Citation:

Hartung, J. & Knapp, G. (2003). An alternative test procedure for meta-analysis. In R. Schulze, H. Holling & D. Böhning (Eds.), *Meta-analysis: New developments and applications in medical and social sciences* (pp. 53-69). Hogrefe & Huber.

4

An Alternative Test Procedure for Meta-Analysis

Joachim Hartung and Guido Knapp
Department of Statistics[†]
University of Dortmund

Summary

In this chapter, we show that a meta-analysis carried out in the random effects model is preferable to the fixed effects model. Especially in the normal mean case, as our simulation study indicates, the test of association in the FE model does not yield satisfactory results. If one prefers to use the commonly used methods, the choice between the FE and the RE model, which leads to the choice of the test statistic for the hypothesis of no association, is better based on the sign of the method of moments estimator of the between-study variance than on the test of homogeneity. But the use of the alternative test statistic originally proposed in (Hartung, 1999) is preferable concerning the significance level to all commonly used methods. The test is always carried out in the RE model, but it yields sufficiently good results if no heterogeneity is present. So, one does not have to choose between the FE and the RE model in advance. In the case of a small between-study variance, a combined test procedure involving the commonly used test in the FE model and Hartung's alternative test statistic may still improve the actual significance level of the test towards the prescribed one.

[†]Project "Meta-Analysis in Biometry and Epidemiology" (SFB 475) of the Deutsche Forschungsgemeinschaft (DFG).

4.1 INTRODUCTION

In this chapter, we focus our attention on the tests of association in the meta-analytic framework, that is, we want to judge if an overall treatment effect exists given the stochastically independent study-specific estimates of the treatment effect. This test is carried out either in the fixed effects model of meta-analysis assuming a homogeneous treatment effect over the studies or in the random effects model if heterogeneity of the study-specific treatment effects is present. Before applying the test of association one usually decides which of the two models one takes. Even in recent literature one can find the proposal that the choice of the model should be based on the test of homogeneity, cf. for instance Normand (1999). But the test of homogeneity in this context often has too low power to detect a deviation from the hypothesis of homogeneity and the false use of the fixed effects model, if heterogeneity is present, can lead to a dramatic increase of the Type I error rate of the commonly used test of association as pointed out for instance in Ziegler and Victor (1999). Moreover, the commonly used tests of association in the fixed effects model and in the random effects model, respectively, may lead to a large number of unjustified significant evidences even if one carries out the analysis in the correct model. In the fixed effects model this was shown by Li, Shi, and Roth (1994) and also Böckenhoff and Hartung (1998) in the normal mean case.

We will now consider an alternative test statistic for the test of association in the random effects model of meta-analysis proposed by Hartung (1999) and show that this test provides satisfactory results concerning the actual significance level in the fixed effects model as well as in the random effects model, so that with this test a choice between the two models, in advance, is unnecessary. Furthermore, we will discuss decision rules for the test of association, which combines the commonly used tests and this alternative test, and investigate these decision rules, whether they yield a further improvement with respect to the prescribed significance level.

The outline of the chapter is as follows: In the next two sections we first describe the theoretical foundations of the meta-analysis in a fixed effects and a random effects approach, respectively. In Section 4.4, the commonly used methods for a practical application in the fixed effects and random effects model are presented. Section 4.5 contains some results of the theoretical deficiency of the commonly used tests in the both models. In Section 4.6, the alternative test statistic in the random effects model proposed by Hartung (1999) is presented and in Section 4.7 some decision rules are discussed, which combine the commonly used tests and the alternative test. In a simulation study, of which the results are given in Section 4.8, the discussed tests are compared concerning their actual significance levels in the normal mean case. Finally, some conclusions are given.

4.2 THE HOMOGENEOUS FIXED EFFECTS MODEL

Let us consider k independent studies and let us denote by $\theta_1, \ldots, \theta_k$ the (one-dimensional) parameters of interest, where each parameter stands for the treatment effect in a study. For example, the parameter θ_i, $i = 1, \ldots, k$, may represent the mean and the standardized difference of means, respectively, for continuous outcome variables or the risk difference, the logarithmic odds ratio, and the relative risk, respectively, for binary outcome variables. In each study an estimate of the parameter θ_i, say $\hat{\theta}_i$, is available, and all study-specific estimators $\hat{\theta}_i$, $i = 1, \ldots, k$, are stochastically independent. Assuming that the parameters of interest are fixed and homogeneous, that is, it holds $\theta_1 = \cdots = \theta_k = \theta$, and the study specific estimators $\hat{\theta}_i$ are at least approximately normally distributed and unbiased or at least consistent, then the so-called (homogeneous) fixed effects model (FE model) of meta-analysis is given by

$$\hat{\theta}_i \sim \mathcal{N}\left(\theta, \sigma^2(\hat{\theta}_i)\right), \quad i = 1, \ldots, k, \tag{4.1}$$

where $\sigma^2(\hat{\theta}_i)$ denotes the variance of the estimator $\hat{\theta}_i$ in the ith study.

In model 4.1, the best linear unbiased estimator of the common mean θ is given by

$$\tilde{\theta}_{FE} = \sum_{i=1}^{k} \frac{v_i}{v} \hat{\theta}_i, \quad v = \sum_{i=1}^{k} v_i, \tag{4.2}$$

with $v_i = [\sigma^2(\hat{\theta}_i)]^{-1}$ the inverse of the variance of the study-specific estimator $\hat{\theta}_i$ in the ith study. The estimator $\tilde{\theta}_{FE}$ is also the maximum likelihood estimator (MLE) of θ in model 4.1 if the normal distribution exactly holds and the variances $\sigma^2(\hat{\theta}_i)$ are known.

The assumption of homogeneity of the parameters can formally be checked using the test statistic

$$Q = \sum_{i=1}^{k} v_i \left(\hat{\theta}_i - \tilde{\theta}_{FE}\right)^2, \tag{4.3}$$

which is at least approximately χ^2-distributed with $(k-1)$ degrees of freedom under the hypothesis of homogeneity (Cochran, 1954; Normand, 1999).

If all assumptions in model 4.1 are fulfilled, the estimator $\tilde{\theta}_{FE}$ from Equation 4.2 has the following distributional property:

$$\tilde{\theta}_{FE} \sim \mathcal{N}\left(\theta, \frac{1}{v}\right). \tag{4.4}$$

So, from 4.4 an (approximate) $(1-\alpha)$-confidence interval for the common parameter θ is given by $\tilde{\theta}_{FE} \mp u_{1-\alpha/2}/\sqrt{v}$, where u_γ denotes the γ-quantile of the standard normal distribution. Furthermore, the two-sided test rejects the

hypothesis of no association $H_0 : \theta = 0$ at level α if

$$\frac{\left(\tilde{\theta}_{FE}\right)^2}{1/v} = \frac{\left(\sum_{i=1}^k v_i \hat{\theta}_i\right)^2}{v} > \chi^2_{1;1-\alpha},$$

where $\chi^2_{\nu;\gamma}$ denotes the γ-quantile of the χ^2-distribution with ν degrees of freedom.

If the assumption of homogeneity is not valid in model 4.1, that is, it exists at least one pair $\theta_i \neq \theta_j$, $i \neq j$, then the estimator $\tilde{\theta}_{FE}$ from Equation 4.2 is still an unbiased estimator of a weighted average of the θ_i's, namely of $\sum_{i=1}^k v_i \theta_i / v$. So, the above described confidence interval and test can always be used for this weighted average of the parameters. But the usual proceeding, if the hypothesis of homogeneity is not valid, is either to try to identify covariates which stratify studies into homogeneous populations or to carry out the meta-analysis in a random effects model (Normand, 1999). In the next section we will consider the latter proposal.

4.3 THE RANDOM EFFECTS MODEL

In contrast to the homogeneous fixed effects model 4.1, we first allow that the study-specific estimators $\hat{\theta}_i$, $i = 1, \ldots, k$, may possess different expected values θ_i, $i = 1, \ldots, k$, that is, it holds approximately

$$\hat{\theta}_i | \theta_i, \quad \sigma^2(\hat{\theta}_i) \sim \mathcal{N}\left(\theta_i, \sigma^2(\hat{\theta}_i)\right), \quad i = 1, \ldots, k,$$

and for each study-specific mean θ_i we assume that it is drawn from some superpopulation of effects with mean θ and variance τ^2, that is,

$$\theta_i | \theta, \quad \tau^2 \sim \mathcal{N}\left(\theta, \tau^2\right).$$

The parameters θ and τ^2 are referred to as hyperparameters, θ represents the average treatment effect and τ^2 the between-study variation. Given the hyperparameters, the marginal distribution of the estimators $\hat{\theta}_i$ is given by

$$\hat{\theta}_i \sim \mathcal{N}\left(\theta, \tau^2 + \sigma^2(\hat{\theta}_i)\right), \quad i = 1, \ldots, k, \tag{4.5}$$

(cf. Whitehead & Whitehead, 1991; Normand, 1999). If the between-study variance τ^2 is equal to zero then the random effects model (RE model) 4.5 reduces to the FE model 4.1.

In the RE model 4.5, the best linear unbiased estimator of the average treatment effect θ is given by

$$\tilde{\theta}_{RE} = \sum_{i=1}^k \frac{w_i}{w} \hat{\theta}_i, \quad w = \sum_{i=1}^k w_i$$

with $w_i = [\tau^2 + \sigma^2(\hat{\theta}_i)]^{-1}$ the inverse of the variance of the estimator $\hat{\theta}_i$ in the RE model. The estimator $\tilde{\theta}_{RE}$ is also the MLE of θ if the variance components are known and the normal distribution in 4.5 exactly holds.

The estimator $\tilde{\theta}_{RE}$ in model 4.5 possesses the following distributional property:

$$\tilde{\theta}_{RE} \sim \mathcal{N}\left(\theta, \frac{1}{w}\right).$$

Thus, an $(1-\alpha)$-confidence interval for the average treatment effect θ in the RE model is given by $\tilde{\theta}_{RE} \mp u_{1-\alpha/2}/\sqrt{w}$ and the hypothesis of no association, that is, $H_0 : \theta = 0$, is rejected at level α if

$$\frac{\left(\tilde{\theta}_{RE}\right)^2}{1/w} = \frac{\left(\sum_{i=1}^k w_i \hat{\theta}_i\right)^2}{w} > \chi^2_{1;1-\alpha}.$$

4.4 THE COMMONLY USED METHODS IN THE FE AND RE MODEL

In the previous two sections we have summarized the theoretical aspects of the FE and RE model of meta-analysis. For a practical application of the just described inference the involved variances τ^2 and $\sigma^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, are hardly ever known. So, they have to be replaced by appropriate estimators.

Let us first consider the FE model. Normally, an estimator of the variance of $\hat{\theta}_i$, say $\hat{\sigma}^2(\hat{\theta}_i)$, is available in each study. We assume that these estimators $\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, are jointly stochastically independent and at least nearly unbiased for the corresponding variances $\sigma^2(\hat{\theta}_i)$. Using these variance estimators, the feasible estimator of θ in model 4.1 is given by

$$\hat{\theta}_{FE} = \sum_{i=1}^k \frac{\hat{v}_i}{\hat{v}} \hat{\theta}_i, \quad \hat{v} = \sum_{i=1}^k \hat{v}_i, \quad \hat{v}_i = \left[\hat{\sigma}^2(\hat{\theta}_i)\right]^{-1}, \quad i = 1 \ldots, k.$$

In general, this estimator is not an unbiased one of θ. If the estimators $\hat{\theta}_i$ and the variance estimators $\hat{\sigma}^2(\hat{\theta}_i)$ are stochastically independent, which implies that $\hat{v}_i/\hat{v}$ as a function of $\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, is also stochastically independent of $\hat{\theta}_i$, then the estimator $\hat{\theta}_{FE}$ is unbiased for θ in model 4.1. This can be readily seen with $\sum_{i=1}^k \hat{v}_i/\hat{v} = 1$.

For further inference in the FE model, the variance estimators are commonly inserted in the test statistics and the confidence interval given in Section 4.2. Thus, the assumption of homogeneity of the treatment effects in model 4.1 is checked in practice with the test statistic

$$Q_1 = \sum_{i=1}^k \hat{v}_i \left(\hat{\theta}_i - \hat{\theta}_{FE}\right)^2. \tag{4.6}$$

The test statistic for testing the hypothesis of no association is given by

$$T_1 = \frac{\left(\hat{\theta}_{FE}\right)^2}{1/\hat{v}} = \frac{\left(\sum_{i=1}^k \hat{v}_i\hat{\theta}_i\right)^2}{\hat{v}}, \tag{4.7}$$

and the hypothesis is rejected at level α if the observed value of T_1 exceeds the $(1-\alpha)$-quantile of the χ^2-distribution with one degree of freedom.

In the RE model, besides the estimates of the within-study variances $\sigma^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, an estimator of the between-study variance τ^2 has to be deduced. One approach is to use the method of moments estimator, which is derived using the test statistic of homogeneity Q from Equation 4.3. The expected value of Q in the RE model is given by

$$E(Q) = (k-1) + \tau^2\left(v - \sum_{i=1}^k v_i^2/v\right)$$

(cf. DerSimonian & Laird, 1986; Whitehead & Whitehead, 1991). So, the method of moments estimator reads

$$\tilde{\tau}^2 = \frac{Q-(k-1)}{v - \sum_{i=1}^k v_i^2/v}. \tag{4.8}$$

Due to its construction, the estimator $\tilde{\tau}^2$ is unbiased but it can yield negative estimates with positive probability. Moreover, the estimator depends on the unknown variances $\sigma^2(\hat{\theta}_i)$. Thus, for a practical application the feasible estimator of τ^2 is given by

$$\hat{\tau}^2 = \frac{Q_1-(k-1)}{\hat{v} - \sum_{i=1}^k \hat{v}_i^2/\hat{v}}, \tag{4.9}$$

with Q_1 from Equation 4.6, and usually the truncated version of this estimator is used, namely,

$$\hat{\tau}_+^2 = \max\{0, \hat{\tau}^2\}. \tag{4.10}$$

The larger the between-study variance τ^2, the less the probability of $\hat{\tau}^2$ yielding negative estimates. So, for large τ^2 both estimators in Equations 4.9 and 4.10 are nearly identical. Note that feasibility in this sense does not mean unbiasedness for all τ^2.

Other approaches of estimating the between-study variance τ^2 are to use the restricted maximum likelihood approach or a Bayesian approach (Normand, 1999). But we will not consider these approaches in the context of this chapter.

With the between-study variance estimator $\hat{\tau}_+^2$ and the within-study variance estimators $\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, the feasible estimator of the average treatment effect θ in the RE model is given by

$$\hat{\theta}_{RE} = \sum_{i=1}^k \frac{\hat{w}_i}{\hat{w}}\hat{\theta}_i, \quad \hat{w} = \sum_{i=1}^k \hat{w}_i, \quad \hat{w}_i = \left[\hat{\tau}_+^2 + \hat{\sigma}^2(\hat{\theta}_i)\right]^{-1}, \; i = 1\ldots, k.$$

So, the test statistic for testing the hypothesis of no association in the RE model is given by

$$T_2 = \frac{\left(\hat{\theta}_{RE}\right)^2}{1/\hat{w}} = \frac{\left(\sum_{i=1}^{k} \hat{w}_i \hat{\theta}_i\right)^2}{\hat{w}}, \tag{4.11}$$

and the hypothesis is rejected at level α if the observed value of T_2 is larger than the $(1-\alpha)$-quantile of the χ^2-distribution with one degree of freedom.

4.5 THE THEORETICAL DEFICIENCY OF THE COMMONLY USED TESTS IN THE FE AND RE MODEL

In the previous section, we have presented the commonly used test statistics in the FE and RE model (see Equations 4.7 and 4.11) for testing the hypothesis of no association as well as the rule of rejection at a prescribed significance level α. But as already pointed out in Li et al. (1994) and Böckenhoff and Hartung (1998), the actual significance level of the commonly used test in the FE model with normally distributed responses is often much larger than the prescribed level α due to underestimation of the variance of the combined estimator $\tilde{\theta}_{FE}$ so that this phenomenon results in a large number of unjustified significant evidences.

We now summarize the main theoretical results of the work of Böckenhoff and Hartung in the FE model and indicate that the same deficiency also holds in the RE model. First, we need some mathematical tools from Hartung (1976).

Definition 4.5.1. A function $f : \mathbb{R}^k \to \mathbb{R}^\ell$ is called convex if

$$\Big(x, y \in \mathbb{R}^k, \lambda \in [0,1] \Rightarrow f\left(\lambda x + (1-\lambda)y\right) \leq \lambda f(x) + (1-\lambda) f(y)\Big)$$

with the natural semi-ordering, that is,ordering by components.

Definition 4.5.2. A function $f : \mathbb{R}^k \to \mathbb{R}^\ell$ is called quasi-convex if

$$\Big(y \in \mathbb{R}^\ell \Rightarrow \{x \in \mathbb{R}^k | f(x) \leq y\} \text{ is convex}\Big).$$

Definition 4.5.3. A function f is called (quasi-)concave if $(-f)$ is (quasi-)convex.

Lemma 4.5.1. *Let $f : \mathbb{R}^k \to \mathbb{R}^\ell$ be convex [concave] and $T : \mathbb{R}^\ell \to \mathbb{R}^m$ be (quasi-) convex [(quasi-)concave] and increasing by the natural semi-ordering in $\mathbb{R}^m$. Then the composed function $T \circ f$ is (quasi)-convex [(quasi)-concave].*

Lemma 4.5.2. *Let $f : \mathbb{R}^k \to \mathbb{R}^\ell$ be convex [concave] and $T : \mathbb{R}^\ell \to \mathbb{R}^m$ be (quasi-) concave [(quasi-)convex] and decreasing by the natural semi-ordering in $\mathbb{R}^m$. Then the composed function $T \circ f$ is (quasi)-concave [(quasi)-convex].*

Lemma 4.5.3. *If $f : \mathbb{R}_+^\ell \to \mathbb{R}_+$ is quasi-convex [quasi-concave] and $f(\lambda\, x) = \lambda\, f(x)$, $\lambda > 0, x \neq 0$, then f is also convex [concave].*

Jensen's Inequality. For a random variable $\hat{\vartheta}$ it holds
$E(f(\hat{\vartheta})) \geq f(E(\hat{\vartheta}))$ if f is convex, and
$E(g(\hat{\vartheta})) \leq g(E(\hat{\vartheta}))$ if g is concave.

We now return to the meta-analysis and consider the variance estimator $1/\hat{v}$ in the FE model. Note, that in the previous section we have assumed that the within-study estimators $\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, are (nearly) unbiased for $\sigma^2(\hat{\theta}_i)$, that is, $\mathrm{E}(\hat{\sigma}^2(\hat{\theta}_i)) = \sigma^2(\hat{\theta}_i)$, $i = 1, \ldots, k$. So, we can prove the following theorem.

Theorem 4.5.4. *In the FE model the variance estimator $1/\hat{v}$ in average underestimates the variance $1/v$, that is, $E(1/\hat{v}) \leq 1/v$.*

Proof. First consider $\hat{\sigma}^2(\hat{\theta}_i) > 0$, $i = 1, \ldots, k$, then $1/\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \ldots, k$, is a convex function in $\hat{\sigma}^2(\hat{\theta}_i)$.

Furthermore, $\sum : \mathbb{R}^k \to \mathbb{R}$ is a convex and increasing function. So, with Lemma 4.5.1 the function $\sum_{i=1}^k 1/\hat{\sigma}^2(\hat{\theta}_i)$ is convex in $\hat{\sigma}^2(\hat{\theta}_i)$, $i = 1, \ldots, k$.

With Lemma 4.5.2, we yield that the function $\hat{v}^{-1} = \left(\sum_{i=1}^k 1/\hat{\sigma}^2(\hat{\theta}_i)\right)^{-1}$ is quasi-concave, because every monotone function is quasi-convex as well as quasi-concave.

Now, it holds for $\lambda > 0$ that

$$\left(\sum_{i=1}^k \frac{1}{\lambda\hat{\sigma}^2(\hat{\theta}_i)}\right)^{-1} = \left(\frac{1}{\lambda}\sum_{i=1}^k \frac{1}{\hat{\sigma}^2(\hat{\theta}_i)}\right)^{-1} = \lambda\left(\sum_{i=1}^k \frac{1}{\hat{\sigma}^2(\hat{\theta}_i)}\right)^{-1},$$

that means with Lemma 4.5.3 that the function $\left(\sum_{i=1}^k 1/\hat{\sigma}^2(\hat{\theta}_i)\right)^{-1}$ is concave. Applying Jensen's inequality, we obtain

$$E\left(\hat{v}^{-1}\right) = E\left(\sum_{i=1}^k \frac{1}{\hat{\sigma}^2(\hat{\theta}_i)}\right)^{-1} \leq \left(\sum_{i=1}^k \frac{1}{E(\hat{\sigma}^2(\hat{\theta}_i))}\right)^{-1} = \left(\sum_{i=1}^k \frac{1}{\sigma^2(\hat{\theta}_i)}\right)^{-1} = v^{-1},$$

which completes the proof. □

As a consequence of Theorem 4.5.4, we obtain that the distribution of the test statistic T_1 from Equation 4.7 under $H_0 : \theta = 0$ may not be well approximated by a χ^2-distribution with one degree of freedom. Suppose that the feasible estimator $\hat{\theta}_{FE}$ has variance near $1/v$ under H_0, then the probability of T_1 under H_0 to exceed the $(1-\alpha)$-quantile of the χ^2_1-distribution is larger than α, so that the actual significance level of the test is larger than the prescribed one. Such an attitude of the test based on T_1 has been observed in the normal mean case as already mentioned by Li et al. (1994) as well as by Böckenhoff and Hartung (1998). In the latter work, some alternative estimators of the variance of $\tilde{\theta}_{FE}$ are discussed in the normal mean case, which in average overestimate the variance of $\tilde{\theta}_{FE}$ and thereby lead to a correction of the actual significance

level towards the prescribed one. This procedure does not, in general, result in a conservative attitude of the test, as one may expect at first sight, because one has to keep in mind that the variance of the theoretical estimator $\tilde{\theta}_{FE}$ is estimated and not the variance of the feasible estimator $\hat{\theta}_{FE}$. Thus, even with an overestimation of the variance $\sigma^2(\hat{\theta}_i)$ in each study one may obtain too many unjustified significant results but in a smaller number than with the commonly used method (Böckenhoff & Hartung, 1998).

The above argumentation mainly holds if the variance estimator $1/\hat{v}$ underestimates the variances of the feasible estimator $\hat{\theta}_{FE}$. But if the variance estimator $1/\hat{v}$ overestimates the variance of $\hat{\theta}_{FE}$, the resulting test can be very conservative as Hartung and Knapp (2001) have observed, for example, in the log odds ratio case.

It is worthwhile to note that the just described deficiencies of the commonly used test statistic in the FE model are most striking if the sample sizes in the studies are small to moderate depending on the choice of the parameter of interest.

In the RE model, a similar result as in Theorem 4.5.4 can be stated for testing the hypothesis of no association using the test statistic T_2 from Equation 4.11. Suppose we use the untruncated unbiased estimator $\tilde{\tau}^2$ from Equation 4.8 of the between-study variance in T_2, then we can prove, following the lines of the proof of Theorem 4.5.4, that the variance estimator $1/\tilde{w}$, $\tilde{w}_i = [\tilde{\tau}^2 + \hat{\sigma}^2(\hat{\theta}_i)]^{-1}$, $i = 1, \ldots, k$, $\tilde{w} = \sum_{i=1}^{k} \tilde{w}_i$, in average underestimates the variance of the theoretical estimator $\tilde{\theta}_{RE}$. Thus, the test in the RE model may be rather anticonservative if the variance of the feasible estimator $\hat{\theta}_{RE}$ is near $1/w$. But on the other hand, the test can be very conservative if $1/\tilde{w}$ overestimates the variance of $\hat{\theta}_{RE}$. If the truncated estimator $\hat{\tau}_+^2$ from Equation 4.10 is used in the test statistic T_2 in practice, the expected value of the variance estimator is larger than the expected value of the variance estimator with the untruncated estimator $\hat{\tau}^2$ from Equation 4.9, but for growing τ^2 this difference diminishes.

4.6 AN ALTERNATIVE TEST STATISTIC IN THE FE AND RE MODEL

In Section 4.4, we have estimated the variances of the theoretical estimators $\tilde{\theta}_{FE}$ and $\tilde{\theta}_{RE}$ in the FE and RE model by estimating their components separately. Now we consider an estimator of the variance of $\tilde{\theta}_{RE}$ in the RE model 4.5 following the theory of variance components estimation (cf. Rao, 1972; Hartung, 1981). This estimator is a quadratic function of the study-specific estimators $\hat{\theta}_i$, $i = 1, \ldots, k$, and is given in the RE model 4.5 as

$$\tilde{\sigma}^2(\tilde{\theta}_{RE}) = \frac{1}{k-1} \sum_{i=1}^{k} \frac{w_i}{w} \left(\hat{\theta}_i - \tilde{\theta}_{RE}\right)^2,$$

(cf. Hartung, 1999). Note that for $\tau^2 = 0$ the quadratic function $w \cdot (k-1) \cdot \tilde{\sigma}^2(\tilde{\theta}_{RE})$ coincides with the statistic Q from Equation 4.3, which is formally used in the FE model for checking the assumption of homogeneity.

Theorem 4.6.1. *The estimator $\tilde{\sigma}^2(\tilde{\theta}_{RE})$ is an unbiased estimator of the variance of $\tilde{\theta}_{RE}$ in the RE model 4.5.*

Proof. It holds

$$\begin{aligned} E\left(\hat{\theta}_i\right)^2 &= \mathrm{Var}\left(\hat{\theta}_i\right) + \left[E\left(\hat{\theta}_i\right)\right]^2 = w_i^{-1} + \theta^2, \\ E\left(\tilde{\theta}_{RE}\right)^2 &= w^{-1} + \theta^2, \\ E\left(\hat{\theta}_i \tilde{\theta}_{RE}\right) &= \frac{w_i}{w} E\left(\hat{\theta}_i\right)^2 + \sum_{j \neq i}^{k} \frac{w_j}{w} E\left(\hat{\theta}_i\right) E\left(\hat{\theta}_j\right) \\ &= w^{-1} + \theta^2. \end{aligned}$$

Then it follows

$$E\left(\hat{\theta}_i - \tilde{\theta}_{RE}\right)^2 = w_i^{-1} + \theta^2 - 2w^{-1} - 2\theta^2 + w^{-1} + \theta^2 = w_i^{-1} - w^{-1}$$

and

$$E\left(\tilde{\sigma}^2\left(\hat{\theta}_{RE}\right)\right) = \frac{1}{k-1} \sum_{i=1}^{k} \frac{w_i}{w}(w_i^{-1} - w^{-1}) = \frac{1}{k-1}\left(kw^{-1} - w^{-1}\right) = \frac{1}{w}.$$

□

Moreover, it is shown in Hartung (1999) that the quadratic form $w \cdot (k-1) \cdot \tilde{\sigma}^2(\tilde{\theta}_{RE})$ is χ^2-distributed with $(k-1)$ degrees of freedom and stochastically independent of $\tilde{\theta}_{RE}$ in the RE model if the study-specific estimators $\hat{\theta}_i$, $i = 1, \ldots, k$, are exactly normally distributed. Thus, we now consider the random variable

$$\left(\tilde{\theta}_{RE} - \theta\right) \Big/ \sqrt{\tilde{\sigma}^2(\tilde{\theta}_{RE})},$$

which under the above assumption is exactly t-distributed with $(k-1)$ degrees of freedom. Therefore, an alternative test of $H_0 : \theta = 0$ is given by

$$\frac{\left|\tilde{\theta}_{RE}\right|}{\sqrt{\tilde{\sigma}^2(\tilde{\theta}_{RE})}} > t_{k-1;1-\alpha/2}, \tag{4.12}$$

where $t_{\nu;\gamma}$ stands for the γ-quantile of the t-distribution with ν degrees of freedom.

The alternative test in 4.12, however, depends on the unknown between-study variance τ^2 and the unknown within-study variances $\sigma^2(\hat{\theta}_i)$, $i = 1, \ldots, k$. Thus, for a practical application of this test we replace the unknown variance

components by appropriate estimators. Then, the feasible test statistic reads

$$T_3 = \frac{\left|\hat{\theta}_{RE}\right|}{\sqrt{\hat{\sigma}^2(\tilde{\theta}_{RE})}} \tag{4.13}$$

and the hypothesis of no association is rejected if the observed value of T_3 exceeds the $(1 - \alpha/2)$-quantile of the t-distribution with $(k - 1)$ degrees of freedom.

4.7 COMBINED DECISION RULES

In the previous sections, we have always distinguished between the FE and the RE model. For practically conducting a meta-analysis one usually has to choose in advance between these two models. In the literature, there exist different opinions how to deal with this decision problem.

A widespread procedure is to make first a test of homogeneity using the test statistic Q_1 from Equation 4.6 and, if the hypothesis of homogeneity is rejected, one uses the RE model, otherwise the FE model (Normand, 1999). Note that the hypothesis of homogeneity in the FE model, that is, $H_0 : \theta_1 = \ldots = \theta_k$, is equivalent to the hypothesis that the between-study variance τ^2 in the RE model is equal to zero. But the test of homogeneity often has low power against the alternative $\tau^2 > 0$ so that one cannot satisfactorily avoid the false use of the FE model if a between-study variance is present. The effect of a dramatically increasing Type I error by using the test statistic T_1 from Equation 4.7, if heterogeneity is present, is, for example, shown in Ziegler and Victor (1999) and in our simulation study described in the next section.

Whitehead and Whitehead (1991) propose a similar decision making at first sight. They consider the method of moments estimator $\hat{\tau}^2$ from Equation 4.9 of the between-study variance and suggest to use the FE model if $\hat{\tau}^2$ yields a negative estimate, otherwise the RE model. But this procedure is identical to the principle to always use the RE model with the truncated variance estimator $\hat{\tau}^2_+$ from Equation 4.10.

Both procedures have in common that the decision for an analysis in a corresponding model depends on a judgement of the variation between the studies. If one is mainly interested in testing the hypothesis $H_0 : \theta = 0$, the "pre-test" determines the test procedure which has to be used, but a false decision of the "pre-test" may considerably affect the properties of the test procedure. The most crucial point in the just described choice between the two models is given if the true between-study variance τ^2 is relatively small. In this situation one may expect the most false decisions between the models. For growing τ^2 the decision for the RE model becomes more and more certain. Thus, we will consider decision rules for tests of the hypothesis $H_0 : \theta = 0$, which incorporate the test procedure in the FE model as well as in the RE model and depend only in part on a variance estimate of the between-study variance. Moreover,

we make always use of the alternative test statistic T_3 from Equation 4.13 in the RE model.

The first combined decision rule is that the hypothesis $H_0 : \theta = 0$ is rejected if $T_1 > \chi^2_{1;1-\alpha}$ and $|T_3| > t_{k-1;1-\alpha/2}$, that means, we require that the commonly used test in the FE model as well as the alternative test in the RE model has to reject H_0. The idea behind this decision rule is that for a small between-study variance one may correct the significance level of the anticonservative test based on T_1 in the FE model towards the prescribed one, because the alternative test based on T_3 in the RE model may possess a smaller significance level. If the between-study variance grows, the role of the test based on T_1 becomes more and more ignorable and the combined decision rule is nearly identical to the decision rule simply based on the alternative test.

Furthermore, we consider two additional combined decision rules which include the estimation of the between-study variance. The first combined decision rule rejects $H_0 : \theta = 0$ if ($|T_3| > t_{k-1;1-\alpha/2}$ and $\hat{\tau}^2_+ > 0$) or ($T_1 > \chi^2_{1;1-\alpha}$ and $|T_3| > t_{k-1;1-\alpha/2}$, and $\hat{\tau}^2_+ = 0$), that is, we always require that the alternative test in the RE model rejects the hypothesis H_0 irrespective of the estimated value of τ^2, but if the truncated estimator $\hat{\tau}^2_+$ is equal to zero, that is, the usual method of moments estimator yields a negative estimate, the commonly used test in the FE model has also to reject H_0. This combined decision rule is motivated to correct a possible anticonservative attitude of the alternative test statistic in the RE model if the between-study variance is small and the test is anticonservative, while the commonly used test in the FE model is rather conservative in this situation. But if the alternative test statistic in the RE model performs better in case of small τ^2, this decision rule is nearly identical to the previous combined decision rule. Again, for growing τ^2 this decision rule becomes more and more similar to the decision rule based solely on the alternative test statistic T_3 in the RE model.

The second combined decision rule, which incorporates an estimation of τ^2, rejects $H_0 : \theta = 0$ if ($T_1 > \chi^2_{1;1-\alpha}$ and $\hat{\tau}^2_+ = 0$) or ($T_1 > \chi^2_{1;1-\alpha}$ and $|T_3| > t_{k-1;1-\alpha/2}$ and $\hat{\tau}^2_+ > 0$), that means, we always require that the commonly used test in the FE model rejects the hypothesis H_0 irrespective of the estimated value of τ^2, but if the truncated estimator $\hat{\tau}^2_+$ is greater than zero, the alternative test in the RE model has also to reject the hypothesis. This decision rule is similar to the proposal of Whitehead and Whitehead (1991), except that we use the alternative test statistic T_3 instead of the commonly used T_2 in the RE model. Since we also require to carry out the commonly used test in the FE model if the variance estimate of τ^2 is positive, this combined decision rule reduces the anticonservative attitude of this test for growing τ^2 as well as a possible anticonservative attitude of the test based on the alternative test statistic for small between-study variances.

4.8 SIMULATION STUDY

In a small simulation study, we compare the decision rules according to their actual Type I error rate. Table 4.1 summarizes all seven different decision rules we have already discussed and which are considered in the simulation study.

Table 4.1 Tests and Corresponding Decision Rules for Rejecting the Hypothesis $H_0 : \theta = 0$ at Level α With the Test Statistics T_1, T_2, and T_3

Test	Decision Rule: Reject $H_0 : \theta = 0$ if	
ψ_1	$T_1 > \chi^2_{1;1-\alpha}$	
ψ_2	$T_2 > \chi^2_{1;1-\alpha}$	
ψ_3	$\lvert T_3\rvert > t_{k-1;1-\alpha/2}$	
ψ_4	$T_1 > \chi^2_{1;1-\alpha}$,	if $Q_1 \leq \chi^2_{k-1;1-\alpha}$
	$T_2 > \chi^2_{1;1-\alpha}$,	if $Q_1 > \chi^2_{k-1;1-\alpha}$
ψ_5	$T_1 > \chi^2_{1;1-\alpha}$ and $\lvert T_3\rvert > t_{k-1;1-\alpha/2}$	
ψ_6	$\lvert T_3\rvert > t_{k-1;1-\alpha/2}$,	if $\hat{\tau}^2_+ > 0$
	$T_1 > \chi^2_{1;1-\alpha}$ and $\lvert T_3\rvert > t_{k-1;1-\alpha/2}$,	if $\hat{\tau}^2_+ = 0$
ψ_7	$T_1 > \chi^2_{1;1-\alpha}$,	if $\hat{\tau}^2_+ = 0$
	$T_1 > \chi^2_{1;1-\alpha}$ and $\lvert T_3\rvert > t_{k-1;1-\alpha/2}$,	if $\hat{\tau}^2_+ > 0$

Note. For a definition of T_1, T_2, and T_3 see Equations 4.7, 4.11, and 4.13, for Q_1 see Equation 4.6, and for $\hat{\tau}^2_+$ see Equation 4.10.

As an example, we choose the random one-way ANOVA model with heteroscedastic error variances, which is given by

$$y_{ij} = \mu + a_i + e_{ij}, \quad i = 1, \ldots, k; \; j = 1, \ldots, n_i, \tag{4.14}$$

with $a_i \sim \mathcal{N}(0, \tau^2)$ and $e_{ij} \sim \mathcal{N}(0, \sigma_i^2)$, and all random effects are stochastically independent. Instead of the individual data, usually summary statistics are given in publications. We consider the arithmetic mean $\hat{\mu}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ as the estimator of μ in each study. For this estimator, we have the following distributional property:

$$\hat{\mu}_i \sim \mathcal{N}\left(\mu, \tau^2 + \sigma^2\left(\hat{\mu}_i\right)\right), \quad \sigma^2(\hat{\mu}_i) = \sigma_i^2/n_i, \quad i = 1, \ldots, k. \tag{4.15}$$

Furthermore, an unbiased estimator of the error variance in model 4.14 is given by $\hat{\sigma}_i^2 = s_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \hat{\mu}_i)^2/(n_i - 1)$, $i = 1, \ldots, k$, so that an unbiased estimator of the within-study variance $\sigma^2(\hat{\mu}_i)$ is $\hat{\sigma}^2(\hat{\mu}_i) = s_i^2/n_i$. This estima-

tor is stochastically independent of $\hat{\mu}_i$ and it holds that $(n_i - 1)s_i^2/\sigma_i^2$ is χ^2-distributed with $(n_i - 1)$ degrees of freedom.

In the simulation study, we consider four different patterns of sample sizes and error variances, which are given in Table 4.2 for $k = 3$ studies. The first pattern has equal sample sizes and equal error variances, whereas in the second pattern the sample sizes are doubled in each study. In the last two patterns we consider different sample sizes in each study. In pattern 3, the error variances are increasing with growing sample sizes, but the within-study variance σ_i^2/n_i, $i = 1, 2, 3$, is always 0.1. In pattern 4, the error variances are decreasing with growing sample sizes so that the study with the largest sample size has the smallest within-study variance.

Table 4.2 Sample Sizes and Error Variances Used in the Simulation Study

Pattern for $k = 3$	Sample Sizes (n_1, n_2, n_3)	Error Variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$
1	$(5, 5, 5)$	(4,4,4)
2	$(10, 10, 10)$	(4,4,4)
3	$(10, 20, 40)$	(1,2,4)
4	$(10, 20, 40)$	(4,2,1)
Patterns for $k = 9$: Replicated Twice the Patterns for $k = 3$		

In Table 4.3, the results of our simulation study are put together. We present the results for $k = 3$ and $k = 9$ studies, where the patterns for $k = 9$ studies have been constructed by replicating the patterns for $k = 3$ studies twice. As different values of the between-study variance we choose $\tau^2 = 0, 0.1, 1, 10$, and as the estimator of τ^2 we always use $\hat{\tau}_+^2$ from Equation 4.10. Besides the estimated Type I error rates of the test given a prescribed significance level of $\alpha = .05$, the table also contains the estimated proportion of negative estimates of τ^2 using $\hat{\tau}^2$ from Equation 4.9 and the estimated power of the test of homogeneity based on Q_1 from Equation 4.6. Each estimated value in the table is based on 10,000 replications of the corresponding model.

From Table 4.3, we see that the commonly used test ψ_1 in the FE model is rather anticonservative in the normal mean case if the between-study variance is equal to zero, that is, the FE model is the theoretically correct one. Moreover, the Type I error rates increase if the number of studies grows, but if the sample sizes increase for fixed k the Type I error rates decrease. If between-study variation is present, the estimated Type I error rates become still larger and they increase for growing between-study variance. Consequently, one should be very careful in the normal mean case to apply the test ψ_1.

The commonly used test ψ_2 in the RE model, which coincides with the proposal of Whitehead and Whitehead how to deal with the choice between FE and RE model, yields its best results in comparison to the prescribed level if the between-study variance is equal to zero. From the estimated proportions of negative estimates of τ^2, we observe that for $k = 3$ in approximately 60 % of

Table 4.3 Estimated Type I Error Rates (in %) for the Seven Different Two-Sided Tests of $H_0 : \mu = 0$ in Model 4.15, Given $\alpha = .05$, the Estimated Proportion (in %) of Negative Estimates of τ^2, the Estimated Power (in %) of the Test of Homogeneity for $k = 3$ and $k = 9$ Studies

k	Pattern	τ^2	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7	$\hat{\tau}^2$ neg.	Power (Q_1)
3	1	0	19.4	10.3	7.2	14.8	3.7	3.7	7.8	56.4	12.6
		0.1	21.3	11.0	7.2	15.9	4.2	4.2	8.4	53.2	14.6
		1	36.3	14.8	6.8	20.9	5.2	5.2	9.1	32.4	33.8
		10	69.3	18.6	5.1	20.5	5.0	5.0	5.8	6.3	82.7
	2	0	10.6	6.4	5.5	8.9	2.2	2.2	5.1	60.9	8.0
		0.1	14.8	8.2	5.9	11.7	2.8	2.8	6.2	51.7	13.4
		1	37.5	15.6	5.7	19.9	4.7	4.7	8.3	23.5	45.5
		10	73.7	18.7	5.2	19.3	5.2	5.2	5.7	3.7	89.4
	3	0	10.0	5.1	5.5	7.3	1.7	1.7	4.2	57.1	11.8
		0.1	26.4	13.7	7.7	20.1	5.2	5.2	10.7	44.4	20.6
		1	64.0	21.1	8.4	27.0	7.8	7.8	10.8	16.5	61.0
		10	87.7	21.0	6.1	21.5	6.1	6.1	6.2	2.2	93.7
	4	0	6.6	4.1	4.7	5.7	1.3	1.3	3.4	59.7	8.5
		0.1	35.2	16.6	9.1	25.2	6.8	6.8	12.3	38.7	24.9
		1	73.7	21.6	8.0	25.5	7.6	7.6	9.3	10.5	73.5
		10	91.1	22.2	5.8	22.5	5.7	5.7	5.8	1.4	96.0
9	1	0	26.9	9.8	9.4	15.3	7.6	7.7	8.1	32.9	29.8
		0.1	28.9	9.7	9.0	14.9	7.6	7.7	8.0	26.3	35.9
		1	44.5	9.8	6.8	12.8	6.7	6.7	6.7	5.5	74.7
		10	74.9	9.8	5.3	9.8	5.3	5.3	5.3	0.0	99.9
	2	0	12.1	6.5	6.6	9.3	4.6	4.6	5.2	44.7	14.1
		0.1	17.2	8.1	7.2	11.6	5.9	5.9	6.3	30.5	26.4
		1	39.7	8.6	5.4	10.0	5.3	5.3	5.4	2.2	86.2
		10	76.1	9.3	5.4	9.3	5.4	5.4	5.4	0.0	100
	3	0	11.2	5.4	5.3	7.5	3.7	3.7	4.1	40.8	19.5
		0.1	27.8	9.4	6.6	13.8	6.2	6.2	6.5	14.8	51.8
		1	66.7	10.3	6.0	10.7	6.0	6.0	6.0	0.3	97.7
		10	88.3	10.5	5.3	10.5	5.3	5.3	5.3	0.0	100
	4	0	7.5	4.9	5.4	6.6	3.3	3.3	3.8	49.1	11.4
		0.1	36.1	10.4	7.4	14.7	7.2	7.2	7.4	8.0	65.8
		1	73.9	10.0	5.5	10.0	5.5	5.5	5.5	0.1	99.6
		10	91.8	10.7	5.3	10.7	5.3	5.3	5.3	0.0	100

the simulated cases and for $k = 9$ in still approximately 40 % of the cases one actually performs the commonly used test in the FE model. If between-study variation is present, the estimated Type I error rates in the normal mean case increase to more than four times the prescribed level for $k = 3$ studies and to nearly twice the prescribed level for $k = 9$ studies. But this indicates that for an increasing number of studies the actual Type I error rate diminishes.

The alternative test ψ_3 in the RE model yields the best results concerning the actual significance level in comparison to the tests ψ_1 and ψ_2. If $\tau^2 = 0$ is not present, the tests ψ_2 and ψ_3 have rather similar estimated Type I error rates, but if a positive between-study variance exists, the alternative test has estimated Type I error rates often near the prescribed level and only in some cases goes beyond 7 %.

The test ψ_4, which has a decision rule depending on the test of homogeneity, always has an estimated Type I error rate which is greater or equal to the estimated Type I error rate of the test ψ_2. Thus, this combined decision rule does not yield an improvement concerning the actual significance level.

The test ψ_5, which requires the rejection of the hypothesis with the test ψ_1 and with the alternative test ψ_3, has always estimated Type I error rates which are less or equal to the estimated Type I error rates of the test ψ_3. The test ψ_5 is often an essential improvement in comparison to the test ψ_3 if the between-study variance τ^2 is equal to 0.1 or 1, but for $\tau^2 = 0$ the test ψ_5 may become rather conservative.

Due to the fact that the test ψ_1 is anticonservative the test ψ_6 yields nearly the same results as the test ψ_5, as we have pointed out in Section 4.7.

The test ψ_7 yields rather similar results like the test ψ_5 if $k = 9$ studies are considered. The estimated Type I error rates of ψ_7 are slightly greater or equal to the estimated Type I error rates of ψ_5. The relationship between these two tests also holds for $k = 3$ studies, but the difference between the estimated Type I error rates of the tests is much larger. Often, the estimated Type I error rates of ψ_7 are twice as large as the estimated ones of ψ_5.

REFERENCES

Böckenhoff, A., & Hartung, J. (1998). Some corrections of the significance level in meta-analysis. *Biometrical Journal*, *40*, 937–947.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129.

DerSimonian, R., & Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177–188.

Hartung, J. (1976). Some inequalities for a random function. *Teorija Verojatnostei i ee Primenenja*, *21*, 661–665 (See also: (1977) *Theory of Probability and its Application, SIAM, 21*).

Hartung, J. (1981). Nonnegative minimum biased invariant estimation in variance component models. *Annals of Statistics*, *9*, 278–292.

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*, *41*, 901–916.

Hartung, J., & Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, *20*, 3875–3889.

Li, Y., Shi, L., & Roth, H. D. (1994). The bias of the commonly-used estimate of variance in meta-analysis. *Communications in Statistics – Theory and Methods*, *23*, 1063–1085.

Normand, S.-L. T. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine*, *18*, 321–359.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association*, *67*, 112–115.

Whitehead, A., & Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, *10*, 1665–1677.

Ziegler, S., & Victor, N. (1999). Gefahren der Standardmethoden für Meta-Analysen bei Vorliegen von Heterogenität [Some problems of the standard meta-analysis methods in the presence of heterogeneity]. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, *30*, 131–140.