This is chapter 9: Synopsis of statistical methods and Monte Carlo study results (pp. 183-190) from

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

9

Synopsis of Statistical Methods and Monte Carlo Study Results

The statistical methods for meta-analysis of correlations were outlined in this book and classified with respect to a series of characteristics. The main relevant characteristics for the comparison of approaches were identified to be a) the effect size measure used, b) the weighting scheme used, and c) the underlying statistical model. Although these classificatory aspects are not mutually exclusive, they are nevertheless useful to differentiate between approaches with reference to characteristics that cause differences in results.

Effect Size Used in the Approaches The coefficients under examination were

- the untransformed correlation coefficient r,
- the Fisher-z transformed correlation z,
- a bias-corrected mean Fisher-z transformed correlation $\bar{z}_{\text{Hot}}$,
- a bias-corrected untransformed correlation coefficient G, and
- the r-to-d transformed d.

For the untransformed correlation r it was shown that it is biased with respect to ρ and that the variance of the estimator depends on ρ. Fisher-z transformed correlations are biased as well but they have the desirable property that their variance only depends on the sample size and not on the population parameter. Hotelling (1953) has analyzed the bias both of r and z and proposed several corrections which were presented in Part II. Of these corrections, a bias-correction for mean z was considered to be especially attractive for use in meta-analysis and was incorporated as an independent approach in the subsequent comparison of approaches. The UMVU estimator presented by Olkin and Pratt (1958) was also considered and it was shown that its variance does also depends on the population parameter. Finally, the transformation of r to d was also included because of its high relevance for practical applications meta-analysis. This offered the opportunity to examine whether the transformation leads to different meta-analytic results when computations are based on the transformed d instead of r or z. Some of these effect sizes are used in well-known and often applied meta-analytic approaches (r, z, and d) whereas others (bias-corrected $\bar{z}_{\mathrm{Hot}}$ and G) — interestingly those with desirable properties with respect to bias — are not widely known and used.

Many of these effect sizes involve what can be called a "change of space". That is, there is a change from r-space to z-space by application of the Fisher-z transformation and a change from r-space to d-space by the corresponding conversion formula. In more mathematical parlance, this is called change of variable. The initial motivation for the former change of space was to circumvent problems with the rather untractable probability density function of R. The motivation for the latter is simply the need to bring available research findings into a common space to carry out the meta-analytic computations of an approach. Hence, both kinds of transformations are justified in the meta-analytic context. The change of space, basically designating the use of a nonlinear transformation of the correlation coefficient r to either z or d, was hypothesized to be a cause for differences in results between approaches. When meta-analytic computations are carried out in the "transformed space" (z or d) and computational results are transformed back into r-space subsequently, then differences to results from computations based on r can be expected.

Weighting Schemes The weighting scheme used in aggregating effect sizes of k studies was pointed out to be another important characteristic. There are basically two variants of weights in meta-analysis of (transformed) correlations: sample size and reciprocals of the estimator's variance. The former has the rather simple rationale of giving those studies higher weight that provide "more evidence". Of course, larger studies are simultaneously also thought to provide more precise estimates of the parameter in question (assuming consistency). Weighting by the reciprocals of the estimator's variance has a clearer statistical rationale as these weights are optimal in the sense that they provide a pooled estimator with minimum variance. Furthermore, it can be shown that under certain assumptions these weights are also those of the maximum likelihood estimator for the universe parameter in a fixed effects situation (for a proof, see Böhning, 2000, pp. 101–102). Hence, weighting by the reciprocal of the estimator's variance has very desirable statistical properties.

However, it was repeatedly argued in this book that under certain circumstances the optimal weights become suboptimal. The first reason leading to bias in the pooled estimator in the present context is lack of knowledge about the variance, and hence the need to plug in estimates in the weights. The above mentioned dependency of some estimators' variance on the universe parameter and small sample sizes making the individual estimates highly variable exacerbate the problem. This already points to the second cause for bias, namely, the dependency of the variance on the universe parameter. These two causes together produce the undesirable effect of bias in some estimators. It becomes particularly problematic when the fixed effects model is used instead of the random effects model (see below) in a situation where the latter is appropriate. In this case, this dependency leads to bias even when n grows large.

Meta-Analytic Models The presented models were the fixed effects model, the random effects model, mixture models, and hierarchical linear models. The two models of highest relevance for the classification of approaches are the FE and RE model. The difference between the FE and RE model is the conceptualization of the universe of studies as characterized either by a single constant parameter (ρ; FE; homogeneous case) or by a random variable (P; RE; heterogeneous case). In the RE model, the variance of P (heterogeneity variance) is always some positive value, whereas in the FE model it is zero by definition. Hence, approaches categorized as using the FE model do not include estimators for heterogeneity variance whereas in approaches using the RE model they are an integral part. In addition to estimating the heterogeneity variance, it is also used in the weights in approaches using the RE model.

As examples of more general models for meta-analysis, mixture models and HLM were introduced. In the latter case, it was shown that the FE and RE model are special cases, thereby revealing in what respect these two models are special or limited. In contrast to HLM, mixture models include latent variables as causes for heterogeneity of effects. These models were used to conceptualize three situations to which the approaches under examination may be applied.

Situations The first situation ($\mathfrak{S}_1$) represented the homogeneous case for which FE model approaches are appropriate. The second situation ($\mathfrak{S}_2$) was a heterogeneous situation characterized by a discrete distribution in the universe of studies. The examination in this book was limited to a dichotomous latent variable where categories have equal weights, hence a two-point uniform distribution. The third situation ($\mathfrak{S}_3$) was characterized by a continuous latent variable, thus also qualifying as a heterogeneous case, and the subsequent presentation focused on the beta distribution. For both $\mathfrak{S}_2$ and $\mathfrak{S}_3$, RE model approaches are appropriate.

Approaches The specific approaches for meta-analysis of correlations in common use in the social sciences were outlined in Part II and details on the computational procedures were given. In addition, refinements were also presented that have not yet been widely applied. A concise overview of the approaches that provides their classification according to the above mentioned characteristics and also specifies the homogeneity test and whether heterogeneity variance is estimated, is given in Table 9.1.

There are several things to note with regard to the entries in Table 9.1. Firstly, although HOT is characterized by the effect size z, the defining characteristic of this approach is actually a correction of the mean z resulting from aggregation as done in the HOr approach. Secondly, the weights are given as used in the approaches but it can be easily identified which of them are solely based on

Table 9.1 Overview of Approaches

Approach	Effect Size	Weight	Model	Homog. Test	Heterog. Variance
HOr	z	$n-3$	FE	Q	No
HOT	z	$n-3$	FE	–	No
HOd	d	$\hat{\sigma}_D^{-2}$	FE	Q	No
RR	z	n	FE	–	No
HS	r	n	RE?	75% & Q	Yes
DSL	z	$(\frac{1}{n-3}+\hat{\sigma}_\zeta^2)^{-1}$	RE	–	Yes
OP	G	n	FE	–	No
OP-FE	G	$\hat{\sigma}_G^{-2}$	FE	Q	No
OP-RE	G	$(\hat{\sigma}_G^2+\hat{\sigma}_\rho^2)^{-1}$	RE	–	Yes

Note. – = redundant to other approaches or not included in Monte Carlo study.

n and which incorporate estimated variances. Thirdly, although the weights of the HOr approach, for example, are only based on n, these are the optimal weights in the above mentioned sense. This is due to the fact that in the case of Fisher-z transformed correlations the variances are $(n-3)^{-1}$. Hence, such approaches use the optimal weights but do not suffer from the above mentioned problems. Fourthly, apart from a minor difference in testing procedures, RR is basically identical to HOr. The weight as given in Table 9.1 for the RR approach could have also been the same as for HOr according to the proponents of the RR approach. Fifthly, the classification of the HS approach as belonging to the RE model class is not entirely clear. This is indicated by a question mark but it is also recognized that the HS approach is mostly an RE approach in conceptualization. Lastly, for some of the approaches there is no entry in the column labeled "Homog. Test" because first, the test would be identical to others (e.g., HOr, HOT, RR), second, a plausible test is not included in the subsequent Monte Carlo study (OP), or such a test would simply make no sense (DSL, OP-RE).

Estimated Parameters in the Universe of Studies It was shown that differences between approaches in the effect size used are very important with respect to the estimated universe parameter. Whereas μ_ρ, the first moment of the distribution of universe effect sizes, is the estimated parameter for r- or G-based approaches (HS, OP, OP-FE, OP-RE), the parameters are different for Fisher-z-based ($\mu_{\rho z}$) and d-based ($\mu_{\rho d}$) approaches in heterogeneous situations. For the latter approach, however, it was shown that the weighting scheme leads to results for estimates of mean effect sizes to be closer to μ_ρ than $\mu_{\rho d}$. Hence, μ_ρ was considered to be the more sensible standard of comparison in the Monte Carlo study for HOd. Since μ_ρ is considered to be the parameter of interest for most meta-analysts when pooling correlation coefficients, cautions were raised about the use of approaches that do not use r in heterogeneous situations.

Monte Carlo Study In addition to the theoretical analyses of the second part, the results of a comprehensive Monte Carlo study were presented. This was done to comparatively evaluate the outcomes of the various approaches — including those not well-known and examined in previous Monte Carlo studies — in several situations ($\mathfrak{S}_1$ to $\mathfrak{S}_3$). The design was specified to include levels of several design variables (n, k, μ_ρ, σ^2_ρ) likely to arise in practice as well as levels (small n and k) to study and evaluate the performance of the approaches at boundary values. This seemed reasonable as properties of the estimators and tests are known theoretically only in approximation (for n and/or k approaching infinity).

For the design and conducting of the Monte Carlo study, several candidates for the simulation procedures were considered for generating the database of correlation coefficients. The candidates under consideration were a series of approximations to the distribution of R that were examined and evaluated in comparison to the exact density of R. None of the approximations were considered sufficiently good as to be used to generate correlation coefficients in a simulation study. The simulation procedures used in the Monte Carlo study were therefore specified in a computationally rather expensive form. Several predictions for the performance of the approaches mainly based on the previously mentioned differences between approaches (e.g., consequences of different effect sizes used) were explicated and largely confirmed.

An overview of results is presented in Table 9.2. The table provides the results in the form of recommendations for applications of meta-analysis to correlational data. The recommendations in Table 9.2 only apply to applications of the approaches to correlation coefficients and may not be used for other effect size data. Of course, some of the cut-off values listed in the recommendations might seem arbitrarily chosen as it is naturally the case with most cut-off values in the methodological context. Nevertheless, the values have been chosen to reflect the results of the present study as closely as possible.

The table is structured according to the tasks to be performed in a meta-analysis and the situation given. Of course, the situation is something a meta-analyst does ordinarily not know in advance. The statements in the tables have thus to be interpreted as summaries of the performance of the various approaches in the Monte Carlo study and to give an indication which procedure is recommended when a certain situation is given.

As can be seen in the table, there is no single approach performing best for all tasks under all conditions. Instead, approaches seem to perform best overall when their basic model assumptions are met. For some of the tasks in meta-analysis specified in the table nearly all, and for some others none, of the approaches performs at an acceptable level according to conventional criteria. This indicates tasks and conditions for which the approaches evaluated here do not provide adequate statistical tools. This is the case, for example, for the homogeneity test Q in heterogeneous situations.

Table 9.2 Recommendations for Meta-Analysis of Correlational Data

Task	$\mathfrak{S}$	Recommendation
Estimation of μ_ρ	$\mathfrak{S}_1$	All estimators, except OP-FE and OP-RE, are usable when $n > 16$. However, OP shows *no* bias notwithstanding which n, k, or μ_ρ is given. OP is therefore recommended. HOT performs almost as well as OP and is more efficient when μ_ρ is small ($\mu_\rho < .10$). It can thus be considered as a good alternative in this situation.
	$\mathfrak{S}_2$	Only good r-based estimators should be used (OP and HS) to provide estimates of μ_ρ. Among these estimators OP shows *no* bias notwithstanding which n, k, or μ_ρ is given. OP-FE and OP-RE are not good choices. HS seems to be a good alternative to OP when $n > 32$. The estimate of μ_ρ should, however, be interpreted with caution when vastly different universe effect sizes are suspected. To determine whether this may be the case, a homogeneity test might be considered.
	$\mathfrak{S}_3$	Only good r-based estimators should be used (OP and HS) to provide estimates of μ_ρ. All estimators, except OP-FE and OP-RE are usable when $n > 32$. OP is preferable to all other estimators.
Significance tests for $\mu_\rho = 0$	$\mathfrak{S}_1$	DSL and HOT perform best by showing mean rejection rates below α. HOr shows rejection rates closest to α when the null hypothesis is true. Except for HS3, HS4 and OP-RE, all rejection rates are quite close to α, so the choice of test does not make a big difference. When the null hypothesis is false, all tests reach satisfactory power levels very quickly. The choice of a test does not make a substantial difference here as well.
	$\mathfrak{S}_2$	No substantial differences in power between approaches prevail. Random effects approaches are generally more conservative, though differences are marginal.
	$\mathfrak{S}_3$	When the null hypothesis is true, only random effects approaches (especially DSL) perform adequately. All other approaches show rejection rates far too high, even for moderate n (64) and σ^2_ρ (.01), and should not be used here. When the null hypothesis is false, there are only small disadvantages in power by using random effects approaches. Thus, DSL is recommended, deliberately accepting a disadvantage in power.

table continues

continued table

Task	$\mathfrak{S}$	Results and Recommendations
Confidence intervals for μ_ρ	$\mathfrak{S}_1$	HOT and OP reach the desired coverage rates most closely, and show a very stable performance across all levels of the design variables. Thus, both are recommended though with some reservations because of much larger interval widths when n and k are very small (i.e., less than 16). All other approaches, except OP-RE, HOd, HS3, and HS4 also show mean coverage values of about .93 for 95% confidence intervals and may also be useful when bearing this in mind.
	$\mathfrak{S}_2$	HOT and OP reach the desired coverage rates most closely and show a very stable performance across all levels of the design variables. Since only OP estimates μ_ρ, it is recommended.
	$\mathfrak{S}_3$	All approaches show at least some deficiencies and none can be recommended without reservations (note that only r-based estimators were evaluated). Amongst the evaluated approaches, HS3 and HS4 performed best.
Homogeneity test: Q	$\mathfrak{S}_1$	HOr and, with some reservations when $n < 32$, also HS are usable. The transformation of r to d leads to excessive rejection rates which strongly cautions against the use of the HOd approach here.
	$\mathfrak{S}_2$	All approaches show deficiencies in detecting small to medium effects, especially when n or k are small. Thus, reliance on the Q-test for a decision on the conduct of HLM-type procedures or for the choice of model (FE vs. RE) can be a risky business.
	$\mathfrak{S}_3$	None of the approaches show satisfactory power in detecting small to medium variances (.0025 to .0225), especially when n or k are small. Unless $k > 32$, tests are not reliable indicators of heterogeneity.
Homogeneity test: 75%- and 90%-rule	$\mathfrak{S}_1$	Both the 75%- and 90%-rule are not viable alternatives to the Q-test (see above). Rejection rates are generally too high in this situation.
	$\mathfrak{S}_2$	Both the 75%- and 90%-rule are not viable alternatives to the Q-test (see above). Rejection rates are too low unless $n > 64$ and heterogeneity variance is at least medium.

table continues

continued table

Task	$\mathfrak{S}$	Results and Recommendations
	$\mathfrak{S}_3$	Basically the same results as in $\mathfrak{S}_2$ emerged. Hence the same recommendations also apply here.
Estimation of σ^2_ρ	$\mathfrak{S}_1$	For very low k and n below 32 all estimators show strong overestimation. OP-RE is unusable for $n < 16$ in all cases. All estimators provide acceptable estimates for $n > 32$ and k not less than 32.
	$\mathfrak{S}_2$	For very low k and n below 32 DSL and OP-RE show high biases. For even modest $\Delta\rho$ both OP-RE and DSL should not be used. In general, HS performs best in $\mathfrak{S}_2$ though it should be used with some caution when $k < 16$.
	$\mathfrak{S}_3$	OP-RE performs generally poorly and should not be used in this situation. HS and DSL both perform well, but HS performs best.

Note. $\mathfrak{S}_1$ to $\mathfrak{S}_3$ = Assumed situation in meta-analysis.

Overall, the good performance of OP in various situations and for various purposes is remarkable. For estimating the mean effect size, for example, it can be recommended without reservations. However, Table 9.2 also indicates when it should used with strong reservations at best (construction of confidence intervals in $\mathfrak{S}_3$).

For the tasks of testing $\mu_\rho = 0$ and homogeneity tests, approaches do not differ markedly. In the former case they show equally good performance and in the latter they all perform equally badly. For homogeneity tests, the procedures unique to the HS approach are not interesting alternatives. For the purpose of estimating the heterogeneity variance, however, HS emerged as the best approach, though it should be added that DSL is hard to compare because computations and results are in z-space.

Finally, a caveat seems indicated. Note that it is *not* recommended in general to employ any of the methods in $\mathfrak{S}_2$ and to abstain from using HLM procedures. Since appropriate predictors are not always available to the meta-analyst, the methods of meta-analysis as described in this book are the only available option. Hence, an evaluation of their performance as provided here is of vital importance.