This is chapter 10: Discussion and conclusion (pp. 191-196) from

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

10
Discussion and Conclusions

Reviews of meta-analytical methods have generally been very positive, at least in the social sciences (e.g., Kavale, 1995). The ongoing debate about its usefulness as a scientific research tool (cf. Hunter & Schmidt, 1996; Feinstein, 1995) has not hampered its growth in the literature or the willingness to adopt it as a useful tool by researchers. Most critics argue not on purely statistical grounds but attack the application of meta-analytical methods for reasons founded in the philosophy of science or on conceptual grounds from the specific field of application. Some of these lines of criticism seem legitimate, indeed, and meta-analysis is certainly not free of conceptual problems and ambiguities in application. For example, it was pointed out in the introductory chapter that meta-analysis is not a strictly standardized technique for which clearly articulated rules of conduct are available at *any* step of the whole process. Reviews have shown that meta-analyses on the same issue do not provide nearly identical results but are quite different and variable (e.g., Steiner, Lane, Dobbins, Schnur, & McDonnell, 1991). Moreover, doubts have been raised regarding the reliability of implementing meta-analysis in practice (Zakzanis, 1998). As was pointed out by Wanous, Sullivan, and Malinak (1989), judgement calls are important and seem to influence the results and conclusions drawn in meta-analyses on the same topic. Thus, problems pertaining to the application of meta-analysis seem to mainly result because meta-analysis is more than just estimating parameters (see, e.g., Bailar, 1995).

The present examination focused on the *statistical* methods most common for meta-analysis of correlations, that is, the analysis step. This step is probably viewed by many as the immune core of meta-analysis, hence regarded as the step with the least problems or potential for subjective influences. A result of this may be the seemingly generally adopted assumption that it makes no difference which of the available sets of statistical procedures is used. The choice of an approach seems more a question of the field of research in which the methods are applied rather than a question of the statistical model assumed for a research situation. However, it was shown that there are many important implications for the results and therefore potentially also for the conclusions drawn from a meta-analysis due to the choice of one of the available approaches.

Interestingly, it is not an easy task to clearly answer the question of what approaches are actually available, because ambiguities arise in exactly specifying the available ones. One possibility to do this would be to focus only on major presentations of meta-analytic methods for correlations in the literature. This basically leads to three approaches (Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Rosenthal, 1991). In this book, these three major approaches were complemented by further approaches which are partly already included in the treatment by Hedges and Olkin (1985) or represent sometimes minor but consequential differences in statistical procedures. It seems legitimate to call into question such a concept of approaches or the meaningfulness of the very concept of approaches. A good example for not classifying approaches according to author groups or major treatments in the literature is given by comparison of HO*r* and RR.

The reason to classify RR differently in comparison to HO*r* is, in fact, a minor one at best, concentrated on a single aspect of significance testing for the mean effect size (compare procedures on page 60 and page 62). This is not regarded as a compelling reason to differentiate between HO*r* and RR. With the same reasoning it might be argued that HOT, for example, is also not an independent approach but represents only a minor change in the HO*r* procedures. Indeed, this is true. However, this argument extends to other approaches as well. It is argued that criteria to differentiate or classify *statistical approaches* in meta-analysis should better be based more on classes of statistical models and effect size measures, for example, rather than authors, books, or any other historical and seemingly arbitrary reason. This was done in the present book. A slightly extended list of classification aspects includes:

- effect size measure used,
- weighting scheme used,
- FE versus RE models, including conditional RE approaches,
- explanatory (e.g., HLM) versus non-explanatory models, and
- use of observed and/or latent variables (HLM vs. mixture analysis).

These aspects may even be extended by some models not presented in this book, for example, (empirical) Bayes models (see, e.g., Raudenbush & Bryk, 2002). Although such classification aspects are partly overlapping, they enable a distinction between meta-analytic procedures more in line with common statistical distinctions. These aspects also show that there is a wealth of models and procedures for the meta-analysis of correlations that let statements like "...there is only one dominant approach for conducting meta-analysis of correlation research and that is the Hunter and Schmidt (1990) approach" (Huffcutt, 2002, p. 209) appear untenable. In retrospect then, the use of *approaches* in this book is only a vehicle to differentiate between statistical procedures and does not necessarily designate fundamentally different routes to meta-analysis of correlations.

Of the classification aspects listed above, the choice of effect size measure seems to be one accompanied by more fundamental consequences than previously thought. As has been pointed out, r-based approaches do not in principle suffer from changes in estimated parameters in heterogeneous situations. The change in spaces by the Fisher-z or r to d transformation of the original correlation has manifest consequences for interpretation many users of meta-analysis may not be aware of. It was shown that the use of the Fisher-z transformation leads to higher absolute estimates of $\mu_{\rho z}$ as compared to μ_ρ in heterogeneous situations. Interpreting estimates of $\mu_{\rho z}$ as estimates of μ_ρ would simply be a misinterpretation in heterogeneous situations.

It is difficult to assess the severity of this problem in previous practical applications of meta-analysis at least for two reasons. First, the difference between the universe parameters μ_ρ and $\mu_{\rho z}$ estimated in the approaches based on the Fisher-z transformation versus r-based approaches depends on the unknown heterogeneity in the universe of studies. To quantify the difference it would be necessary to know exactly the categorical or continuous distribution in the universe. Such knowledge is, of course, not available and it would be interesting to reanalyze existing meta-analytic databases to examine the differences arising in practice. Second, even if the difference could be quantified, severity is a very subjective aspect. For example, in a situation with a beta distributed random variable with $\mu_\rho = .60$ and $\sigma^2_\rho = .0625$ in the universe of studies, a corresponding $\mu_{\rho z} \approx .67$ is given. The difference of .07 would certainly be judged by some researchers for a certain research question — for example in the personnel selection context — as substantial and in the context of other research questions it might not change interpretation of results and therefore be inconsequential.

Hence, doubts are raised as to whether the Fisher-z transformation should be applied to correlation coefficients in meta-analysis. Arguments put forward in favor of its use often rest on highlighting the bias of r (e.g., Silver & Dunlap, 1987). As was shown in the current book, differences in bias favor r over z but are minuscule in absolute value, anyway. Moreover, in light of the fact that an UMVU estimator G is available, easily computed, and shows excellent performance in terms of bias as reported in the Monte Carlo study, arguments in favor of Fisher-z which are based on the bias of r are not convincing.

Another line of argument against the use of r draws on the dependency of the variance of the estimators on the universe parameter (e.g., James et al., 1986). This is indeed a serious issue not only for r but also for other estimators in meta-analysis and therefore represents a *general* problem for pooled estimators. The optimal weights require the correct variances of the estimators. Since only estimates of these variances are available in practice and these estimates are plugged in the weights in aggregation, a dependency of the variance on the universe parameters, or more precisely on the estimator when estimates are plugged in, induces a bias in the pooled estimator, especially when n is small. This was most clearly evident in the Monte Carlo study for OP-FE and OP-RE, the UMVU estimator weighted by the inverse of its (estimated) variances. Note that the problem not only pertains to these estimators. Since this problem does not arise with suboptimal weights that depend only on n, the use of G weighted by the sample sizes of the studies is recommended here when precise estimation of the universe parameter is of vital interest, as is nearly always the case in meta-analysis of correlations. As may be noted, the approach proposed by Hunter and Schmidt (1990) is also an r-based approach with the sample sizes as weights. The usage of this approach is thus also encouraged. Nevertheless, a better choice than r is to use the UMVU estimator. The recommended approach based on the UMVU estimator is not without problems. There are also certain tasks in meta-analysis for which the approach — as it was specified — does not perform satisfactorily, for example, for testing $\mu_\rho = 0$ in $\mathfrak{S}_3$. In consequence, there is no single best approach amongst the set of examined approaches. Such an approach has yet to be developed. However, taking into account the many possible situations, many tasks, and many boundary conditions (e.g., with respect to n and k) in meta-analysis, it seems unlikely that such a single approach will ever become available.

Problems in interpreting a mean effect size estimate in meta-analysis not only arise in the context of transformations of the correlation coefficient. Interpretation also depends on whether heterogeneity in universe effect sizes is present at all, detected, and modeled. In general, mean effect sizes have an undisputable interpretation in homogeneous situations but not in heterogeneous situations. This does not mean, however, that they are not interpretable in heterogeneous situations. As has been argued, the mean effect size generally has to be interpreted like the grand mean in ANOVA-type analyses. Of course, if heterogeneity is suspected or detected by any of the available tests, then models to explain heterogeneity (e.g., HLM) are certainly indicated to go beyond grand mean interpretations.

In any case, the interpretation of results in meta-analysis has to be done within the framework of a chosen model, another characteristic to differentiate between approaches. Unfortunately, the choice of a model is often done in practice just en passant. As has also been shown with other methods as those used in this book (Hedges & Vevea, 1998) and in different contexts (Overton, 1998), methods generally perform best when their model assumptions are met. This conclusion seems trivial at first glance, but in light of the fact that many of the statistical derivations of procedures used in meta-analysis rest on large-sample theory, it is important to test by simulation methods whether the properties of estimators, for example, also hold for constellations of design characteristics likely to arise in practice. Unfortunately, this information is not of great help for the meta-analyst, who wants to decide which model to adopt, though it is certainly reassuring. A theoretically founded line of reasoning may lead researchers to the choice of a model. Questions about the intended inference, theoretically expected heterogeneity, or simply the number and origin of available studies help in deciding which model to adopt.

Another possibility is to condition the choice of the model on the result of the Q-test. As was shown in the Monte Carlo study as well as in the literature (e.g., Harwell, 1997), the Q-test to detect heterogeneity is not satisfactorily powerful in many situations and heterogeneity may therefore remain undetected. This test is thus not a very good guide for a model decision because it leads to many wrong decisions. Hence, statements like "if the chi square is not significant, this is strong evidence that there is no true variation across studies, but if it is significant, the variation may still be negligible in magnitude" (Hunter & Schmidt, 1990, p. 112) are questionable (see also Harwell, 1997). Proposed alternatives to this test, like the 75%- or 90%-rule do not represent viable alternatives to the Q-test (see also Sánchez-Meca & Marín-Martínez, 1997). Interestingly, the 75%-rule seems to be in widespread use, at least in I/O psychology. Cortina (2003) reviewed 59 quantitative reviews containing not less than 1,647 meta-analyses, of which all appeared in one of the most prestigious journals of I/O psychology, the Journal of Applied Psychology. He found that as many as 57% of the meta-analyses used the 75%-rule as a homogeneity test and only 19% the Q-statistic. Thus, further theoretical developments as well as their empirical evaluation to establish procedures that perform better for this task of meta-analysis are needed. Hartung and Knapp (2003), for example, recently proposed such an alternative test procedure for meta-analysis.

Yet another option would be to explore heterogeneity by application of mixture models (Böhning, 2000; Schlattmann et al., 2003). These models provide a statistically well-founded framework for meta-analysis that is not widely used yet. Though early presentations of these techniques have been given in the psychological literature (Thomas, 1989b; Thompson, 1989; Thomas, 1990b), they have not been adopted very often. The reasons for this fact may lie in unfamiliarity with these models or in perceived technical difficulties. Since easy-to-use software for the application of these models has recently become available (Böhning et al., 1992; Schlattmann et al., 2003), their use is encouraged because they address one of the central questions of meta-analysis quite elegantly, the modeling of heterogeneity.

Apart from suggesting to condition the use of a model on the outcome of a homogeneity test — a so-called conditional random effects procedure — Hedges and Vevea (1998) have proposed to make a choice between the FE and RE model on the basis of the intended inference. The intended inference is a question about properties of the universe of studies to which results are generalized to. These properties may be restricted to characteristics like those of the observed studies (FE model) or more general (RE model). The question of intended inference is not always an easy question to answer since generalization not only depends on the desire of a researcher as Hunter and Schmidt (2000) suggest, but also on a series of other aspects, like those Matt (2003) has described, for example. The shift from applications of FE models to RE models that is strongly encouraged in the literature (e.g., Erez et al., 1996) is not without problems, as the presented Monte Carlo study results suggest. Especially when the number of studies is small ($k < 32$), the most important aspect of RE models, the heterogeneity variance, can not be estimated with acceptable precision. Note that a number of 32 studies is far from unusual in practice and even in some Monte Carlo studies considered to be large (e.g., Field, 2001).

Another aspect addressed in the present Monte Carlo study with potentially far-reaching implications is the conversion of correlations to standardized mean differences *d*. Transformations of effect sizes are necessary in most applications because of different designs and analysis methods used in the primary studies to address the same research question. The implicit assumption of applying the transformation is that computations based on the transformed effect size (e.g., *d* from *r*) lead to equivalent results in comparison to computations based on the untransformed effect size (e.g., *r*). In other words, the transformation does not introduce any bias or distortion of results. If the equivalence were given, then it would not matter whether meta-analytic computations were carried out with *r* or *d* as an effect size, the results would be the same. However, the *r* to *d* transformation leads to changes in results in meta-analysis as reported in the Monte Carlo study. This clearly challenges the assumption of an inconsequential application of this transformation. Since the influences of weights that depend on the universe parameter are also involved in explanations of results, the origin of the deviant results by using *d* is not entirely clear. The derivation of the transformation formula, however, rests on assumptions that seem questionable. Of course, as has repeatedly been highlighted, there would be no need to apply the transformation to a database consisting only of *r* in practice. Instead, this would be ordinarily necessary only for a subset of studies. The results presented in this book suggest that it is wise to at least conduct a sensitivity analysis to assess the effect of the transformed effect sizes on the results.

To conclude, the choice of an approach to meta-analytically synthesize correlation coefficients as presented in this book does make a difference. Some approaches are better than others for various tasks but a single best set of procedures has yet to be established. The present book has nevertheless pointed out some procedures that should be used with caution and others that seem under-utilized and deserve more attention in methodological developments and applications.