140	.0096	.0205
HS4	.1256	.0645	.0539	.0330	.0261
	.0701	.0219	.0123	.0050	.0196
OP	.0853	.0560	.0531	.0475	.0079
	.0341	.0129	.0115	.0078	.0046
OP-RE	.2224	.0849	.0582	.0433	.0491
	.1392	.0320	.0139	.0066	.0347
DSL	.0503	.0430	.0434	.0356	.0036
	.0101	.0079	.0080	.0057	.0011

Note. The total number of values described by these statistics is 42. Proportion for tests at $\alpha = .05$ are given in the first row of each approach and for tests at $\alpha = .01$ in the second row.

viously leads to slightly higher rejection rates in comparison to the nominal α-level.

Amongst the HS variants two groups can be identified: HS1 and HS2 versus HS3 and HS4. This corresponds to standard errors proposed for homogeneous (HS1 and HS2) and heterogeneous (HS3 and HS4) situations. Since HS3 is assumed to be adequate both for homogeneous and heterogeneous situations (Osburn & Callender, 1992), special attention may be paid to the results of this particular variant. The results in Table 8.7 show that, for the group of HS-variants proposed for homogeneous situations such as the present one, HS1 is closest on the α-levels whereas HS2 overshoots. Usage of the tests from the second group leads to rejection rates being too high overall. This predominantly occurs in cases where a low number of studies are aggregated.

OP performs as well (or bad) as RR and HO*d* whereas the random effects approaches behave quite differently. DSL performs as expected from theory,

Table 8.8 Rejection Rates for Testing the Mean Effect Size in $\mathfrak{S}_1$, $\mu_\rho \neq 0$, $\alpha = .05$

	μ_ρ								
Approach	.10	.20	.30	.40	.50	.60	.70	.80	.90
HO*r*	.7856	.9228	.9690	.9875	.9958	.9990	.9999	1	1
HOT	.7761	.9162	.9650	.9852	.9946	.9986	.9998	1	1
HO*d*	.7965	.9300	.9732	.9897	.9966	.9992	.9999	1	1
RR	.7945	.9287	.9724	.9894	.9965	.9992	.9999	1	1
HS1	.7907	.9275	.9724	.9895	.9966	.9992	.9999	1	1
HS2	.8035	.9348	.9762	.9913	.9974	.9995	1	1	1
HS3	.8094	.9357	.9753	.9901	.9965	.9989	.9998	1	1
HS4	.7978	.9291	.9718	.9884	.9958	.9987	.9997	.9999	1
OP	.7985	.9325	.9751	.9908	.9972	.9994	.9999	1	1
OP-RE	.8162	.9385	.9765	.9904	.9965	.9989	.9997	.9999	1
DSL	.7680	.9122	.9625	.9836	.9937	.9980	.9996	.9999	1

Note. The total number of values described by these statistics is 42 for each μ_ρ.

showing rejection rates below the nominal α due to overestimates of random effects variance in this situation (see also Section 8.5). OP-RE in contrast, shows very high rejection rates, a fact that would not be expected for random effects approaches. The reason for this finding is the bad performance of OP-RE in cases of $n < 64$. A combination of small n and *high* k exacerbate this malperformance. Again, bad results from estimation of the mean effect size transfer to bad results in significance testing.

The case of $\mu_\rho \neq 0$ can be considered to enable an examination of the test results with regard to their power.[3] That is, rejection rates for the null hypothesis are presented when it is actually false. Hence, they should be interpreted as rate estimates of correctly rejecting the null hypothesis. Table 8.8 provides an overview of the results for increasing μ_ρ and $\alpha = .05$.

As can be seen in the table, all the tests seem to rapidly attain satisfactory[4] levels of .80 when aggregated across k and n. Of course, approaches showing higher rejection rates when the null hypothesis is true generally perform better in this context. Results for the rejection rates when $\alpha = .01$ are not presented. They show a similar performance for the approaches as compared to those in the presented case.

As would be expected from theory, rejection rates are larger for higher levels of *n and k*. Satisfactory power levels are rapidly reached even for modest values of $k = 32$, for example. The general trends are illustrated in Figure 8.12.

[3]This term is used somewhat loosely in the present context because no alternative hypotheses are explicitly considered. The values to be presented for the rejection rates may nevertheless be regarded as some approximation to the power function of the tests (cf. Barnett, 1981).
[4]According to Cohen (1988, 1992).

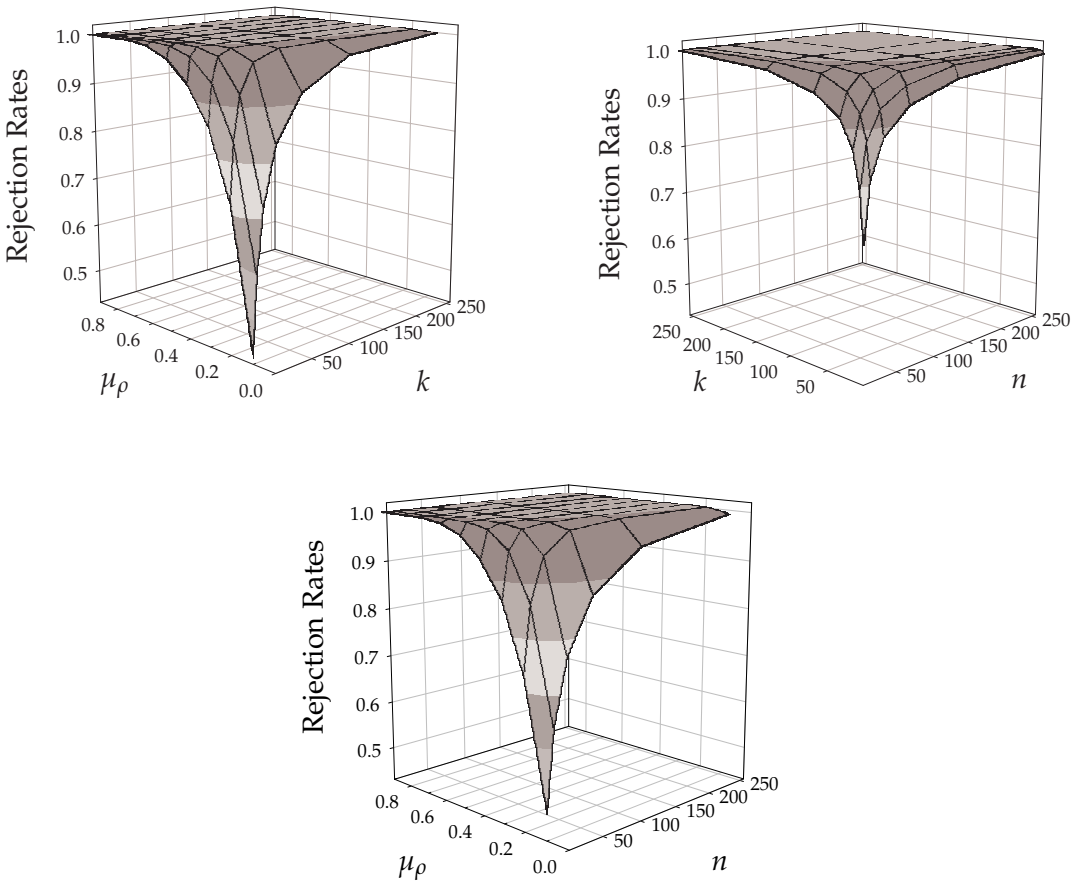

Figure 8.12 Rejection rates for testing the mean effect size in $\mathfrak{S}_1$, $\mu_\rho \neq 0$, $\alpha = .05$, HO*r* approach.

Here, the results for only one approach (HO*r*) are depicted since the trends are the same for all approaches and the surfaces would not be discriminable. In general, differences between approaches are not very large in testing the null hypothesis in $\mathfrak{S}_1$. HO*r* exhibits a performance closest to nominal α-levels when the null hypothesis is true and all approaches reach satisfactory levels for power rather quickly.

Nonetheless, the three panels in Figure 8.12 also point to cases for which power might not be satisfactorily high. An additional table is provided in the appendix (Table C.2) which is especially informative to qualify the results in Table 8.8 with respect to k and n. It shows rejection rates for selected levels of design variables supposed to be of highest interest with respect to regions in the figure where power is not very high. The results basically underscore the general impression gained from the figure. Power can be low for small effect sizes in the universe ($\rho = .10$), especially when n and k are very low. Even when n is at a level presumably considered as moderate or sufficient by many researchers ($n = 126$) and which can be observed quite often in correlational studies in the behavioral sciences, detection of a small universe effect size can be conducted without reaching satisfactorily high levels of power. For exam-

Table 8.9 Rejection Rates for Testing the Mean Effect Size in $\mathfrak{S}_2$ for Selected $\mu_\rho \neq 0, \alpha = .05$

	μ_ρ							
Approach	.05	.10	.15	.20	.30	.40	.50	.60
HO*r*	.5815	.7889	.8787	.9270	.9730	.9915	.9978	.9995
HOT	.5703	.7792	.8707	.9205	.9691	.9896	.9971	.9993
HO*d*	.5936	.7976	.8856	.9313	.9739	.9900	.9961	.9990
RR	.5910	.7957	.8840	.9301	.9736	.9908	.9973	.9994
HS1	.5847	.7917	.8818	.9290	.9735	.9910	.9974	.9994
HS2	.6025	.8066	.8930	.9381	.9790	.9939	.9986	.9997
HS3	.5875	.7669	.8679	.9083	.9563	.9768	.9914	.9978
HS4	.5725	.7544	.8582	.9004	.9512	.9738	.9899	.9974
OP	.5944	.7995	.8879	.9335	.9758	.9918	.9976	.9995
OP-RE	.6145	.7891	.8792	.9165	.9542	.9609	.9817	.9963
DSL	.5444	.7309	.8366	.8822	.9385	.9653	.9872	.9966

Note. The total number of values described by these statistics varies between 42 and 168 for each μ_ρ due to some differences occurring more often than others for the combination of design variable levels. For the omitted values of $\mu_\rho \geq .70$ all values are practically equal to 1.

ple, power is lower than .80 for all approaches in cases where $k = 4$, $n = 128$, and $\rho = .10$. Overall, the approaches do not vary greatly in behavior with respect to power in $\mathfrak{S}_1$. Although some approaches (e.g., DSL) show lower rejections rates than others (e.g., OP), differences are small in comparison.

Rejection Rates in $\mathfrak{S}_2$ Next, we turn to the test results in $\mathfrak{S}_2$. The null hypothesis μ_ρ is always false in $\mathfrak{S}_2$ because at least one $\rho \neq 0$. Accordingly, the results of the Monte Carlo study only reflect the power of the tests. In Table 8.9 the mean rejection rates are presented for varying values of μ_ρ.

Once more, rejection rates show only small differences between approaches. The overall trends as presented in Table 8.9 are similar in comparison to the results in $\mathfrak{S}_1$. As far as general trends across the design variables are concerned, only a selection of figures is presented here to illustrate the largest differences between the approaches.

The series of graphs in Figure 8.13 depict the dependencies of rejection rates on the design variables n, k, and μ_ρ for the approaches HO*r* and DSL. All other approaches show a performance "in between" the ones presented. As can be seen by comparison of the left and right panels, the random effects approach leads to more conservative test results especially for small k and intermediate μ_ρ. This is due to incorporation of heterogeneity variances in the standard errors of the tests making them more conservative than fixed effects approaches. This difference occurs for DSL and also OP-RE not shown in the figure. In comparison to the homogeneous situation $\mathfrak{S}_1$ the tests are not as powerful

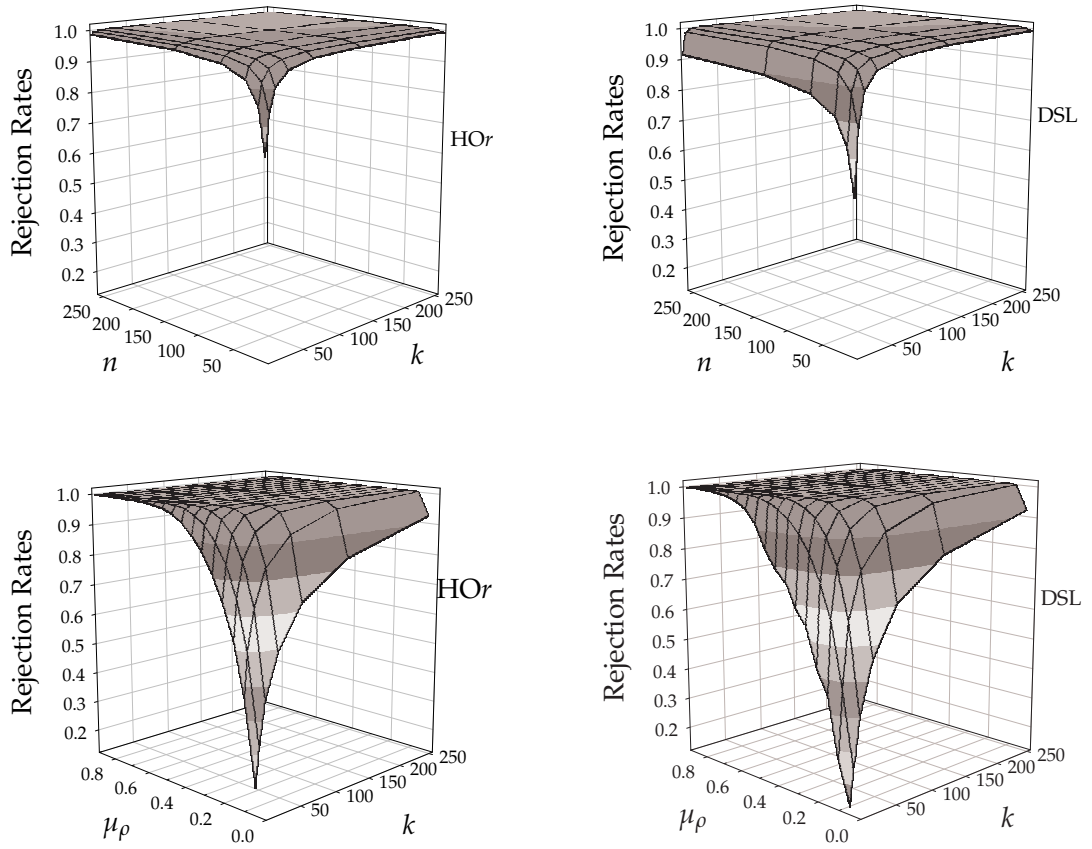

Figure 8.13 Rejection rates for testing for the mean effect size in $\mathfrak{S}_2$, $\mu_\rho \neq 0, \alpha = .05$, HO*r* and DSL approach.

for small effects (e.g., $\rho = .10$). The lower panels in Figure 8.13 in particular point to the result of inadequate power for the approaches for the boundary regions of the design. The interested reader might wish to consult Tables C.3 and C.4 where detailed results for these boundary regions are presented. In short, even when the study sample size seems reasonable for the aggregated studies in a meta-analysis ($n = 128$) and a number of studies — rather typical for some meta-analyses and considered by some as "large" — of $k = 32$ is available, small effects of $\mu_\rho = .05$ are not detectable with a power of .80 by any of the approaches. However, the power raises very quickly for all approaches for higher effects in the universe of studies. In sum, performance of the approaches in testing the generally adopted null hypothesis $\mu_\rho = 0$ in $\mathfrak{S}_2$ is very similar. In cases where power problems seem to prevail, none of the approaches seem to offer a considerable advantage over the others.

Rejection Rates in $\mathfrak{S}_3$ The last part of results for significance tests is presented for $\mathfrak{S}_3$, where two cases are distinguished. As in $\mathfrak{S}_1$, results for $\mu_\rho = 0$ will first be given followed by the results for $\mu_\rho \neq 0$. Table 8.10 provides a condensed overview of the results for the case when the null hypothesis is true.

Table 8.10 Rejection Rates for Testing the Mean Effect Size in $\mathfrak{S}_3$, $\mu_\rho = 0$

	Statistic				
Approach	Max.	Mean	Median	Min.	SD
HO*r*	.3599	.1410	.1026	.0472	.0880
	.3046	.0775	.0364	.0089	.0795
HOT	.3596	.1360	.0994	.0338	.0912
	.3044	.0749	.0349	.0054	.0806
HO*d*	.3539	.1427	.1043	.0576	.0828
	.2946	.0769	.0369	.0127	.0754
RR	.3519	.1408	.1013	.0562	.0828
	.2928	.0753	.0359	.0119	.0751
HS1	.3518	.1381	.1003	.0485	.0845
	.2926	.0747	.0358	.0095	.0754
HS2	.3603	.1510	.1145	.0595	.0821
	.3043	.0837	.0449	.0144	.0766
HS3	.1347	.0698	.0583	.0465	.0250
	.0760	.0246	.0144	.0079	.0201
HS4	.1289	.0639	.0541	.0318	.0255
	.0743	.0220	.0131	.0040	.0195
OP	.3519	.1425	.1041	.0566	.0818
	.2928	.0773	.0391	.0129	.0743
OP-RE	.2139	.0906	.0814	.0499	.0335
	.1336	.0351	.0260	.0094	.0236
DSL	.1005	.0551	.0527	.0349	.0111
	.0494	.0143	.0114	.0055	.0079

Note. The total number of values described by these statistics is 210. Proportion for tests at $\alpha = .05$ are given in the first row of each approach and for tests at $\alpha = .01$ in the second row.

As can be seen from the results in Table 8.10, the performance of the approaches can roughly be categorized in two groups. On the one hand, fixed effects approaches like HO*r*, HO*d*, HS1, and OP, for example, show relatively large inflated mean Type I error rates. These approaches show adequate rejection rates only in minimum and also have relatively high standard deviations. On the other hand, random effects approaches like OP-RE and especially DSL perform adequately overall in this situation. The HS variants HS3 and HS4 perform like the random effects approaches though not as well as DSL, for example. Hence, the violation of basic assumptions of the fixed effects approach in $\mathfrak{S}_3$ leads to differences in rejection rates. This was not as clear in $\mathfrak{S}_2$, though this is a heterogeneous situation too.

Apart from the overall performance, it should be mentioned that the results markedly differ across levels of the design variables. The results for varying levels of these variables are therefore presented next. In Figure 8.14, a selection

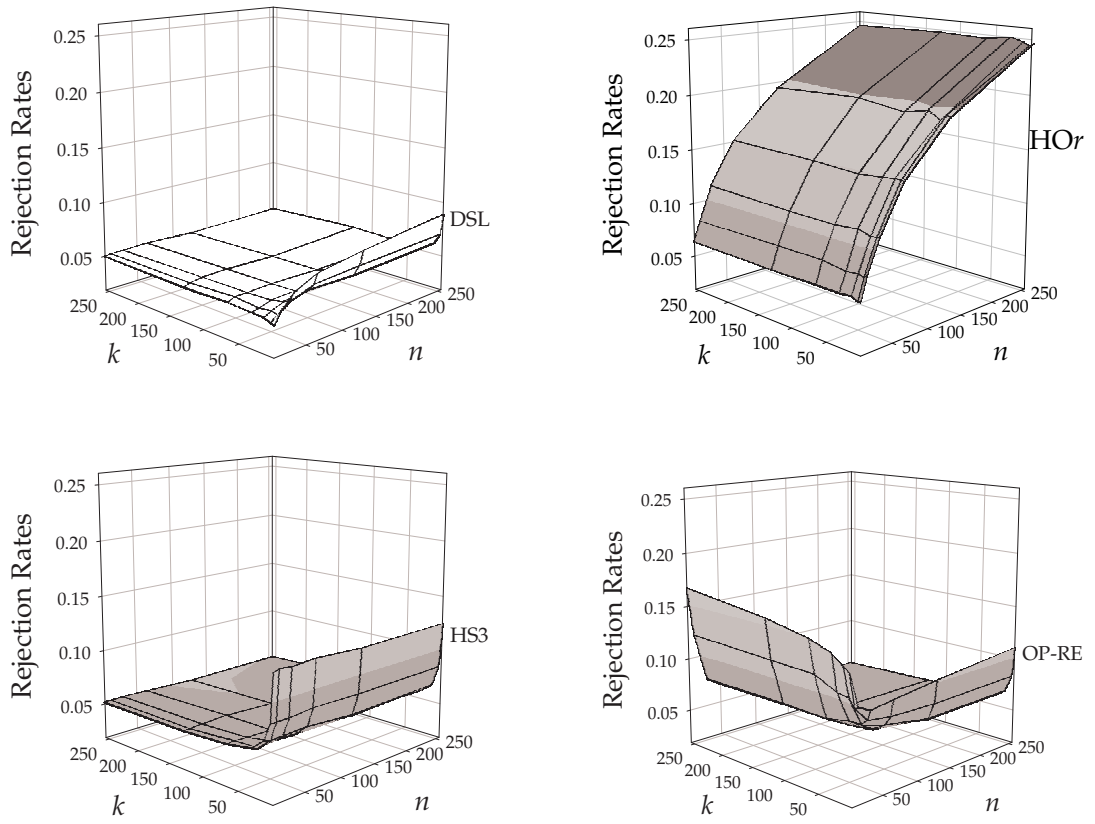

Figure 8.14 Rejection rates for testing the mean effects size in $\mathfrak{S}_3$ by n and k, $\mu_\rho = 0$, $\alpha = .05$.

of approaches is depicted that represents prototypical trends of the results for rejection rates in $\mathfrak{S}_3$ when the null hypothesis is true.

The upper left panel shows that the DSL approach leads to rejection rates corresponding to the nominal α for most of the values of k and n, the only exception is a slight elevation of rejection rates for k less than 16. Nonetheless, the rejection rates for DSL are not very high in any region of the n and k combinations under investigation. The lower left panel shows that HS3 (and HS4 which performs equally well) yields inflated rejection rates for small k invariably across values of n. Notwithstanding these elevated Type I errors, the overall performance of this approach seems acceptable here. OP-RE in the lower right panel shows too high rejection rates for a small number of studies and also for values of small k. Although this approach suffers from inadequate performance in estimating the mean effect size, the rejection rates in this situation do not seem unacceptable for moderate values of n and k.

In marked contrast to these results, *all* other other approaches (HOr, HOT, HOd, RR, HS1, HS2, and OP), for which HOr is depicted in the upper right panel as a representative, show a totally different trend across values of n. The rejection rates steadily increase with higher values of n but show no variation across values of k. In effect, the performance of these tests in $\mathfrak{S}_3$ becomes worse the higher the n of the studies. This demonstrates that the standard errors of

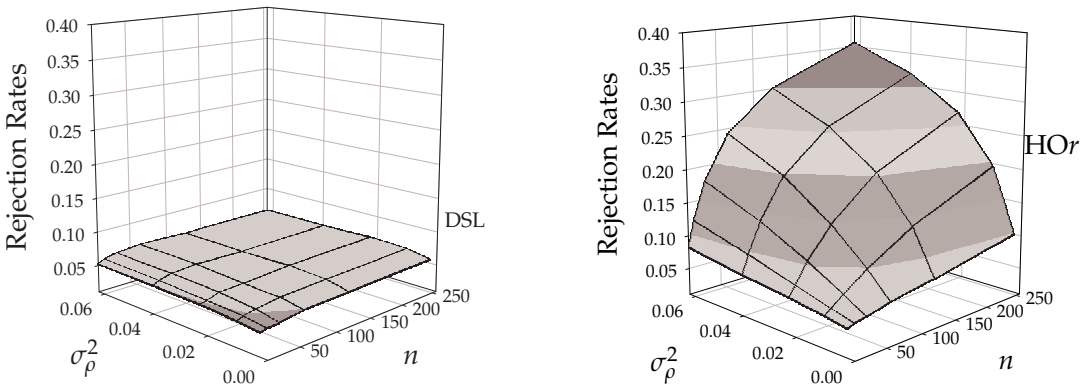

Figure 8.15 Rejection rates for testing the mean effects size in $\mathfrak{S}_3$ by n and σ^2_ρ, $\mu_\rho = 0$, $\alpha = .05$.

these approaches are too small in this heterogeneous situation, not reflecting variation with respect to σ^2_ρ.

In addition to the rejection rates getting higher with n, Figure 8.15 shows that rejection rates are also higher for larger values of σ^2_ρ — at least for HOr and the other fixed effects approaches. The variation of the rejection rates across values of n and σ^2_ρ is depicted in Figure 8.15.

For comparison, the rejection rates for the DSL approach are shown in the left panel and the results for the fixed effects approaches as represented by HOr in the right panel. Evidently, the rejection rates of the HOr approach quickly become far too large even for moderate values of n and σ^2_ρ. Because this approach did not show remarkable bias across levels of the design variables, this can be interpreted as an effect of the standard error estimates. In marked contrast, DSL shows a very good performance across values of σ^2_ρ. There are no elevations of the rejection rate surface in the left panel of Figure 8.15. Taking further into account that DSL also showed rejection rates close to nominal α, apart from in cases of very low n, it is certainly the best approach of those under consideration for testing the mean effect size in $\mathfrak{S}_3$-type situations.

Next, the results for the case $\mu_\rho \neq 0$ will be presented to assess the power of the approaches in $\mathfrak{S}_3$. The main findings are illustrated in an array of graphs in Figure 8.16.

In this figure, only two approaches are depicted that illustrate the different results for fixed versus random effects approaches. The shaded surface in the three panels portrays the slightly more powerful rejection rates for the fixed effects approaches (e.g., HOr). However, the differences to the random effects approach (e.g., DSL) — shown as a white surface lying underneath but close to the one of the FE approaches — are quite small across all design dimensions. Altogether, the figures show only minor differences in power *between* the approaches. With regard to the levels of the design variables, it is noteworthy that different variances in the universe of studies σ^2_ρ do not have a strong effect on testing the mean effect size. Although there is indeed a drop in rejection rates in the upper right panel of Figure 8.16, the effect is not strong for both types of

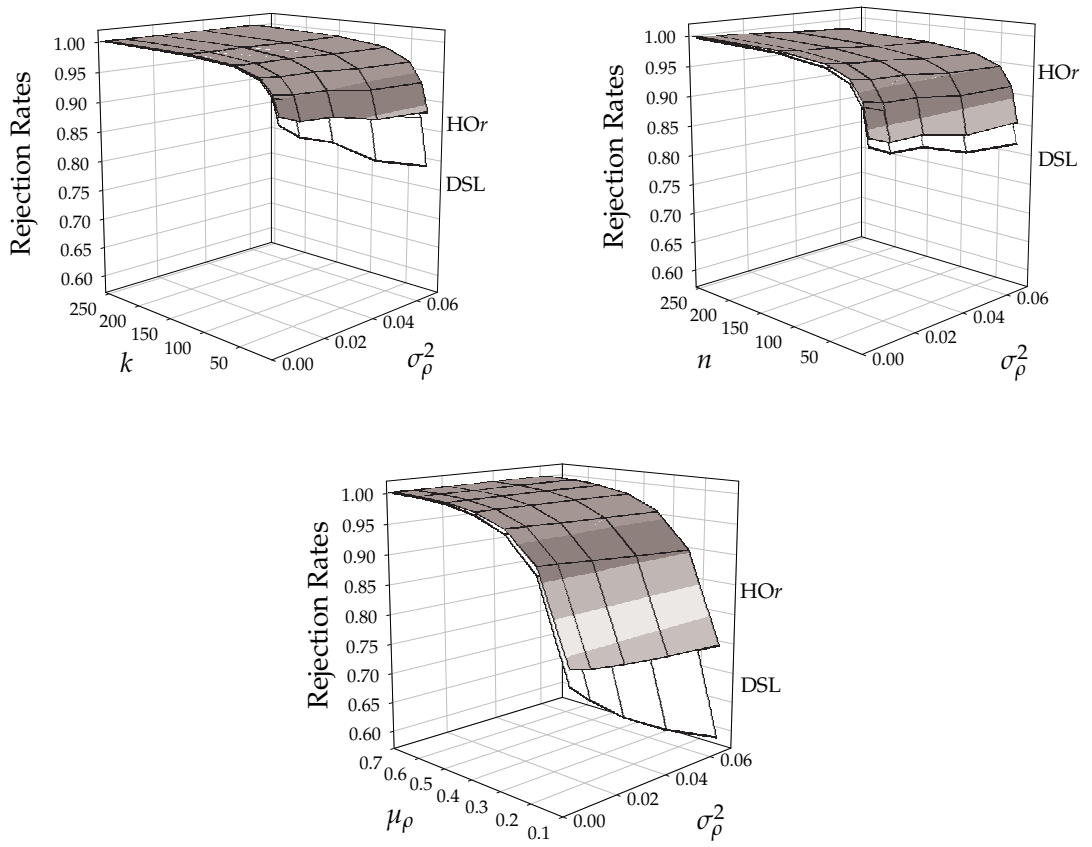

Figure 8.16 Rejection rates for testing the mean effect size in $\mathfrak{S}_3$, $\mu_\rho \neq 0$, $\alpha = .05$.

approaches. In contrast, low levels of k, n and small effects are predictors for low levels of power in $\mathfrak{S}_3$ as well.

With reference to the absolute values shown in the figures, it must again be emphasized that the results shown are always aggregates across the other design dimensions. For example, the DSL approach does not always attain a rejection rate of at least .65 as the array of graphics may suggest at first glance. In a worst-case-scenario of $n = 8$, $k = 4$, $\mu_\rho = .10$, and $\sigma_\rho^2 = .0625$ the estimator only shows a minimum rejection rate for the levels of the design variables of .1353, thereby highlighting the point that the figures are intended for comparison of approaches only, as outlined in the introduction of this chapter.

In sum, the results of the tests for the approaches are best when their basic model assumptions with respect to the fixed versus random effects model of meta-analysis are met (see also Hedges & Vevea, 1998). Differences in $\mathfrak{S}_1$ and $\mathfrak{S}_2$ do not seem to be very large between the approaches, but in $\mathfrak{S}_3$ — when μ_ρ is zero — there are tremendous differences in rejection rates. Hence, in a state of ignorance about the true situation, the potential loss in power caused by applying a random effects approach seems to be justifiable. In light of the errors potentially committed by applying fixed effects models in $\mathfrak{S}_3$ it seems advisable to accept slightly lower power levels.

8.4 CONFIDENCE INTERVALS

Confidence intervals for the approaches will be evaluated with respect to the rate to which they cover the universe parameter the estimators are supposed to estimate. These rates will be labeled as *coverage rates* in the following. They are often considered as a rather important aspect of the quality of meta-analytical approaches. This is evidenced, for example, by the fact that Brockwell and Gordon (2001) based their empirical comparison of approaches (fixed-effects method for log odds ratios, DSL, and a conditionally random effects procedure) almost exclusively on the results for coverage rates and interval widths. Since high coverage rates may be achieved by unduly large confidence interval widths, the mean interval widths will also be presented along with the coverage rates to establish a better foundation for evaluation. The coverage rates were computed as proportions over 10,000 iterations. The confidence limits are also aggregates over iterations so that they need not be exactly symmetrical about the mean effect sizes. Information given for the widths of intervals was computed from these confidence limits. In all cases, only 95%-confidence limits were investigated.

Coverage Rates and Interval Widths in $\mathfrak{S}_1$ Overall statistics for the coverage rates and 95%-confidence interval widths in $\mathfrak{S}_1$ are presented in Table 8.11. They will again be complemented by some graphical representations of the approaches' performance across levels of design variables, after a short discussion of the overall findings.

Of all the approaches, HOT reaches the highest coverage rates for a 95%-confidence interval. All other approaches show coverage rates lower than the standard of .95. In comparison to the overall interval widths of the other approaches, however, HOT also shows larger values. Hence, the high coverage rates may be obtained by virtue of larger interval widths. A second approach with rather good performance is OP. In this case however, relatively good coverage rates are not coupled with high interval widths.

The overall coverage rates for HS3, HS4, and OP-RE, for example, shown in Table 8.11 are too low to be acceptable. Interval widths are not simultaneously very small and minimum coverages show that these approaches show unacceptable performance at least in same regions of the design.

There are several determinants of the coverage rates of the approaches, so that differences between approaches are not easily interpretable. One possible reason for coverage rates being lower than expected is the bias of the estimators. For example, HO*r* and HOT are subject to exactly the same procedures for construction of the intervals, the only difference between these estimators is the correction of the estimator proposed by Hotelling (1953). This makes it possible to trace the reason for the lower coverage rates of HO*r* back to the bias in the estimator because the corrected version HOT shows appropriate rates. An analogous comparison is also possible for HS1 and OP. The standard errors are computed for OP the same way as for HS1 (compare Equations 5.7

Table 8.11 Coverage Rates and Confidence Interval Widths in $\mathfrak{S}_1$

	Statistic				
Approach	Max.	Mean	Median	Min.	SD
HO*r*	.9552	.9225	.9467	.1714	.0820
	.8244	.1208	.0702	.0029	.1383
HOT	.9715	.9548	.9530	.9449	.0059
	.8244	.1228	.0726	.0029	.1404
HO*d*	.9534	.8911	.9276	.0187	.1118
	.6830	.1070	.0617	.0023	.1182
HS1	.9545	.9202	.9446	.1877	.0801
	.7136	.1142	.0700	.0029	.1242
HS2	.9549	.9174	.9407	.2892	.0732
	.6436	.1105	.0707	.0029	.1159
HS3	.9521	.8941	.9214	.3035	.0713
	.6005	.1056	.0689	.0029	.1086
HS4	.9706	.8946	.9227	.1961	.0799
	.6657	.1090	.0677	.0029	.1166
OP	.9539	.9360	.9438	.8363	.0198
	.7095	.1125	.0667	.0029	.1223
OP-RE	.9584	.6525	.7786	.0001	.3113
	.8574	.1270	.0766	.0030	.1440
DSL	.9674	.9317	.9550	.1742	.0823
	.9098	.1301	.0729	.0030	.1523

Note. The total number of values described by these statistics is 420. Statistics for coverage rates are given in the first row of each approach and statistics for the widths of the confidence intervals in the second row.

and 5.11) but estimators differ. Hence, the benefit of an estimator's small bias is recognizable in the given context too.

In addition to the potential bias of an estimator, differences in standard error computations also contribute to the differences in rates and widths of the intervals. However, standard errors are not readily comparable between all estimators, except for the case of HS1 to HS4. Here, the mean effect size is exactly the same for all variants but standard errors differ. Amongst the HS variants, HS1 and HS2 show better performance than HS3 and HS4 in $\mathfrak{S}_1$. Recall again that HS3 was proposed to show good performance both in homogeneous and heterogeneous situations. The comparison between HOT and DSL shows that smaller interval widths do not necessarily lead to lower coverage rates, though this is a strong tendency in the results. The reason for better performance of HOT on both accounts is its smaller bias. This fact again underscores the importance of small bias in estimators, even when differences in accuracy between estimators appeared to be relatively small. The consideration of the minima of coverage rates also strongly emphasizes the importance

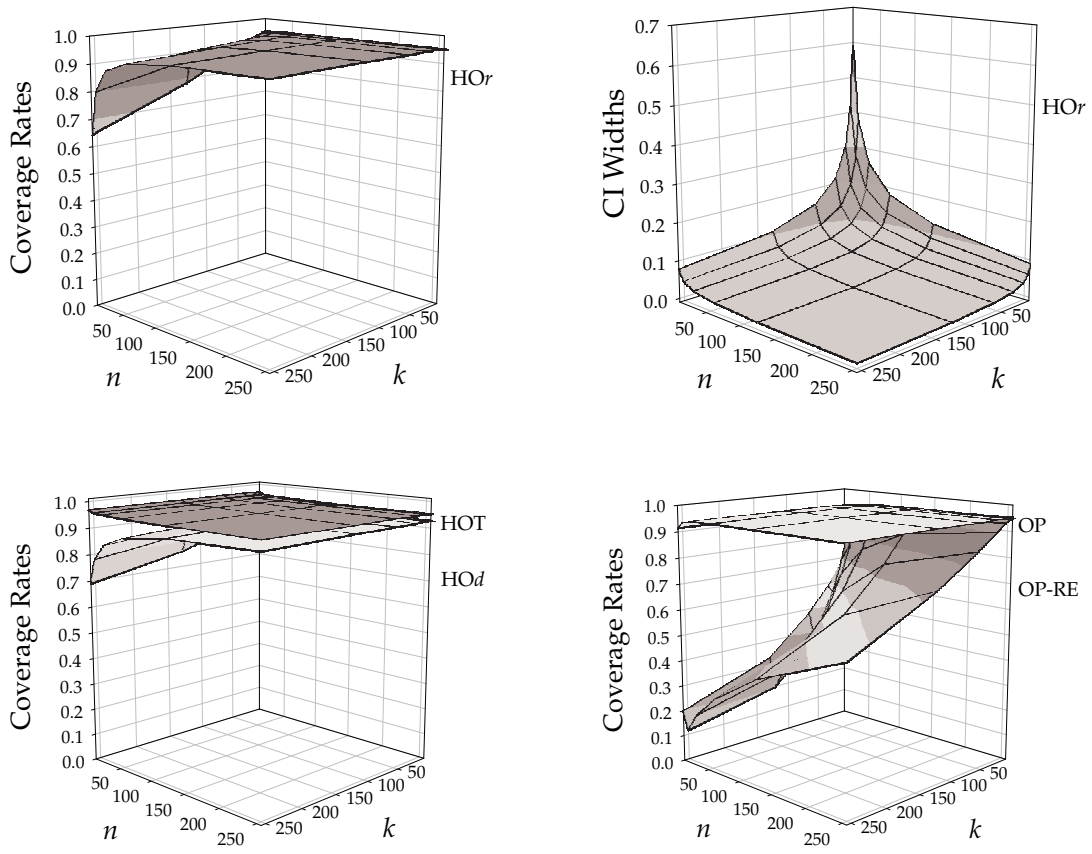

Figure 8.17 Coverage rates and confidence interval widths in $\mathfrak{S}_1$ by k and n.

of small biases. Remarkably, the estimators having shown the smallest bias exhibit, even in minimum across *all* situations considered in the design, very good (HOT) or fairly good (OP) coverage rates.

To illustrate the constellations of levels of design variables leading to poor performance of some approaches, a series of graphs is presented in Figure 8.17. They depict the dependencies of coverages and interval widths on n and k.

The first set of graphs in Figure 8.17 (upper left and both lower panels) shows the coverage rates for a selection of approaches and a separate graph for the interval widths (upper right panel). As in the figures presented in previous sections, approaches are omitted that show very similar surfaces in comparison to the ones depicted and might not be discriminated even when included in the graphs. The selection of approaches is chosen to illustrate the main trends: HO*r* also represents HS1 to HS4, RR, and DSL. HOT, HO*d*, OP, and OP-RE are depicted separately. Although HS3 and HS4 show smaller coverage rates as shown in Table 8.11, they can also be subsumed under HO*r* because of their similarity in trends. For the confidence interval widths only one graph is shown because all approaches nearly have the same surfaces. Interestingly, interval widths do not depend differently on levels of n and k for all approaches. Interval widths grow large for all approaches only in cases of both small sample sizes and a small number of studies. The shrinkage in widths seems to be

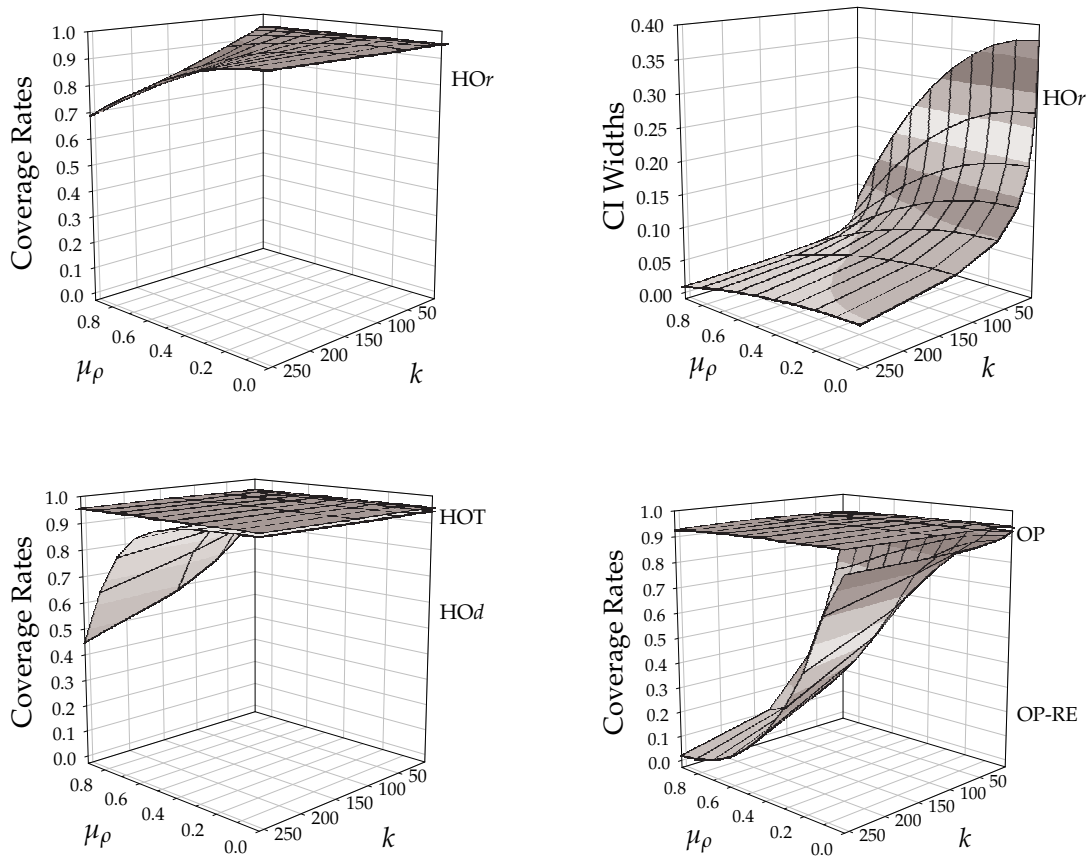

Figure 8.18 Coverage rates and confidence interval widths in $\mathfrak{S}_1$ by k and μ_ρ.

approximately the same when holding n constant and focusing on a growing number of studies and vice versa.

As is evident from Figure 8.17, small n in combination with large k leads to diminishing coverage rates for all approaches except OP and HOT. This again suggest that biases are the cause for poor performance because standard errors are smallest for large k — as is also evidenced by the smaller interval widths — and biases are largest with small n (see Section 8.2.1.1). Cautions are raised by this finding against the use confidence intervals of most procedures to construct confidence intervals in $\mathfrak{S}_1$ when n is small and k is large. Nevertheless, the excellent coverage rates for HOT and OP as evidenced by their flat surface in Figure 8.17 makes them first choice in $\mathfrak{S}_1$ not only for the purpose of estimating the mean effect size but also for the construction of confidence intervals.

For further insight into the dependencies of coverage rates and interval widths on levels of design variables, Figure 8.18 provides the results for levels of k and μ_ρ. It can again be expected that the results of the coverage rates mirror performance of the estimators' bias.

As before, interval widths as shown in the upper right panel in Figure 8.18 do not markedly differ between approaches. Hence, one graph seems to suf-

fice to portray all relevant information for a comparison of approaches. The gradient of the confidence interval width surface is again in agreement with expectations from statistical theory. Intervals are smaller for large k and — for the case of correlation coefficients as effect sizes — widths become smaller with larger μ_ρ. It might be surprising that this should also be the case for Fisher-z based approaches because it was highlighted in several sections that the standard errors of the mean effect size estimators for these approaches do *not* depend on the parameter itself (see, e.g., Equation 5.2 on page 58). This is true, however, in z-space and the results depicted in the figures are all in r-space. A certain interval width of .09 in z-space — which approximately results for the case $n = 32$ and $k = 16$ — corresponds to an interval width of approximately .04 at a mean effect size level of .90 and to a width of .12 at the level of .60. Hence, the change of spaces from z to r in the present case makes the shape of the surface appear reasonable also for Fisher-z based approaches.

The upper left and lower panels in Figure 8.18 depict the coverage rates for the approaches. As can be seen for most approaches, lower coverage rates result for combinations of vary large μ_ρ and *high* k. For OP-RE this phenomenon is certainly due to its large bias but for HOr and HOd biases were shown to be quite small. Especially for HOd rather low coverage rates are shown in the critical design region.

The performance of OP and HOT again stands in marked contrast to those of other approaches. The coverage rates of both approaches is again depicted as a surface at the level of approximately .95. Almost the same picture emerges in the final set of graphs in Figure 8.19. As in the previous figures, the coverage rates and interval widths are shown for combinations of n and μ_ρ in three panels and the upper right one shows the interval widths.

The interval widths basically show the same trends for all approaches and a surface is shown in Figure 8.19 which very much resembles that in the previous figure, only shown from a different angle of view. The widths of confidence intervals are largest for all approaches in combinations of small n and small μ_ρ, as would be expected. The coverage rates decline for combinations of very large μ_ρ and rather small n. This shows again the deleterious effect of small interval widths and large biases.

Coverage Rates and Interval Widths in $\mathfrak{S}_2$ Next, the heterogeneous situation with two different values in the universe of studies is treated. First of all, it should again be noted that the coverages are evaluated with respect to the parameters the estimators are supposed to estimate, just as in Section 8.2.1.2. This is important for a comparative evaluation of approaches in this context. That is, the universe values to be covered by the confidence limits are *different* for the Fisher-z based and r-based approaches with differences in universe parameters (μ_ρ vs. $\mu_{\rho z}$) being larger, the higher the difference is between the two universe values of ρ (i.e., $\Delta\rho$). Furthermore, as in Section 8.2.1.2, the coverage rates for HOd in $\mathfrak{S}_2$ were evaluated with respect to μ_ρ and not to $\mu_{\rho d}$. For an explanation as to why this is the case, see Section 5.5.

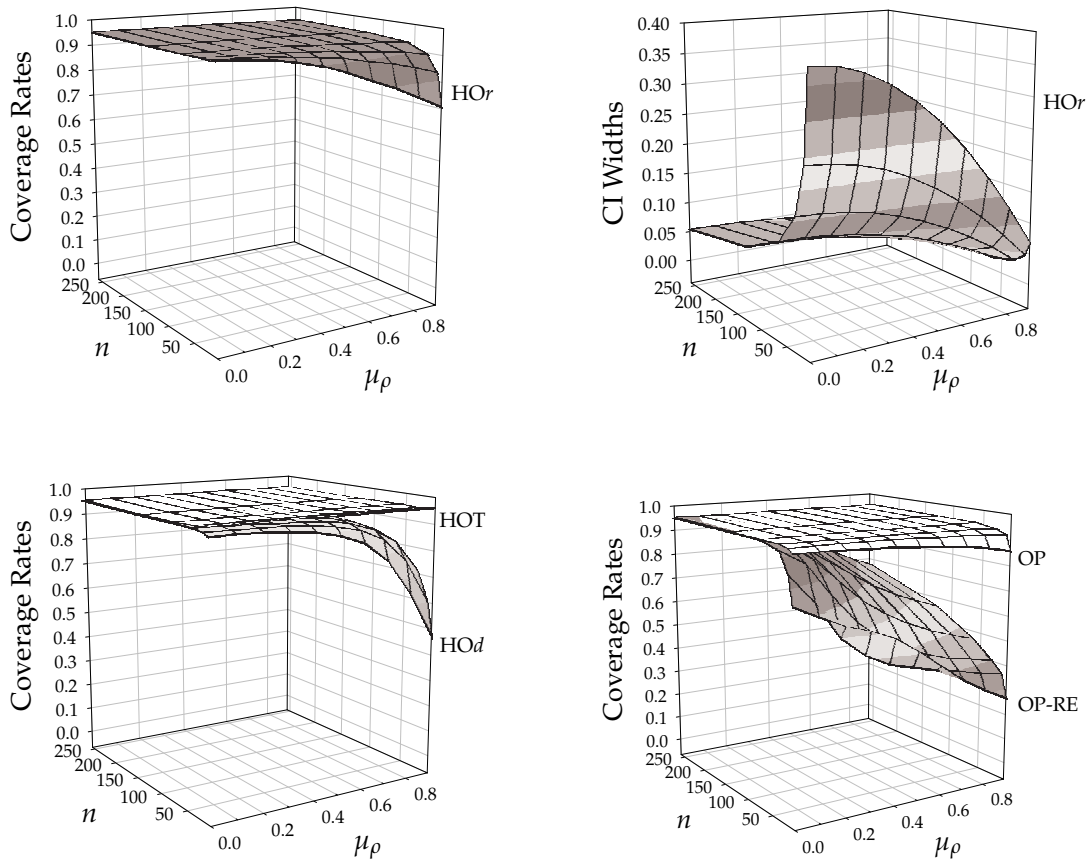

Figure 8.19 Coverage rates and confidence interval widths in $\mathfrak{S}_1$ by n and μ_ρ.

The overall results on coverage rates and confidence interval widths for 95%-Intervals are first presented. Table 8.12 shows descriptive statistics for a comparative evaluation of the approaches.

The values presented in Table 8.12 suggest that the interval widths are considerably larger for all approaches, not only random effects approaches. Nevertheless, for the latter the intervals are approximately twice as wide as for the fixed effects approaches. In the extreme, this leads to intervals larger than one and to coverage rates of up to one in maximum (e.g., DSL).

The results shown in Table 8.12 again demonstrate that OP and particularly HOT approximately attain the desired coverage rates without having excessively large interval widths. Additionally, even the minimum values for the coverage rates indicate a very good performance of the approaches in all cases under investigation. In contrast, the minimum values for all other approaches suggest that there are situations in which they perform very poorly. Other fixed effects approaches like HS1 and HS2, however, also attain good mean overall coverage rates without excessively large interval widths, but the minima for these approaches indicate bad performance at least in some cases under consideration. The highest mean coverages are shown for DSL, but this

Table 8.12 Coverage Rates and Confidence Interval Widths in $\mathfrak{S}_2$

	Statistic				
Approach	Max.	Mean	Median	Min.	SD
HO*r*	.9559	.9266	.9464	.2237	.0639
	.8223	.1219	.0760	.0041	.1321
HOT	.9714	.9544	.9529	.9247	.0057
	.8226	.1244	.0776	.0041	.1353
HO*d*	.9751	.8883	.9311	.0000	.1385
	.6508	.1160	.0764	.0035	.1170
HS1	.9751	.9345	.9470	.2595	.0514
	.7130	.1213	.0790	.0043	.1229
HS2	.9555	.9238	.9410	.3468	.0510
	.6415	.1144	.0758	.0044	.1120
HS3	1	.9718	.9958	.4515	.0486
	.9494	.2060	.1500	.0131	.1690
HS4	1	.9738	.9961	.3653	.0479
	.9993	.2104	.1514	.0130	.1745
OP	.9743	.9407	.9457	.8303	.0200
	.7088	.1198	.0785	.0043	.1209
OP-RE	1	.8627	.9714	.0012	.2249
	1.2930	.2538	.1780	.0149	.2219
DSL	1	.9784	.9971	.2611	.0516
	1.1109	.2280	.1556	.0127	.2034

Note. The total number of values described by these statistics is 1890. Statistics for 95% confidence intervals are given in the first row of each approach and statistics for the width of the confidence intervals in the second row.

is easily explained by the fact that the interval widths are unduly large and should therefore not lead to an overly positive evaluation.

The set of graphs provided in Figure 8.20 illustrates the trends of the coverages and interval widths in $\mathfrak{S}_2$. Again, only a small selection of approaches is depicted to show overall trends in cases where some classes of approaches do not differ markedly in performance. As representatives, DSL and HO*r* are given. DSL stands for the random effects approaches as well as HS3 and HS4, whereas HO*r* roughly represents all other approaches except for OP and HOT. The latter two both show flat surfaces at a height of approximately .95 across all levels of design variables for the coverage rates and interval width surfaces similar to those of HO*r* shown in Figure 8.20. Hence, these two approaches perform uniformly best in all cases but, again, it should be noted that HOT does so with respect to $\mu_{\rho z}$ and OP with respect to μ_{ρ}. Taking this into account, OP seems to be the approach of choice in the present context.

The upper panels in Figure 8.20 show coverage rates and interval widths by n and k. As expected, DSL shows larger widths of intervals than HO*r* and also

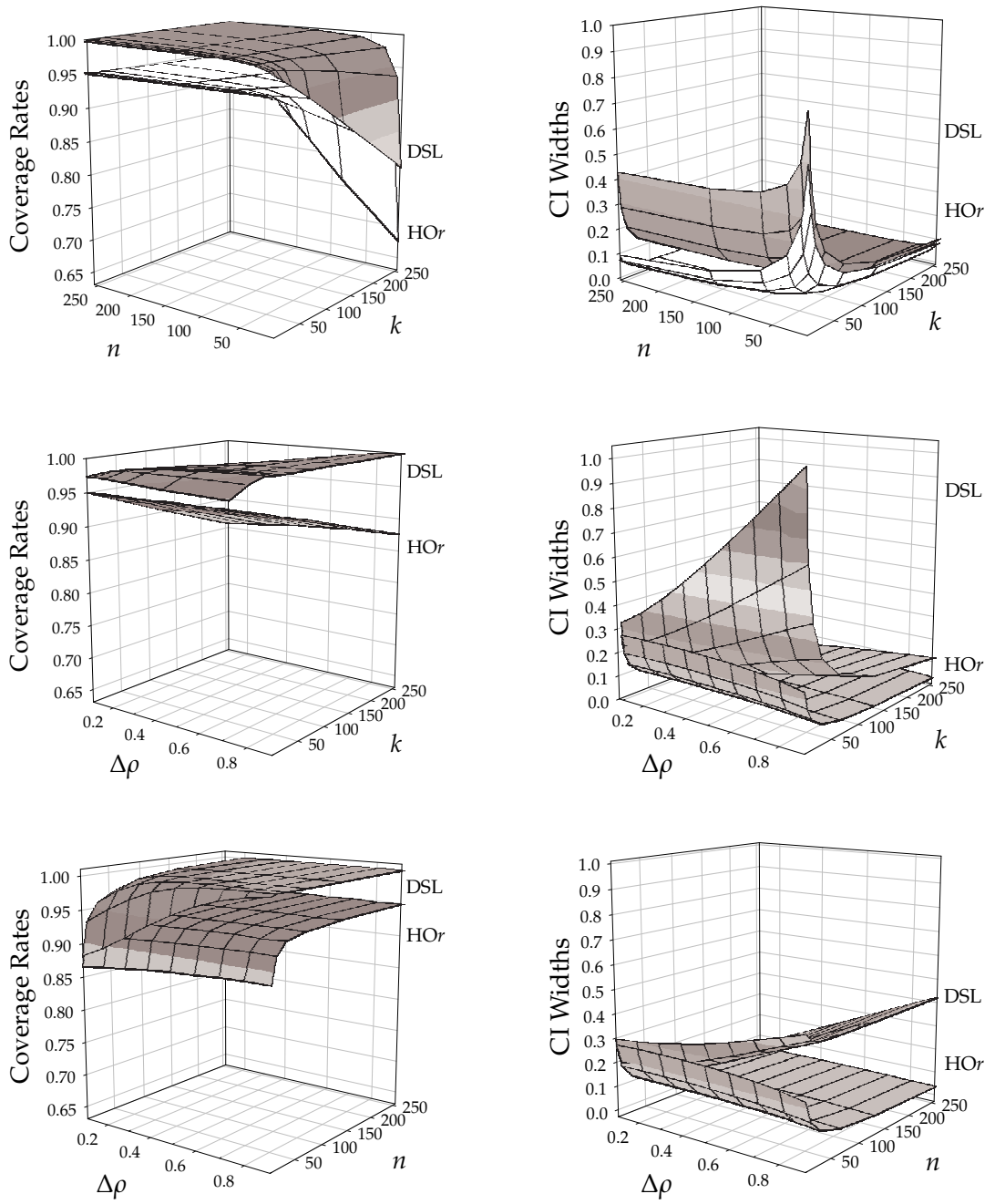

Figure 8.20 Coverage rates and confidence interval widths in $\mathfrak{S}_2$.

higher coverage rates. The coverage rates for DSL actually approach a value of one even for small values of n and k. Thus, they are higher than can be expected for the construction of 95% confidence intervals. This also stands in contrast to the coverages of HOr attaining a value of .95 in limit. However, the effect of bias emerges again in extreme combinations of high k and small n for both approaches. As can be seen in the mid- and lower panels, DSL does react to different values of $\Delta\rho$ in contrast to HOr which performs almost equally

Table 8.13 Coverage Rates and Confidence Interval Widths in $\mathfrak{S}_3$

	Statistic				
Approach	Max.	Mean	Median	Min.	SD
HO*d*	.9418	.7014	.7628	.0320	.2039
	.6942	.1179	.0773	.0024	.1223
HS1	.9507	.7215	.7842	.1897	.1944
	.7140	.1235	.0805	.0029	.1272
HS2	.9361	.7106	.7731	.1992	.1876
	.6429	.1174	.0783	.0032	.1166
HS3	.9532	.9044	.9262	.4588	.0542
	.6863	.1700	.1299	.0127	.1280
HS4	.9707	.9082	.9322	.4002	.0566
	.7573	.1747	.1327	.0126	.1355
OP	.9425	.7246	.7915	.1913	.1922
	.7100	.1221	.0800	.0029	.1254
OP-RE	.9970	.8059	.8696	.0024	.1953
	.9283	.2041	.1522	.0150	.1637

Note. The total number of values described by these statistics is 1848. Statistics for coverage rates are given in the first row of each approach and statistics for the widths of the confidence intervals in the second row.

across all levels of $\Delta\rho$. As a fixed effects approach, HO*r* therefore does not reflect the additional variability introduced by larger universe parameter differences. However, the results also suggest that DSL does overreact on these differences in the sense of overestimating the heterogeneity variance. An examination of this impression will not be presented here but postponed to an in-depth assessment of the estimators of heterogeneity variance in Section 8.6.

Coverage Rates and Interval Widths in $\mathfrak{S}_3$ For a full evaluation of all approaches in $\mathfrak{S}_3$ it would have been necessary to implement expected values of the beta distribution in z-space (i.e., μ_ζ) as a standard for comparison for the approaches that use the Fisher-z transformation. As already noted, this was considered not to be feasible. Accordingly, the following presentation has to be restricted to approaches for which μ_ρ can be used as a standard for comparison. As before, a table of overall results is presented for a comparison of the performance of the approaches. Table 8.13 gives descriptive results for the available approaches in this situation.

First, it is noted that none of the approaches yields the desired coverage rate of .95 in mean or median. Somewhat surprisingly, HS3 and HS4 stand out here with best performance amongst the approaches under consideration. Although these approaches also show higher interval widths in relation to the fixed effects approaches, they attain better coverage rates than OP-RE with smaller mean confidence intervals. In contrast to the previous situations, OP does *not* show acceptable performance. Mean and median coverage rates are

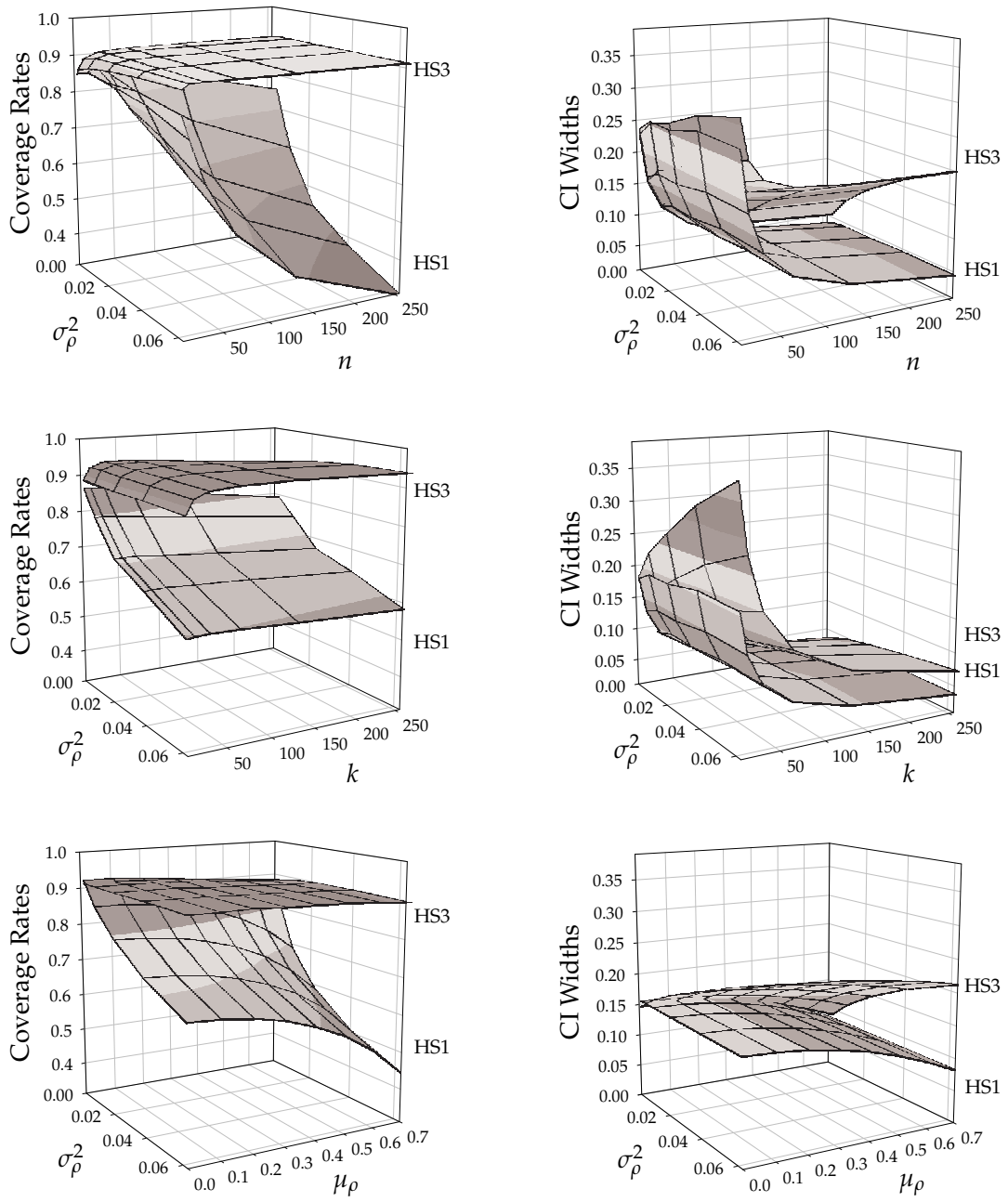

Figure 8.21 Coverage rates and confidence interval widths in $\mathfrak{S}_3$.

too small and the minimum coverage rate also shows that there are cases of very bad performance for this approach. This is astonishing given the estimator's brilliant performance with respect to mean effect size estimation. Hence, the additional variability in the universe of effect sizes is not adequately reflected in the computation of the standard errors in this FE approach, leading to unacceptable performance in the construction of confidence intervals.

The information presented in Table 8.13 shows that there are conditions for all approaches in which they perform rather poorly. The series of graphs in Figure 8.21 shows the results for some combinations of the design variables.

It can be gathered from the graphs in Figure 8.21 that HS3 generally retains its coverage rates across the levels of the design variables whereas the results for HS1 vary strongly. Here, HS3 also represents the results for HS4, and HS1 is depicted to stand for all other approaches available. The reason for this finding lies in the adjustment of interval widths in HS3 for high values of σ_ρ^2. As can be seen in the lower right panel, for example, interval widths are becoming larger the higher the heterogeneity variance (σ_ρ^2) is. This adequately reflects additional uncertainty in estimating the limits of an interval which covers the parameter of interest with a probability of .95. In contrast, HS1 (and all other FE approaches) evidences much more stable confidence interval widths for all values of σ_ρ^2. The minimum values for coverage rates of about .40 are only attained by HS3 and HS4 in very extreme cases of $n = 8$, $k = 256$ and large μ_ρ in combination with large σ_ρ^2. The coverage rates rapidly increase with growing n in this case and already show a value of .75 for $n = 16$ in the same case.

In summing up the results on coverage rates and interval widths, it can be stated that in situations $\mathfrak{S}_1$, $\mathfrak{S}_2$, and $\mathfrak{S}_3$ both HOT and OP showed very good performance in absolute terms and in comparison to other approaches. Since HOT is a Fisher-z based approach, OP should be preferred at least in situations of type $\mathfrak{S}_2$. The picture of results is different in $\mathfrak{S}_3$. Although HOT is not available for a comparative evaluation, as an FE approach it is not suspected to show good performance, especially not when furthermore taking the use of the Fisher-z transformation in this approach into account. OP showed disappointing performance in $\mathfrak{S}_3$. The only well performing approaches emerged to be HS3 and HS4. Although they did not reach coverage rates as prescribed by the $1 - \alpha$ level of the confidence intervals, they appeared to be best amongst the approaches under consideration. Hence, for different situations varying recommendations can be given for the purpose of constructing confidence intervals.

8.5 HOMOGENEITY TESTS

Tests of the homogeneity of effect sizes play a central role in meta-analysis and are conducted for various purposes (see Chapter 4). The present section is devoted to an evaluation of these tests in the three situations of the Monte Carlo study. Note that not all approaches and refinements provide distinct tests so that only tests based on the Q-statistic as described in Chapter 5 are available. The subsections are divided into standard methods, that is, the Q-test for the various approaches on the one hand, and the HS methods on the other. Since Hunter and Schmidt (1990; Hunter et al., 1982) also provide a standard Q-test in addition to the tests unique to their approach, HS appears in both sections. The special tests that Hunter and Schmidt provide are only widespread in I/O psychology and have a completely different statistical rationale than the Q-

Table 8.14 Rejection Rates for the Q-Test in $\mathfrak{S}_1$

	Statistic				
Approach	Max.	Mean	Median	Min.	SD
HO*r*	.0623	.0492	.0499	.0095	.0058
	.0175	.0105	.0104	.0011	.0019
HO*d*	1	.3572	.2191	.0512	.3162
	1	.2357	.0808	.0098	.3080
HS	.9412	.0878	.0535	.0190	.1105
	.8853	.0337	.0115	.0011	.0873
OP-FE	1	.2018	.0859	.0005	.2524
	1	.1319	.0266	.0001	.2430

Note. The total number of values described by these statistics is 420. Proportion for tests at $\alpha = .05$ are given in the first row of each approach and for tests at $\alpha = .01$ in the second row.

test. For this reason and for greater focus on the peculiarities of results for these distinct procedures they will be separated from the Q-tests.

8.5.1 Homogeneity Tests Based on the Q-Statistic

For the results on the homogeneity tests, situations $\mathfrak{S}_1$ versus $\mathfrak{S}_2$ and $\mathfrak{S}_3$ provide the two most relevant classes of situations. $\mathfrak{S}_1$ is the homogeneous case and $\mathfrak{S}_2$ as well as $\mathfrak{S}_3$ both represent different heterogeneous cases. The following subsections are structured in correspondence with this distinction, where $\mathfrak{S}_1$ is used to investigate Type I error rates and the heterogeneous situations are relevant to examine the power of the tests based on the Q-statistic.

8.5.1.1 Homogeneous Situation $\mathfrak{S}_1$: Type I Errors The first examination of results is concerned with overall performance of the proposed tests. The results for tests both on a significance level $\alpha = .05$ as well as $\alpha = .01$ are presented in Table 8.14.

As is shown in Table 8.14, only HO*r* and with strong reservations also HS approximately reach the desired significance levels. All other approaches show unacceptably large Type I error rates. Although HS performs better than HO*d* and OP-FE, the minima, maxima, and standard deviations indicate very large variability of test results in comparison to the Fisher-z-based approach HO*r*. To investigate this variability, the rejection rates are depicted in the following series of graphs by combinations of design variables. Figure 8.22 permits an inspection of the different surfaces across all the dimensions of the design.

Approaches are again clustered for better visibility of the surfaces. OP-FE is omitted for its general poor performance. The upper left panel illustrates that excessive rejection rates occur for combinations of large k and small n for HO*d*. HS also shows its largest values in this case. In marked contrast to these findings, the upper right panel with a rescaled vertical axis indicates that HO*r*

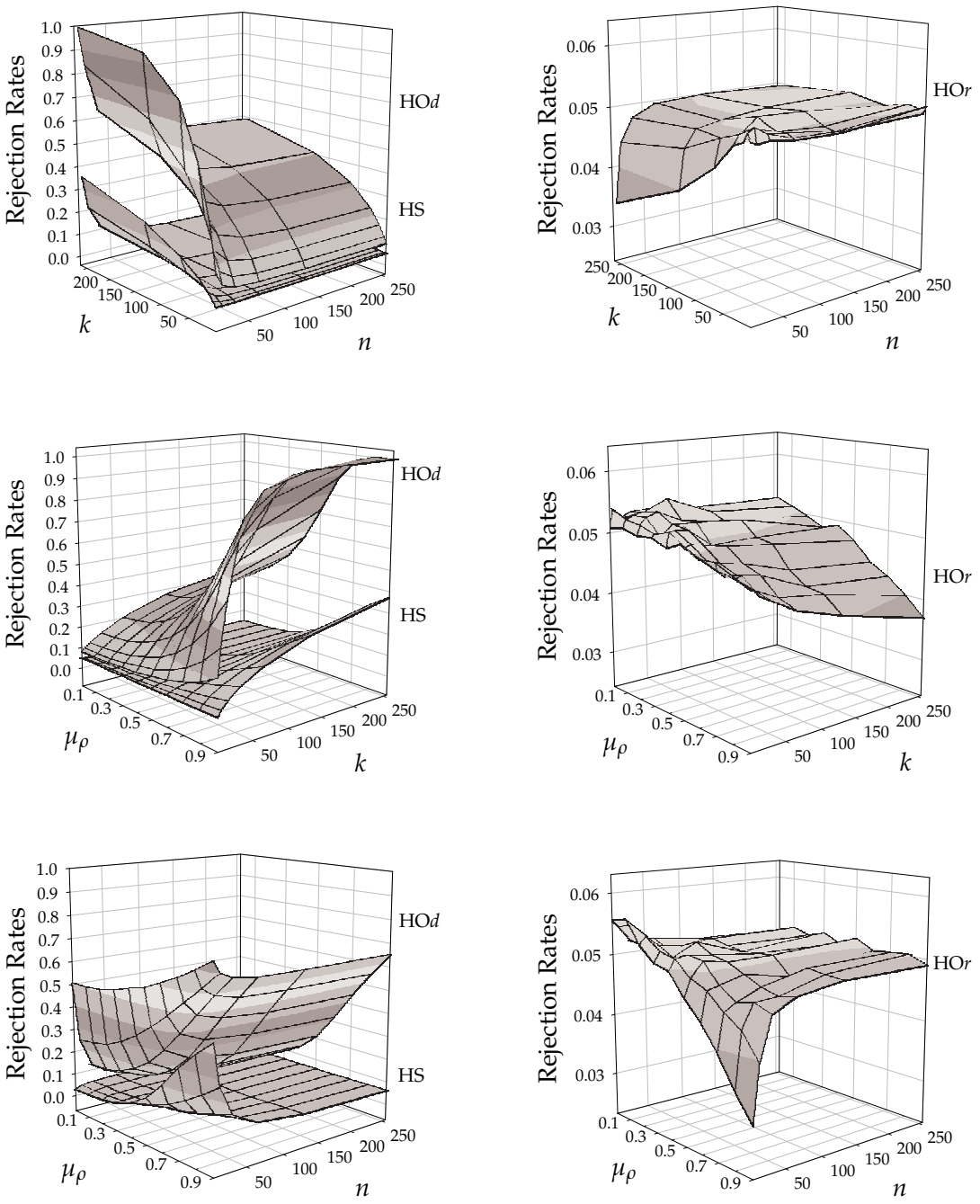

Figure 8.22 Rejection rates for the Q-test in $\mathfrak{S}_1$, $\alpha = .05$.

performs very well in $\mathfrak{S}_1$. Although HOr deviates from the nominal α in the same cases where HOd and HS perform worst, it actually shows *low* rejection rates indicating good performance when the null hypothesis is true as is the case in $\mathfrak{S}_1$.

The mid-panels in Figure 8.22 show a similar picture. Excessive rejection rates for HOd occur for large k and μ_ρ. HS also performs poorly in such cases, but HOr performs adequately in most situations. The same relative perfor-

mance is observed in the lower panels for the three approaches depicted. A worst-case scenario is given in the lower left panel for a combination of low sample sizes coupled with a high value for the universe parameter. Hence, it becomes clear that HO*d* performs most poorly overall for high values of μ_ρ when results are aggregated across values of k, small n seems to even exacerbate this problem.

The question arises how the distinct results of the approaches can be explained. In a Monte Carlo study based comparison of the HS and HO*r* approach, Alexander et al. (1989) showed similar differences between these approaches. They actually used a slightly different HS-estimator for μ_ρ in computing the Q-statistic that is equivalent to the one used in the present context with constant n for all studies. As an aside, in contrast to the present study they used different n for each study simulated. The fact that the results presented here agree with those reported by Alexander et al. lends support to the claim that a constant n for all studies does not lead to limitations in interpretation in the given context. The same is true in comparison to Field's study (2001), which also used varying n within studies and reported similar results. With reference to Snedecor and Cochran (1967), Alexander et al. (1989) attributed the observed differences to the nonnormal distribution of the correlation coefficients. Applying this explanation to the present results can explain the high rates of HS for large values of μ_ρ, but does not readily explain the values reported for HO*d* being even more deviant from the nominal α-level. As pointed out in Sections 3.3 and 5.5, it is the transformation of r to d that may be the cause for intensification of the variability of the d values about the estimated mean effect size. Additionally, the weights used to compute Q also vary with d. They are smaller for higher d and thereby also introduce a further component that amplifies variability in values to be summed to the Q-statistic. All in all, the transformation of r to d results in homogeneity tests not suitable for application.

8.5.1.2 Heterogeneous Situations $\mathfrak{S}_2$ and $\mathfrak{S}_3$: Power The first heterogeneous situation in which rejection rates of the homogeneity tests will be examined is $\mathfrak{S}_2$. Results for the rejection rates by values of k and selected $\Delta\rho$ are presented in Table 8.15.

The results in Table 8.15 show relatively low rejection rates when $\Delta\rho$ is small for all approaches. As expected, rejection rates rise for higher values of k and $\Delta\rho$. From the findings in the previous subsection, it is expected that HO*d* will also show higher rejection rates in $\mathfrak{S}_2$. This is indeed the case but the high Type I error rates in $\mathfrak{S}_1$ should be kept in mind when evaluating the performance of HO*d*. A notable result shown in Table 8.15 is the relatively low power to detect small differences between ρ_1 and ρ_2. Even in a meta-analysis of 256 studies, the power to detect such effects is not impressively high. Moderate differences between universe effect sizes of .30 are also only detected with an appreciable number of studies (more than 16) for approaches with acceptable Type I error rates in the homogeneous situation. In a situation with a very small number of studies — potentially occurring in a meta-analysis when subgroups of studies

Table 8.15 Rejection Rates for the Q-Test by k and $\Delta\rho$ in $\mathfrak{S}_2$

k	$\Delta\rho$	HOr	HOd	HS	OP-FE
	.1	.2112	.2972	.2089	.2242
	.3	.5964	.6738	.5897	.6068
4	.5	.7858	.8539	.7766	.7911
	.7	.8994	.9485	.8890	.8910
	.9	.9736	.9930	.9628	.9442
	.1	.2560	.3781	.2606	.3021
	.3	.6667	.7655	.6655	.7141
8	.5	.8451	.9193	.8402	.8779
	.7	.9445	.9832	.9389	.9517
	.9	.9949	.9999	.9920	.9841
	.1	.3190	.4878	.3337	.4133
	.3	.7372	.8566	.7427	.8279
16	.5	.8999	.9695	.8993	.9505
	.7	.9775	.9978	.9766	.9897
	.9	.9999	1	.9997	.9992
	.1	.3922	.6167	.4199	.5473
	.3	.8025	.9335	.8136	.9230
32	.5	.9436	.9940	.9459	.9897
	.7	.9950	1	.9953	.9996
	.9	1	1	1	1
	.1	.4685	.7491	.5134	.6882
	.3	.8591	.9824	.8742	.9788
64	.5	.9743	.9998	.9775	.9996
	.7	.9998	1	.9999	1
	.9	1	1	1	1
	.1	.5426	.8639	.6055	.8239
	.3	.9055	.9984	.9227	.9973
128	.5	.9935	1	.9952	1
	.7	1	1	1	1
	.9	1	1	1	1
	.1	.6096	.9425	.6910	.9293
	.3	.9424	1	.9587	1
256	.5	.9995	1	.9998	1
	.7	1	1	1	1
	.9	1	1	1	1

Note. Proportion for tests at $\alpha = .05$.

are examined — the power is only acceptable for large differences between the universe parameters. This is the case for all of the approaches in Table 8.15. Nevertheless, the results shown in this table indicate a very similar overall performance of the approaches in $\mathfrak{S}_2$.

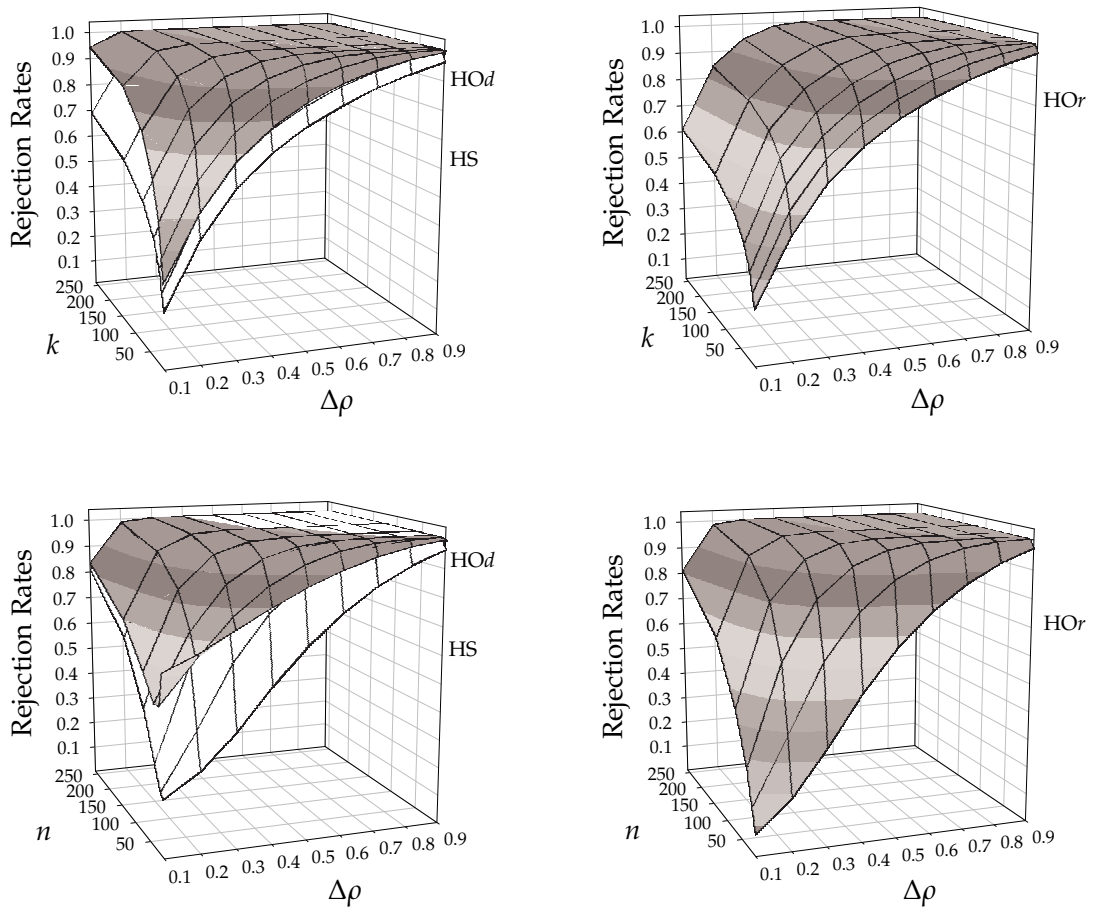

Figure 8.23 Rejection rates for the Q-test in $\mathfrak{S}_2$, $\alpha = .05$.

These results have to be qualified, however, by including the additional design variable n. The lower panels in Figure 8.23 show rejection rates across different values of n. The upper panels depict the results of Table 8.15 but values omitted from the table are added to the graphs. The lower panels indicate that rejection rates also depend on n. In general, the shapes of the surfaces are again quite similar, not favoring any of the approaches in particular. The results in $\mathfrak{S}_2$ show that medium effects sensu Cohen (1988, 1992) of .30 are only detected with acceptable power when n and k are at least 32. Whereas this may be considered a customary condition for n in most fields of correlational research, this is not the case for k. Small effects (.10) are hardly detected by the Q-test unless n and/or k are quite large.

In sum, for some constellations of the design variables' levels the probability to detect differences between universe parameters can be quite low. Although including many studies in a meta-analysis raises power, even a large number does not guarantee sufficient power. The present case can be interpreted as a situation arising from an unobserved dichotomous explanatory variable. Since it is not always the case that such variables can be observed, indications of their existence are of great interest to the meta-analyst. Because the use of

explanatory models is sometimes conditioned upon the results of homogeneity tests, the results point to cases in which such conditional procedures are problematic. Of course, the present examination is restricted to a two-point distribution in the universe of studies, and different results may emerge for more unobserved classes. The more general case of the homogeneity test performance with a continuous mixing distribution is therefore also of interest.

The rejection rates for the approaches in $\mathfrak{S}_3$ are shown in Table 8.16 for varying values of k and σ_ρ^2, and also in an array of graphs in Figure 8.24.

As was the case in $\mathfrak{S}_2$, rejection rates generally rise for higher values of k and σ_ρ^2. In contrast to $\mathfrak{S}_2$, a continuous distribution is given in the universe of studies and homogeneity tests are supposed to indicate variances of this distribution different from zero. As the results in Table 8.16 show, this universe variance in effect sizes is detected by the approaches only with acceptable rates when k is at least 16 and variances are large. Small variances are likely to go unrecognized even in meta-analyses with large k. Though HOd shows the highest power among the approaches under investigation, this comes at the cost of excessive rejection rates in $\mathfrak{S}_1$. Figure 8.24 provides an overview of changes in rejection rates for varying values of σ_ρ^2, n, and k.

The upper panels in 8.24 show that for k and σ_ρ^2 the rejection rates are only satisfactory when both values are relatively high. The mid-panels also indicate decreasing rejection rates for very small n and the lower panels show that these trends do not strongly depend on values of μ_ρ. Hence, almost irrespective of the size of μ_ρ in the universe of studies, an appreciable number of studies is needed to detect even moderate heterogeneity at a power level conventionally considered as acceptable. In sum, all tests show somewhat unsatisfactory rejection rates in $\mathfrak{S}_3$ and cannot safely be taken as indicants of heterogeneity under all configurations of the design variables.

Again, this result is quite important if the Q-test is considered as a decision-making device for the choice between fixed and random effects models as in the so-called conditional random effects model. The results of the Q-test may lead researchers to an unwarranted application of the random effects model in $\mathfrak{S}_1$ especially when using HOd. As a consequence, a loss of power for significance testing would result. Alternatively, the application of the Q-test may lead to the application of fixed effects models in heterogeneous situations like $\mathfrak{S}_2$ and $\mathfrak{S}_3$. In the latter case, tests and confidence intervals would result in unduly small widths for intervals and overpowered tests for most approaches.

8.5.2 The Hunter-Schmidt Approach to the Test of Homogeneity: The 75%- and 90%-rule

In Section 5.3, the 75%-rule by Hunter and Schmidt (1990) was introduced. In short form, it states that if 75% of the observed variance of effect sizes can be explained by artifacts — especially sampling error of the estimator — then the rest of the variance in observed effect sizes can be attributed to unobserved artifacts and homogeneity is therefore given. As the indicant of homogeneity

Table 8.16 Rejection Rates for the Q-Test by k and σ_ρ^2 in $\mathfrak{S}_3$

k	σ_ρ^2	HOr	HOd	HS	OP-FE
	.0025	.1873	.2720	.1844	.1988
	.01	.3220	.3988	.3162	.3416
4	.0225	.4712	.5452	.4614	.4888
	.04	.5528	.6214	.5421	.5818
	.0625	.6473	.7120	.6331	.6720
	.0025	.2515	.3709	.2547	.2960
	.01	.4409	.5462	.4378	.5031
8	.0225	.6087	.7008	.5999	.6690
	.04	.6949	.7794	.6828	.7684
	.0625	.7824	.8552	.7661	.8458
	.0025	.3221	.4846	.3348	.4153
	.01	.5483	.6839	.5513	.6652
16	.0225	.7111	.8210	.7064	.8189
	.04	.7916	.8884	.7812	.9016
	.0625	.8675	.9406	.8534	.9510
	.0025	.4002	.6132	.4253	.5535
	.01	.6361	.8063	.6478	.8091
32	.0225	.7885	.9143	.7887	.9292
	.04	.8628	.9603	.8556	.9755
	.0625	.9264	.9859	.9160	.9933
	.0025	.4806	.7441	.5198	.6969
	.01	.7119	.9084	.7327	.9200
64	.0225	.8501	.9745	.8550	.9848
	.04	.9164	.9936	.9123	.9977
	.0625	.9650	.9989	.9584	.9998
	.0025	.5548	.8574	.6095	.8297
	.01	.7782	.9736	.8051	.9810
128	.0225	.8997	.9973	.9067	.9988
	.04	.9555	.9998	.9536	1
	.0625	.9892	1	.9857	1
	.0025	.6228	.9389	.6927	.9341
	.01	.8327	.9970	.8639	.9980
256	.0225	.9382	1	.9457	1
	.04	.9831	1	.9817	1
	.0625	.9989	1	.9983	1

Note. Proportion for tests at $\alpha = .05$.

in this procedure, the ratio of the estimated sampling error over the observed variance of effect sizes is considered. The ratios are compared to a value of .75 for the 75%-rule and to .90 for the 90%-rule, respectively. The 90%-rule is usually considered to be more suitable for Monte Carlo studies like the present one, where no artifacts are part of the design (see Cornwell & Ladd, 1993; Sack-

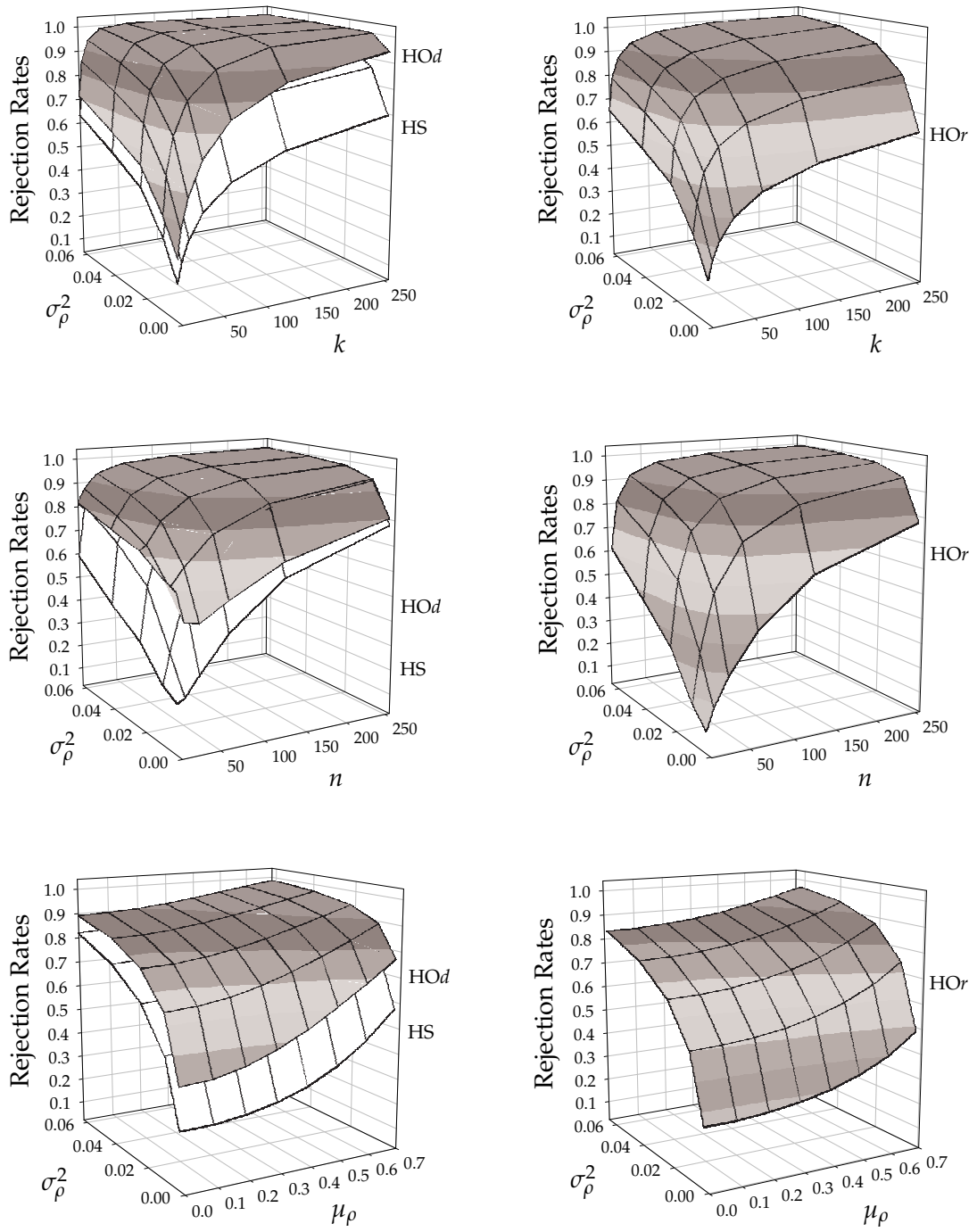

Figure 8.24 Rejection rates for the Q-test in $\mathfrak{S}_3$, $\alpha = .05$.

ett et al., 1986), and for this reason is also included in the results. If the ratios are larger than or equal to the mentioned values, homogeneity is assumed to prevail. In analogy to the hypothesis tests for homogeneity already presented, the rates of rejecting the hypothesis of homogeneity by using these rules are assessed. Since no artifacts are present in the Monte Carlo study, the situations correspond to cases in which all possible artifacts have been corrected for.

Table 8.17 Rejection Rates for 75%- and 90%-Rule in $\mathfrak{S}_1$

	Statistic				
	Max.	Mean	Median	Min.	SD
HS-75%	.7339	.1035	.1101	0	0.0916
HS-90%	.9637	.2687	.2622	.0660	0.1194
HS-ratio	6.0165	1.5040	1.0859	.6759	0.9538

Note. The total number of values described by these statistics is 420. HS-75% = Proportion of meta-analyses indicating heterogeneity according to 75%-rule, HS-90% = Proportion of meta-analyses indicating heterogeneity according to 90%-rule, HS-ratio = Ratio of estimated variance due to sampling error ($\hat{\sigma}_e^2$) over observed variance of effect sizes ($\hat{\sigma}_r^2$).

The rejection rates in $\mathfrak{S}_1$ for applying both rules along with descriptive statistics for the values of the ratio are provided in Table 8.17.

Since the 75%- and 90%-rule are not tests in a formal statistical sense it is not clear what the standards of comparison are. Adopting the procedures applied in previous Monte Carlo studies on the subject (e.g., Cornwell & Ladd, 1993; Sackett et al., 1986; Sagie & Koslowsky, 1993), the tests are expected to falsely indicate heterogeneity only in 5% of the cases in analogy to standard statistical tests. By applying this criterion to the results in $\mathfrak{S}_1$ in Table 8.17 it is recognized that neither of the rules attains a value of 5% and both rules indicate heterogeneity in a homogeneous situation far too often. Although the mean value of the HS-ratio is clearly larger than one, this does not necessarily mean that only a small portion of the ratios reaches values smaller than the criteria. As is evident from the standard deviation, there is also high variability among the ratios leading to the relatively high rejection rates. Results not shown here indicate that the minima reported in Table 8.17 only occur in cases of maximum k and n (both 256). Because the value in the denominator of the ratio is simply the observed variance of the effect sizes and this variance actually *is* sampling error in $\mathfrak{S}_1$, the results point to underestimates of the sampling error variance by the term in the numerator.

Additional information on the changes in the rejection rates across values of n and μ_ρ in $\mathfrak{S}_1$ can be seen in Figure 8.25. Both rules are depicted in this graph and represented by different surfaces.

The tendency for very large rejection rates to occur for large effects and small n is clearly visible. Moreover, both surfaces maintain a height that indicates rejection rates generally too high for both rules, though the 75%-rule performs better in $\mathfrak{S}_1$, — a fact that is trivial — it does not perform satisfactorily. This is rather surprising at first glance since an assumption that about 75% of observed variance can simply be ignored and attributed to some unobserved causes of data turbulences seems quite liberal and favoring homogeneity. Ironically then, the seemingly liberal rules lead to a false rejection of the hypothesis of homogeneity far too often.

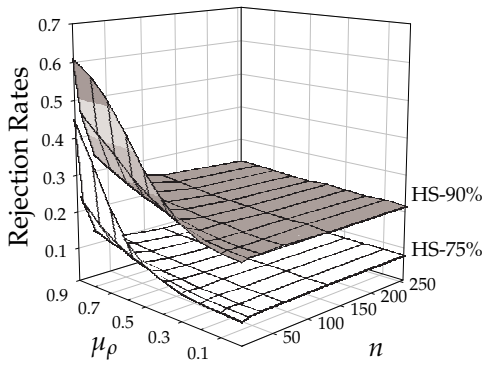

Figure 8.25 Rejection rates for the 75%- and 90%-rule in $\mathfrak{S}_1$ by n and μ_ρ.

Table 8.18 Rejection Rates for 75%- and 90%-Rule in $\mathfrak{S}_2$

	Statistic				
	Max.	Mean	Median	Min.	SD
HS-75%	1	.7814	.9995	0	.3219
HS-90%	1	.8635	1	.0838	.2284
HS-ratio	6.9055	.5447	.3944	.0122	.5962

Note. The total number of values described by these statistics is 1890. HS-75% = Proportion of meta-analyses indicating heterogeneity according to 75%-rule, HS-90% = Proportion of meta-analyses indicating heterogeneity according to 90%-rule, HS-ratio = Ratio of estimated variance due to sampling error ($\hat{\sigma}_e^2$) to observed variance of effect sizes ($\hat{\sigma}_r^2$).

None of the rules therefore seems to represent a viable alternative to the Q-test in $\mathfrak{S}_1$. A trivial consequence of the high rejection rates in $\mathfrak{S}_1$ is a better performance in heterogeneous situations. Hence, it should again be kept in mind that these "tests" do not perform well in $\mathfrak{S}_1$ when inspecting the results for other cases.

The results for the next two situations, $\mathfrak{S}_2$ and $\mathfrak{S}_3$ are shown in Tables 8.18 and 8.19, respectively. The results for $\mathfrak{S}_2$ shown in Table 8.18 indicate a smaller mean ratio and high rates of rejecting the assumption of homogeneity, as would be expected in a heterogeneous situation and by the high baseline of rejection rates in $\mathfrak{S}_1$.

If the conventional level of 80% rejection rates is considered satisfactory for such a "test" and applied to evaluate the results, both rules approximately reach this criterion overall. The results for $\mathfrak{S}_3$ in Table 8.19 lead to the same conclusion based on the mean values of rejection rates.

However, minimum values and standard deviations also indicate that there are considerable differences across levels of the design variables. In contrast to $\mathfrak{S}_1$, the ratios increase for larger values of n, k, and μ_ρ, reaching their maximum when all design variables take on their highest values. Examples for the

Table 8.19 Rejection Rates for 75%- and 90%-Rule in $\mathfrak{S}_3$

	Statistic				
	Max.	Mean	Median	Min.	SD
HS-75%	1	.7076	.9035	0	.3391
HS-90%	1	.8130	.9719	.0795	.2479
HS-ratio	5.0551	.6991	.5736	.0163	.6650

Note. The total number of values described by these statistics is 1848. HS-75% = Proportion of meta-analyses indicating heterogeneity according to 75%-rule, HS-90% = Proportion of meta-analyses indicating heterogeneity according to 90%-rule, HS-ratio = Ratio of estimated variance due to sampling error ($\hat{\sigma}_e^2$) to observed variance of effect sizes ($\hat{\sigma}_r^2$).

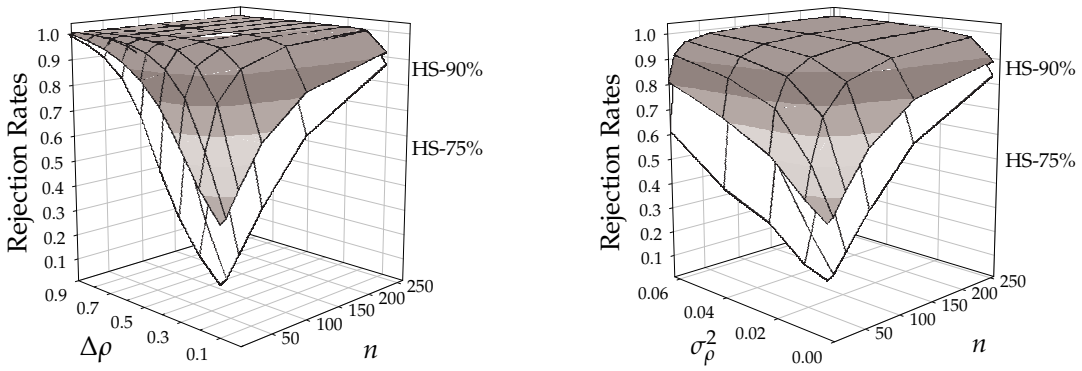

Figure 8.26 Rejection rates for the 75%- and 90%-rule in $\mathfrak{S}_2$ and $\mathfrak{S}_3$.

change in rejection rates for both rules as they occur across levels of the design variables are presented in Figure 8.26.

The general trends look similar to those reported for significance tests in previous sections with smaller rejection rates for lower n, $\Delta\rho$, and σ_ρ^2, respectively. The graphs in Figure 8.26 support the notion of some deficiencies for both rules when the levels of design variables are not at least of medium value.

In sum, the 75%- and 90%-rule of Hunter and Schmidt do not perform much better in all three situations in comparison to homogeneity tests presented in preceding subsections. Results not provided here show that, in general, power to detect heterogeneity can become quite low for combinations of low n, $\Delta\rho$, and σ_ρ^2, respectively. Due to the very high rejection rates in $\mathfrak{S}_1$ and low power in many conditions in heterogeneous situations, the rules should be used with caution. Especially when n and the assumed heterogeneity variance are rather small, decisions about the application of random effects approaches or explanatory models and conclusions concerning the generalizability of an effect should not be solely based on the results of the 75%- or 90%-rule.

8.6 ESTIMATION OF HETEROGENEITY VARIANCE

The estimation of the variance of effect sizes in the universe of studies is an important part of random effects models and also of the HS-type meta-analysis. It is a parameter of interest in itself, like the expected value in the universe of studies. It may, however, also be used in further computations in meta-analysis. For example, heterogeneity variance is used to construct so-called *credibility intervals* as proposed by Hunter and Schmidt (1990). Credibility intervals are constructed analogously to confidence intervals but use the standard deviation of the heterogeneity variance instead of the standard error of the estimator to arrive at estimates for the interval limits. Credibility intervals are not part of the Monte Carlo study and are therefore not considered here.

The most prominent estimators of heterogeneity variance in applications of meta-analysis in psychology, DSL and HS, will be evaluated in this section. In addition, the estimator OP-RE presented in Subsection 5.4.2 will also be evaluated to assess its performance in relation to the standard approaches.

As was the case in the context of estimating μ_ρ, the estimated parameter in the various situations will first be considered. In $\mathfrak{S}_1$, there simply is no variance to be estimated, that is, it is zero. The behavior of the estimators will be examined in two versions. First, the results for the truncated variance estimator will be reported, and second, the results for the non-truncated version thereafter. Recall from Section 5.4.1 that the truncated variance estimator in the DSL approach is $\hat{\sigma}^2_{\zeta+} = \max\{0, \hat{\sigma}^2_\zeta\}$. That is, negative variance estimates which may arise in practice are set to zero for the truncated estimator. The non-truncated version does not set negative estimates to zero. Of course, an analogue procedure is applied in r-space when HS and OP-RE are considered: $\hat{\sigma}^2_{\rho+} = \max\{0, \hat{\sigma}^2_\rho\}$.

Since the DSL estimator of heterogeneity variance is based on Fisher-z transformed correlation coefficients, the corresponding parameter is also in z-space. This is mainly of importance for situations $\mathfrak{S}_2$ and $\mathfrak{S}_3$, where the universe variances have to be computed in order to assess biases. In $\mathfrak{S}_2$, the variance of the universe effect sizes is computed as follows:

$$\sigma^2_\zeta = \frac{\left(\zeta_1 - \mu_\zeta\right)^2 + \left(\zeta_2 - \mu_\zeta\right)^2}{2}.$$

For $\mathfrak{S}_3$, no simple form to compute the variance in z-space resulting from a beta distributed variable P is available. Thus, variances have to be determined via

$$\sigma^2_\zeta = \int_{-1}^{1} \left(\tanh^{-1} r\right)^2 f(r)dr - \mu^2_\zeta,$$

where $f(r)$ denotes the beta probability density function. μ_ζ and σ^2_ζ are in z-space. Note that μ_ζ is given by $\mu_\zeta = \int_{-1}^{1} \tanh^{-1}(r) f(r)dr$. The various values as used in the Monte Carlo study can be found in Tables A.1 and A.2 in the appendix.

Table 8.20 Bias of $\hat{\sigma}^2_\rho$ (HS & OP-RE) and $\hat{\sigma}^2_\zeta$ (DSL) in $\mathfrak{S}_1$

	Statistic				
	Max.	Mean	Median	Min.	SD
HS-nt	.0110	.0014	.0001	−.0001	.0027
HS	.0433	.0042	.0013	0	.0068
OP-RE-nt	.1289	.0099	.0001	−.0144	.0249
OP-RE	.1289	.0130	.0015	0	.0264
DSL-nt	.0009	−.0013	−.0001	−.0164	.0031
DSL	.0622	.0070	.0024	.0001	.0116

Note. The total number of values described by these statistics is 420. -nt designates non-truncated estimators.

8.6.1 Homogeneous Situation $\mathfrak{S}_1$

In the homogeneous situation, the estimators presented in Chapter 5 generally overestimate the heterogeneity variance. This is due to the truncation of the resulting estimates at a value of zero when values less than zero are encountered. To assess whether the non-truncated versions actually estimate the universe parameter precisely and how far off the truncated versions are from zero, both versions are provided in the following presentation. The truncated versions therefore correspond to the estimators used in practice and the non-truncated versions are only given for comparison. The non-truncated estimators are labeled by the additional suffix *-nt*.

In Table 8.20 results for the biases of the estimators in $\mathfrak{S}_1$ are presented. The values are computed in analogy to the biases of the estimators of μ_ρ (see Section 8.2.1).

Unfortunately, the biases of variances in Table 8.20 and also those presented in the following are not directly comparable because the values for HS and OP-RE are given in r-space and those of DSL in the space of z. Nevertheless, in the given situation one would expect the biases of DSL to be uniformly larger to a certain degree than the variances of HS and OP-RE due to the characteristics of the different spaces. Recall from Section 3.1 that the Fisher-z transformation stretches the values of r particularly in the boundary regions (see also Figure 3.1) and therefore leads to larger variances in z-space as compared to r-space. Trivially, the truncated values are always at least as large as their non-truncated counterparts.

As evidenced by the minimum values in Table 8.20, some remarkable negative estimates indeed emerge in some cases. Interestingly, the maxima of both versions for the estimators do not always agree. This is due to rare cases in which very large variances occur for the estimates and a large portion of variance estimates is less than zero. The values reported in Table 8.20 indicate some deficiencies associated with OP-RE in relation to DSL and HS. The OP-RE estimator shows maximum values far too large to be acceptable. The mean

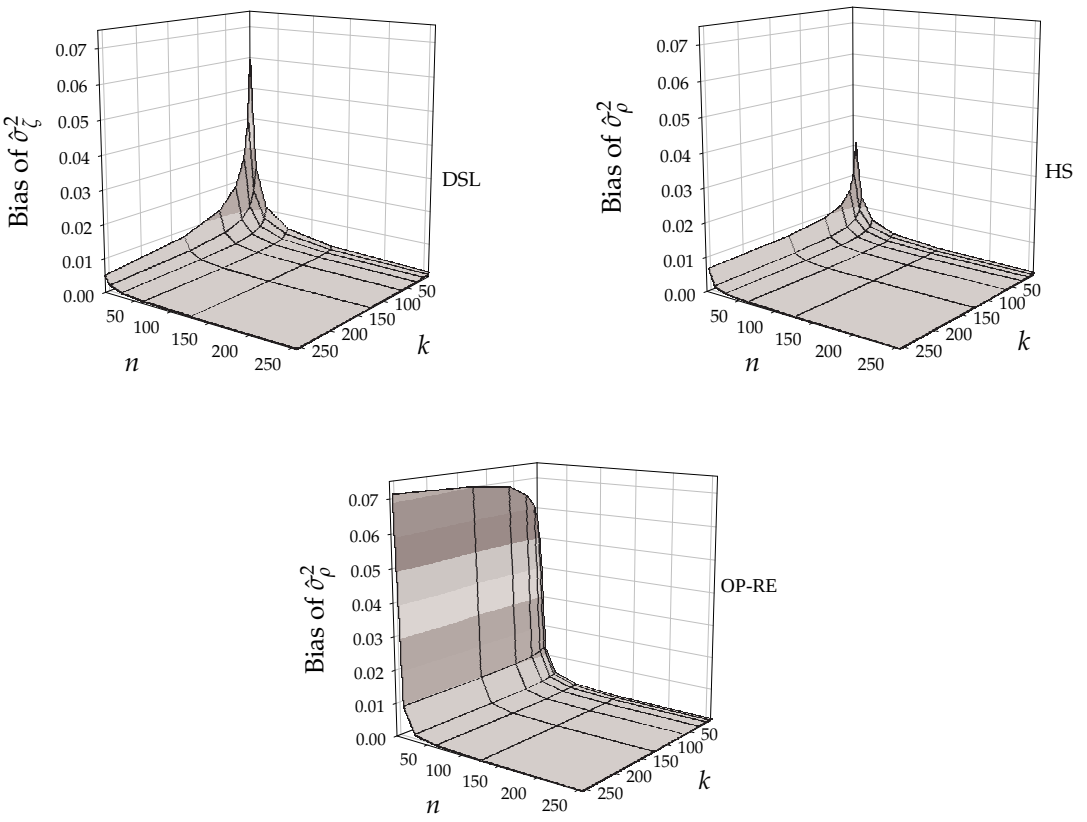

Figure 8.27 Bias of $\hat{\sigma}^2_\rho$ (HS & OP-RE) and $\hat{\sigma}^2_\zeta$ (DSL) in $\mathfrak{S}_1$ by k and n.

and median values shown, however, indicate rather good performance of the approaches overall.

To elucidate under which constellations of the design variables the estimators perform better or worse, a series of graphs is presented in Figure 8.27. Again, an array of graphs shows the estimators' performance across combinations of the design variable levels of k and n.

The estimators' biases shown in Figure 8.27 are only given for the truncated versions to focus on findings relevant for the application of the methods in practice. In general, all panels indicate good performance of the estimators for large values of the design variables. However, DSL obviously overestimates σ^2_ζ when n and k are very small and also retains a positive bias for all values of k when n is very small. This is due to the truncation of the variances. The same shape of surface emerges for HS in the upper right panel but the biases for combinations of a small number of studies and very small sample sizes appear smaller than those of DSL. Since these two estimators operate in different spaces (r vs. z), it is not perfectly clear which estimator actually shows larger bias in comparison. The lower panel gives the results for OP-RE and indicates a very poor performance of the estimator for small values of n across all values of k. Only when n grows larger and reaches a value of approximately 64 does the estimator show acceptable performance.

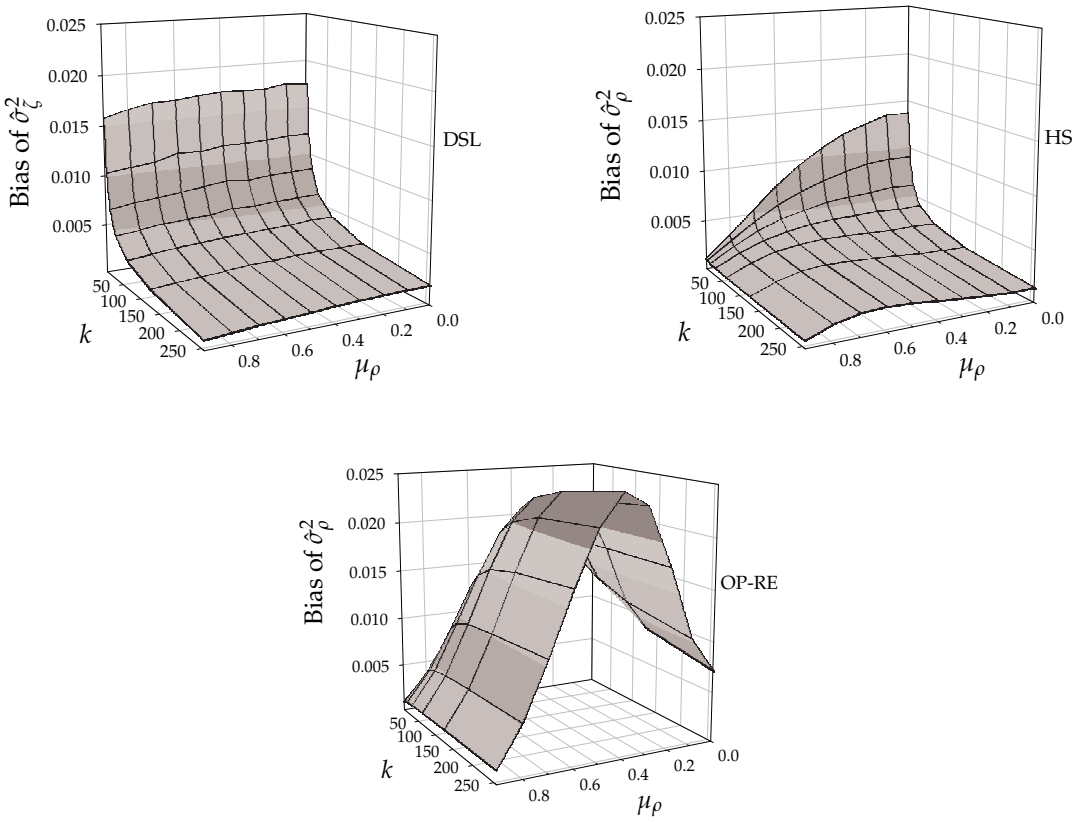

Figure 8.28 Bias of $\hat{\sigma}^2_\rho$ ((HS & OP-RE) and $\hat{\sigma}^2_\zeta$ (DSL) in $\mathfrak{S}_1$ by k and μ_ρ.

It is, of course, also of interest whether biases of the estimators vary across values of the universe effect size. Figure 8.28 provides graphs for the design dimensions k and μ_ρ.

For DSL and HS, both upper panels in Figure 8.28 show an improved performance for larger values of k. DSL shows a relatively stable performance across all values of μ_ρ, but it is acknowledge that the slope of the surface indicates slightly better performance for larger values of μ_ρ. The results depicted in the figure suggest that at least a modest number of studies (approximately 32) have to be available when using this approach for a sufficiently precise estimation of the heterogeneity variance (i.e., very close to zero). HS, in contrast, performs best when μ_ρ is large. This tendency is most obvious for a small number of studies. Unfortunately, for values of ρ suspected to occur often in practice (around .40) the bias still seems non-negligible. Although the absolute values seem small on the vertical axis, it should be remembered that a value of .01 corresponds to a standard deviation of .10. Hence, there seems to be non-trivial bias for the HS estimator for small values of k and moderate to low μ_ρ in the universe of studies. As is the case for DSL, when the number of studies is 32 or larger, the bias seems negligible for HS.

Unlike these first two approaches, OP-RE strongly varies in biases across levels of μ_ρ, notwithstanding how many studies are aggregated, with a max-

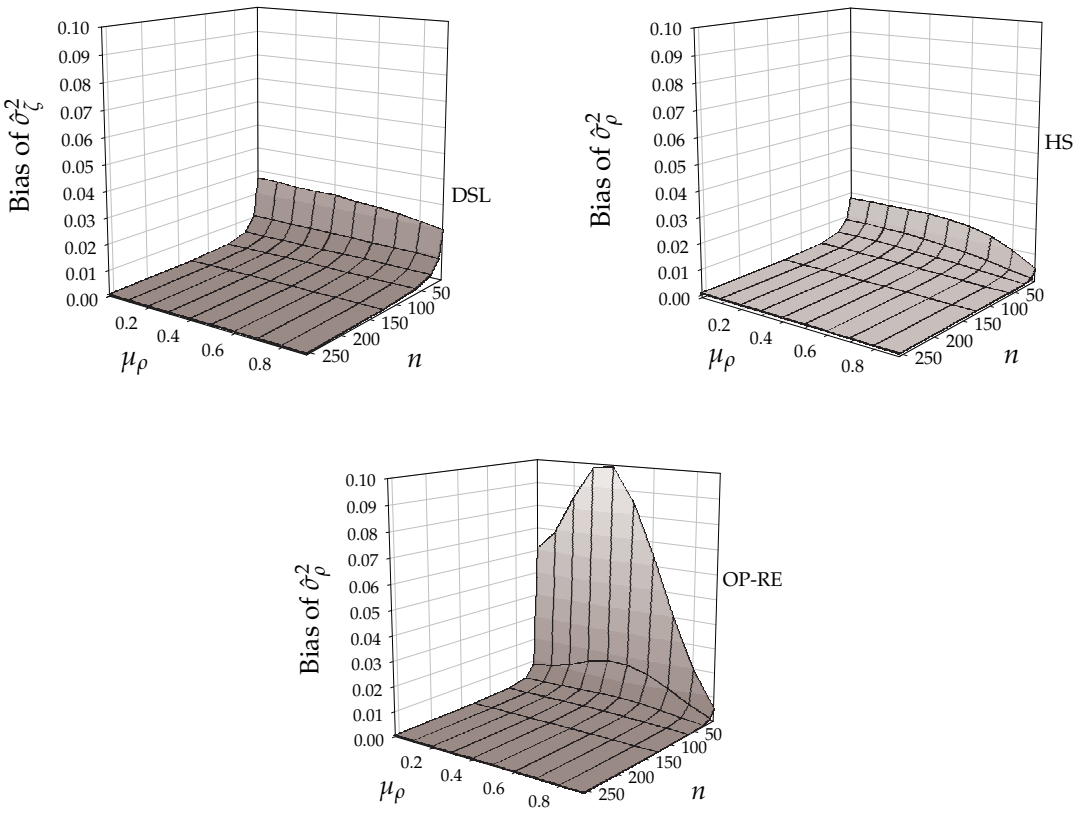

Figure 8.29 Bias of $\hat{\sigma}^2_\rho$ (HS & OP-RE) and $\hat{\sigma}^2_\zeta$ (DSL) in $\mathfrak{S}_1$ by n and μ_ρ.

imum bias at a value of approximately $\mu_\rho = .40$. It is again suspected that this phenomenon is caused by the weighting scheme of OP-RE. The region of maximum bias falls near the point of $\mu_\rho = .347$ where the biggest *change* in the variance of G occurs (see Section 3.1). A big change in variance transfers to big differences in weights since the variance estimates are used in the weighting scheme of the OP-RE approach. If this were true, then the bias should diminish for larger sample sizes. This is indeed the case as the lower panel in Figure 8.29 shows. This figure completes the results for the biases of $\hat{\sigma}^2_\rho$ and $\hat{\sigma}^2_\zeta$ in $\mathfrak{S}_1$.

Again, it can clearly be seen that for very low values of n none of the estimators shows acceptable performance but performance quickly gets better and reaches acceptable levels for sample sizes supposed to be encountered most often in practice (32 or larger). Poor performance for the approaches only occurs for very small n. Though DSL in the upper left panel does not seem to reach small biases for growing n as fast as HS, the reader is again cautioned against such a comparative interpretation because of the different spaces in which DSL and the other approaches operate. Overall, at least the estimators HS and DSL seem to show acceptable performance in $\mathfrak{S}_1$ when n and k are not very small.

Table 8.21 Bias of $\hat{\sigma}^2_\rho$ (HS & OP-RE) and $\hat{\sigma}^2_\zeta$ (DSL) in $\mathfrak{S}_2$

	Statistic				
	Max.	Mean	Median	Min.	SD
HS-nt	.0665	.0034	.0009	−.0253	.0087
HS	.0665	.0044	.0014	−.0253	.0096
OP-RE-nt	.2739	.0440	.0184	−.0130	.0560
OP-RE	.2739	.0452	.0189	−.0018	.0559
DSL-nt	.2559	.0121	.0030	−.0115	.0254
DSL	.2566	.0150	.0051	−.0099	.0262

Note. The total number of values described by these statistics is 1890. -nt designates non-truncated estimators.

8.6.2 Heterogeneous Situations $\mathfrak{S}_2$ and $\mathfrak{S}_3$

The heterogeneity variance estimators become especially important in cases where $\sigma^2_\rho \neq 0$. In such cases, it can be evaluated whether the truncated versions of the estimators still provide overestimates, as is the case for some combinations of levels of design variables in $\mathfrak{S}_1$. Additionally, the two situations $\mathfrak{S}_2$ and $\mathfrak{S}_3$ enable an evaluation of the estimators for a discrete distribution in the universe of studies and for a continuous distribution. For the latter, it should be kept in mind that the beta distribution strongly deviates from normality the larger μ_ρ is. This is considered to be more adequate for *r*-space in comparison to a truncated or otherwise distorted normal distribution, for example, as was used in other Monte Carlo studies (e.g., Overton, 1998; and probably also Field, 2001).

Table 8.21 provides overall results of the three estimators in both versions available. As can be seen, differences between the truncated and non-truncated versions of the estimators do not differ substantially. The focus will therefore be exclusively laid on the truncated estimators.

All three approaches differ in biases. HS is close to the variances to be estimated amongst the approaches under consideration. Mean and median values indicate a good overall performance but minima and maxima also show that there are conditions under which the estimator over- or underestimates the universe variance of effect sizes. OP-RE, in contrast, generally overestimates variances, in some cases to a very large degree. DSL shows a slight tendency for overestimation as indicated by the values in the table but clearly not as strong as OP-RE. The measures of central tendency for DSL close to zero suggest a performance similar to HS. Yet, the maximum values for DSL also suggest that the tendency for overestimation can be strong in some cases. Unfortunately, this is no unequivocal indicator for strong overestimation because it is the bias computed in *z*-space. To elucidate conditions under which the estimators do not perform very well, a series of graphs is provided once more.

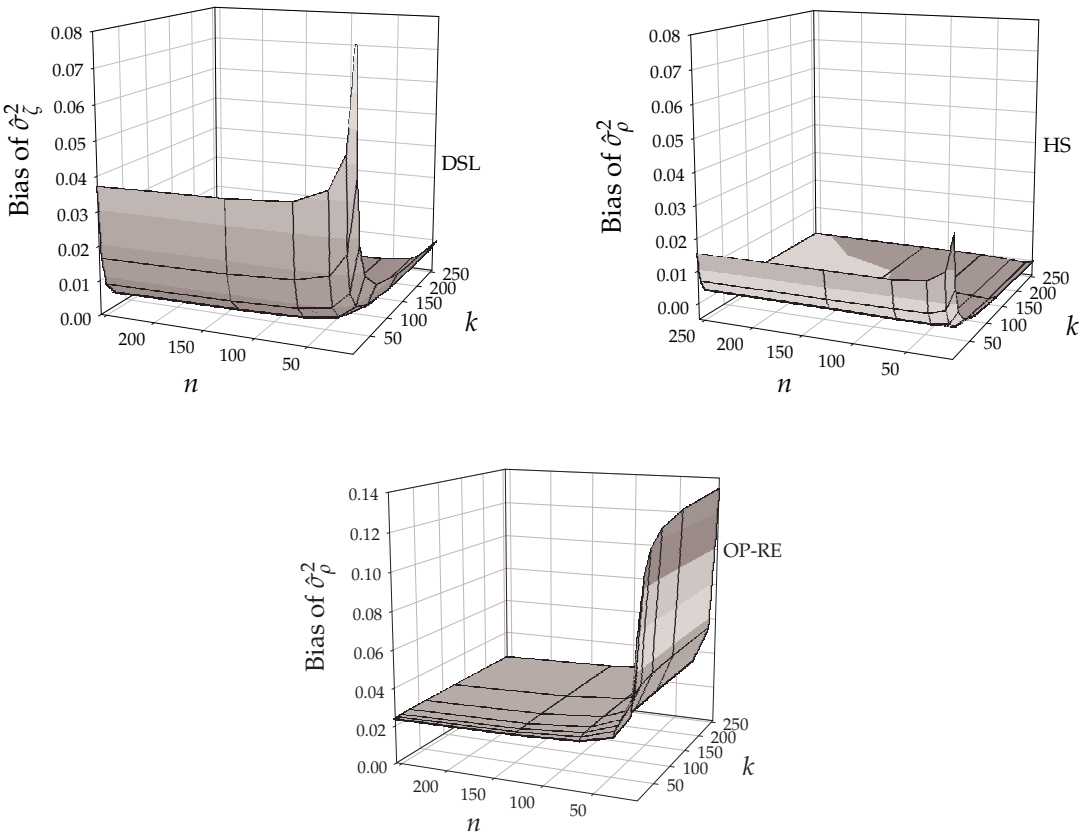

Figure 8.30 Bias of $\hat{\sigma}^2_\rho$ (HS & OP-RE) and $\hat{\sigma}^2_\zeta$ (DSL) in $\mathfrak{S}_2$ by n and k.

The conditions for largest biases of DSL and HS, indicated by the upper panels in Figure 8.30, are again cases of low n and especially k. The combination of both low n and k represents the worst case in terms of bias. Unlike the results presented for $\mathfrak{S}_1$, the biases for these approaches are generally high for k less than 16, irrespective of n. Absolute values for biases are also different for each of these estimators in comparison to the results in $\mathfrak{S}_1$.

A very different shape of surface emerges again for the bias of OP-RE. Results for this estimator indicate a poor performance for low n whereas biases decline for larger n, irrespective of k. The surfaces of DSL and OP-RE do not approximate a value of zero bias for larger k and n, respectively. Note that this is actually the case for HS, which can be regarded as performing best in this respect. The biases of DSL and OP-RE instead converge to some nonzero positive value. This is due to the fact that for both estimators biases also very strongly vary for different values of $\Delta\rho$. This is illustrated in Figure 8.31 where the estimators operating in r-space are shown in one panel. The upper panels provide biases for $\Delta\rho$ by n and the lower two for $\Delta\rho$ by k.

Both panels illustrate the rising bias both for DSL and OP-RE for larger values of $\Delta\rho$. The results explain why the values to which these approaches converge (as shown in Figure 8.30) are larger than zero. The larger the difference between universe values in $\mathfrak{S}_2$, the larger are the biases of DSL and OP-RE.

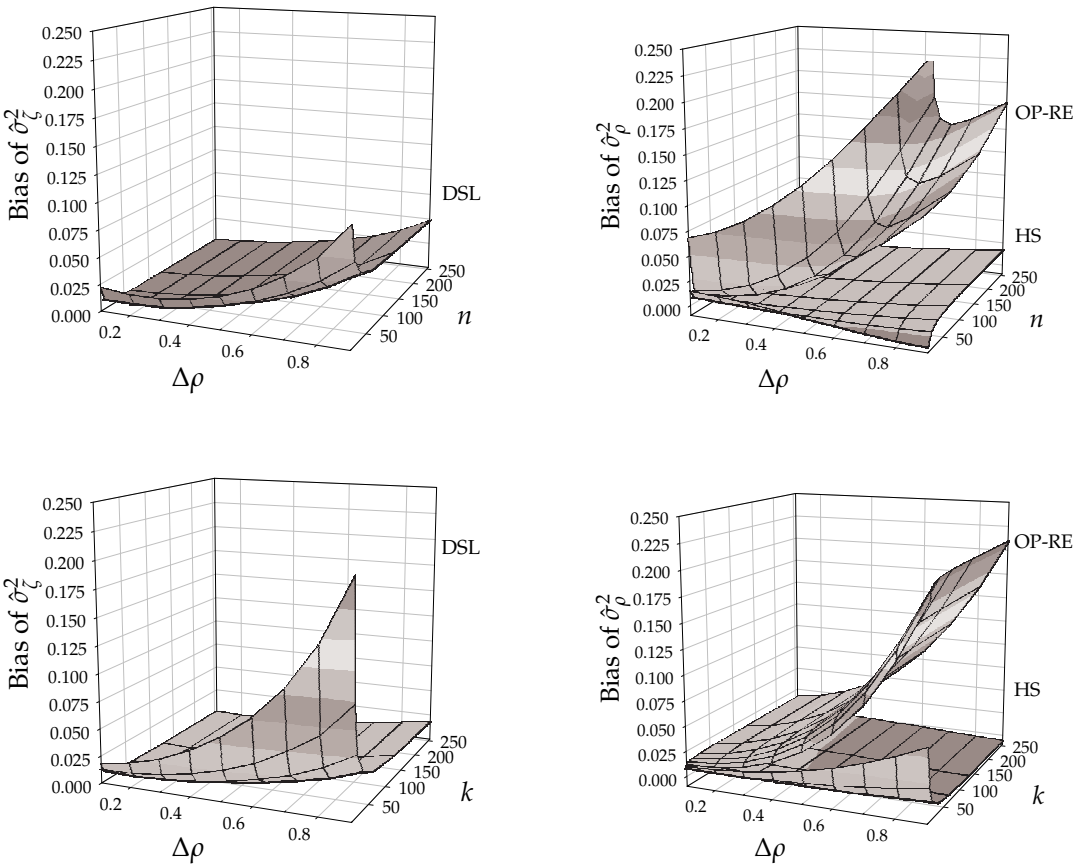

Figure 8.31 Bias of $\hat{\sigma}^2_\rho$ (HS & OP-RE) and $\hat{\sigma}^2_\zeta$ (DSL) in $\mathfrak{S}_2$ by n and $\Delta\rho$ as well as k and $\Delta\rho$.

In the case of OP-RE this is suspected to be caused by the weighting scheme, whereas in the case of DSL — though biases are not directly comparable in absolute terms — this is proposed to be a result of transformation into z-space. In contrast to the performance of these two approaches, HS shows a very good performance in $\mathfrak{S}_2$ and seems to be the approach of choice amongst the ones available in this situation. Cautions against the use of the HS estimator appear reasonable in cases of small k (i.e., $k < 16$), especially when large differences between effect sizes in the universe if studies are suspected.

Finally, the results for biases in $\mathfrak{S}_3$ are presented, the situation with a continuous distribution in the universe of studies. Table 8.22 provides overall results first.

The results in Table 8.22 seem to indicate a much better overall performance of DSL as compared to the previous situation. However, due to difficulties in directly comparing variances in situations of type $\mathfrak{S}_2$ and $\mathfrak{S}_3$ as well as complications arising from interpreting absolute values for biases in z-space, the indication of a better performance are not strong. HS shows very small mean bias whereas OP-RE again strongly overestimates universe variances of effects sizes. Again, maximum and minimum values indicate varying performance of

Table 8.22 Bias of $\hat{\sigma}^2_\rho$ (HS & OP-RE) and $\hat{\sigma}^2_\zeta$ (DSL) in $\mathfrak{S}_3$

	Statistic				
	Max.	Mean	Median	Min.	SD
HS-nt	.0121	−.0001	−.0001	−.0193	.0036
HS	.0433	.0014	0	−.0193	.0059
OP-RE-nt	.2161	.0237	.0062	−.0162	.0373
OP-RE	.2161	.0255	.0074	−.0147	.0376
DSL-nt	.0243	.0008	.0003	−.0154	.0029
DSL	.0648	.0047	.0010	−.0129	.0099

Note. Valid values for all entries are 1848. -nt designates non-truncated estimators.

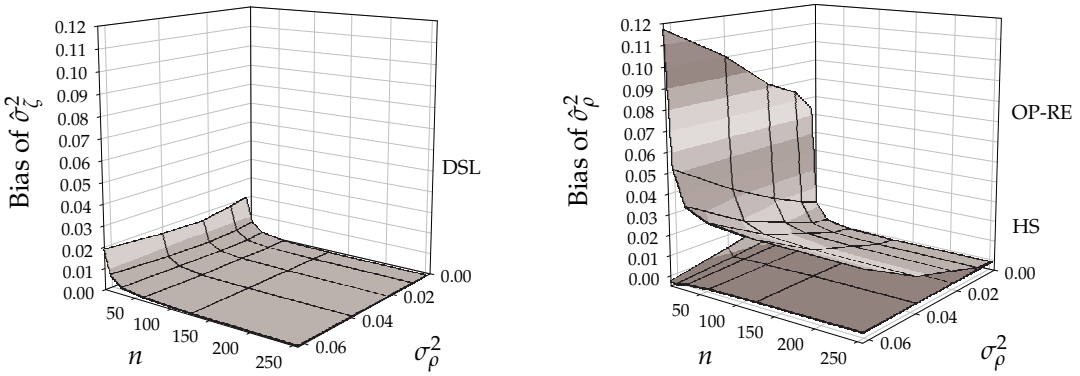

Figure 8.32 Bias of $\hat{\sigma}^2_\rho$ (HS & OP-RE) and $\hat{\sigma}^2_\zeta$ (DSL) in $\mathfrak{S}_3$ by n and σ^2_ρ.

the approaches across levels of the design variables. Interestingly, the maxima for the truncated versions of the HS and DSL estimators occur in cases similar to $\mathfrak{S}_1$, that is, for smallest values of k, n, and σ^2_ρ. In all other cases both approaches show smaller biases.

Graphs are finally presented to assess the performance of the approaches in various design regions. Figure 8.32 illustrates the results for n and σ^2_ρ.

The left panel in this figure shows that overestimation is larger for DSL only for very small n. With sample sizes larger than 16, the bias seems negligible. In the right panel of Figure 8.32, both HS and OP-RE are depicted. The high biases of OP-RE for small n are clearly visible. Since OP-RE is r-based and does not use the Fisher-z transformation, the absolute values for biases can be deemed extremely large. Furthermore, biases strongly raise for OP-RE with increasing values of σ^2_ρ. This is neither true for DSL nor HS. The results for HS indicate that biases are only elevated for small n and small σ^2_ρ, a case that approaches homogeneity. This is however, not visible in the right panel of Figure 8.32 since the surface of OP-RE covers this region. As was highlighted in the context of presenting the results in $\mathfrak{S}_1$, the biases for HS can be considered as non-negligible in some extreme cases in this design region. Nevertheless,

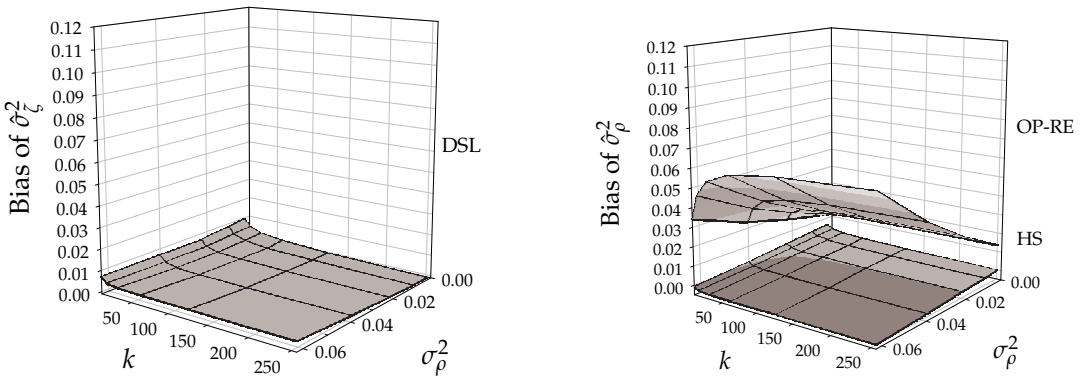

Figure 8.33 Bias of $\hat{\sigma}^2_\rho$ (HS & OP-RE) and $\hat{\sigma}^2_\zeta$ (DSL) in $\mathfrak{S}_3$ by k and σ^2_ρ.

biases are generally relatively small for the HS approach in $\mathfrak{S}_3$. Although DSL also shows a rather good performance in $\mathfrak{S}_3$, HS appears as the most recommendable approach of the three approaches under examination for this situation, since it shows the smallest biases and provides estimates in the space of r.

The next two panels in Figure 8.33 give a very similar impression of the relative performance of the approaches across values of k and σ^2_ρ. For high universe variances, biases of DSL and OP-RE are rather large. On the other hand, HS shows small biases in most cases, albeit the values evidently also vary across levels of k and σ^2_ρ. The trend of larger biases across values of σ^2_ρ is the opposite as compared to the other two approaches. High variances seem to be estimated with appreciable precisions whereas low variances are overestimated. This is due to the truncation in the HS estimator. Results for the non-truncated version, not shown here, indicate almost zero bias in all regions of the design, in particular also those for which values are slightly elevated in Figure 8.33.

The last graphs provided to assess biases are given in Figure 8.34. They underscore the generally good performance of HS, as is evident in the upper right panel. Although biases are not zero across all levels of the design variables, the absolute values are very small. DSL is depicted in the upper left panel and does not show a clear trend of bias across levels of σ^2_ρ and μ_ρ. Nevertheless, absolute biases are also small in absolute value for this approach. OP-RE again shows some variation in biases across levels of the design variables with largest biases occurring for combinations of large μ_ρ and large σ^2_ρ. Due to the large biases shown in all design level combinations in $\mathfrak{S}_3$, it is certainly no interesting alternative to the other two estimators.

In sum, despite small overestimation of the truncated version of the HS estimator in $\mathfrak{S}_1$, it seems to provide the best estimator of heterogeneity variance amongst the three approaches examined. The cases where HS shows overestimation of heterogeneity variance are not likely to be encountered often in practice, but are of interest to find the boundary values for levels of design variables in order to caution against potential problems in estimation. The bias

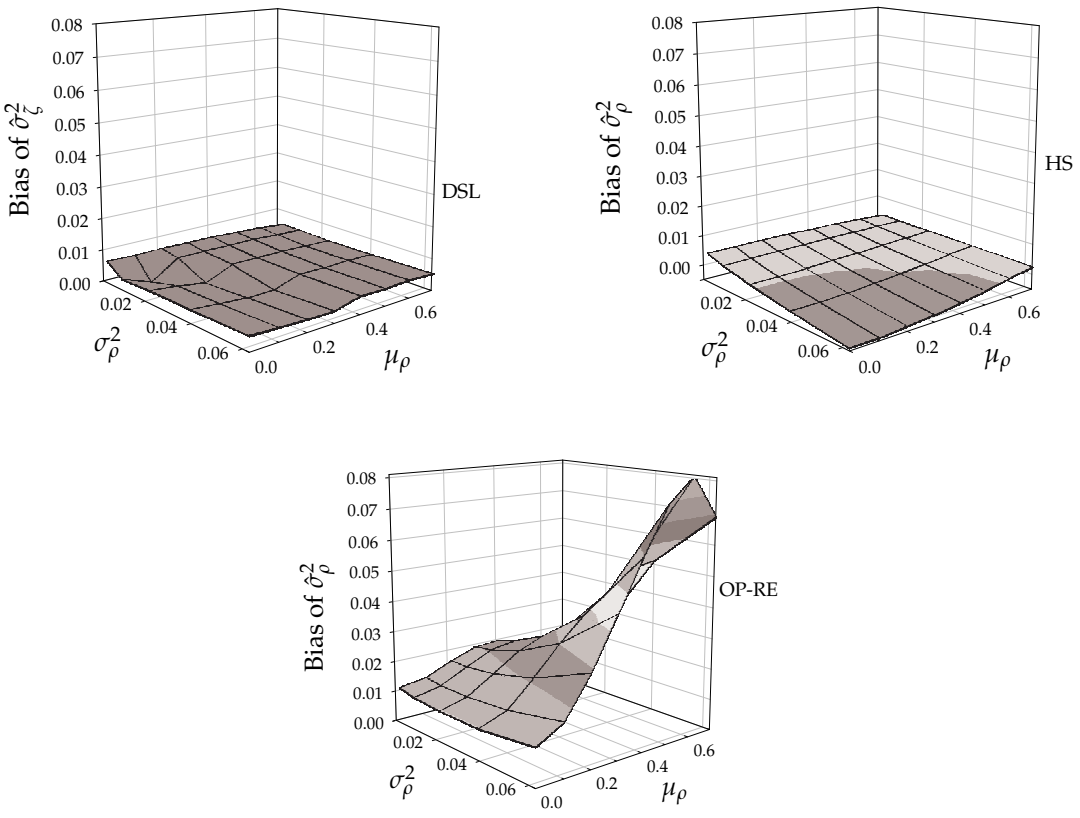

Figure 8.34 Bias of $\hat{\sigma}^2_\rho$ (HS & OP-RE) and $\hat{\sigma}^2_\zeta$ (DSL) in $\mathfrak{S}_3$ by μ_ρ and σ^2_ρ.

of DSL was only examined in z-space, so some reservations with respect to a negative evaluation are in order. The performance of this estimator nonetheless showed variation in the Monte Carlo study that does not let it appear as a promising alternative in comparison to the simple HS estimator. Furthermore, there is no option available to date to transform the results of the DSL estimator into r-space. Hence, variance estimates are in z-space and hard to interpret. This is another limitation of this approach which makes its use in practical applications of meta-analysis unattractive.