Heteroscedastic heterogeneity in meta-analysis

Symposium on 'Recent Advances in Meta-Analysis'
Department of Medical Statistics
University Medical Center Göttingen

Wolfgang Viechtbauer
Maastricht University

2026-05-29

Standard Meta-Analysis

- have $i = 1, \ldots, k$ effect size estimates:

$$y_i | \theta_i \sim N(\theta_i, v_i)$$

where θ_i is the true effect in the ith study and v_i is the sampling variance ('within-study variance')

- **goal**: estimate the size of the (average) effect
- **equal-effects model**: (common-effect model)

$$\theta_i = \theta$$

- **random-effects model**:

$$\theta_i = \mu + u_i$$

where $u_i \sim N(0, \tau^2)$ and τ^2 is the amount of heterogeneity ('between-study variance')

Example: Writing-to-Learn Interventions

- Bangert-Drowns et al. (2004) [1] meta-analyzed $k = 48$ studies examining the **effectiveness of writing-to-learn interventions** for improving educational achievement
- difference between intervention and control groups (on final grade or test score) given as **standardized mean differences** (positive values = better performance in the intervention group)

```
# load the metafor package
library(metafor)

# copy the dataset to dat
dat <- dat.bangertdrowns2004
```


Example: Equal-Effects Model

```
# fit the equal-effects model
res.ee <- rma(yi, vi, data=dat, method="EE")
res.ee
```

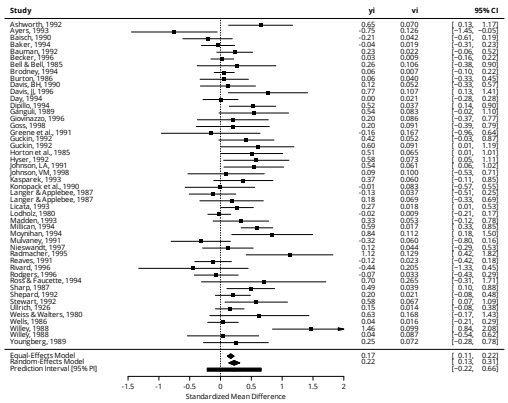
```
Equal-Effects Model (k = 48)

I^2 (total heterogeneity / total variability):   56.12%
H^2 (total variability / sampling variability):  2.28

Test for Heterogeneity:
Q(df = 47) = 107.1061, p-val < .0001

Model Results:

estimate      se    zval    pval   ci.lb   ci.ub
  0.1656  0.0269  6.1499  <.0001  0.1128  0.2184
```


Example: Random-Effects Model

```
# fit the random-effects model
res.re <- rma(yi, vi, data=dat)
res.re
```

```
Random-Effects Model (k = 48; tau^2 estimator: REML)

tau^2 (estimated amount of total heterogeneity): 0.0499 (SE = 0.0197)
tau (square root of estimated tau^2 value):      0.2235
I^2 (total heterogeneity / total variability):   58.37%
H^2 (total variability / sampling variability):  2.40

Test for Heterogeneity:
Q(df = 47) = 107.1061, p-val < .0001

Model Results:

estimate      se    zval    pval   ci.lb   ci.ub
  0.2219  0.0460  4.8209  <.0001  0.1317  0.3122
```


Forest Plot



Testing for Heterogeneity

- are the true effects actually **heterogeneous**?
- want to test:

$$H_0\colon \theta_1 = \theta_2 = \ldots = \theta_k \equiv \theta$$

- **Q-test**: standard test for heterogeneity [e.g., 2, 3]

$$Q = \sum w_i (y_i - \hat{\theta})^2$$

where $w_i = 1/v_i$ and $\hat{\theta} = \sum w_i y_i / \sum w_i$

- $Q \sim \chi^2_{k-1}$ under H_0 (and the assumptions of the model)
- in the example: $Q(47) = 107.1061, p < .0001$

Saturated Meta-Regression Model

- consider the **saturated meta-regression model**:

$$y_i = \beta_1 s_{i1} + \ldots + \beta_k s_{ik} + \varepsilon_i = \sum_{i=1}^{k} \beta_i s_{ij} + \varepsilon_i$$

where s_{ij} is equal to 1 for the ith study and 0 otherwise

- model that allows each study to have its own study-specific true effect (without imposing any constraints or distributional assumptions)

Likelihood Ratio Test

- let ll_{ee} denote the log likelihood for the equal-effects model and ll_{fe}^{sat} for the saturated meta-regression model
- then the likelihood ratio test statistic

$$X^2 = -2(ll_{ee} - ll_{fe}^{sat})$$

is exactly identical to the Q-test

Likelihood Ratio Test

```
res.sat <- rma(yi, vi, mods = ~ 0 + factor(id), data=dat, method="FE")
res.sat

Fixed-Effects with Moderators Model (k = 48)

Test for Residual Heterogeneity:
QE(df = 0) = 0.0000, p-val = 1.0000

Test of Moderators (coefficients 1:48):
QM(df = 48) = 144.9279, p-val < .0001

Model Results:

               estimate      se     zval    pval    ci.lb    ci.ub
factor(id)1      0.6500  0.2646   2.4568  0.0140   0.1314   1.1686
factor(id)2     -0.7500  0.3550  -2.1129  0.0346  -1.4457  -0.0543
...
factor(id)47     0.0400  0.2950   0.1356  0.8921  -0.5381   0.6181
factor(id)48     0.2500  0.2683   0.9317  0.3515  -0.2759   0.7759
```


Likelihood Ratio Test

```
# compute the X^2 statistic manually
-2 * c(logLik(res.ee) - logLik(res.sat))

[1] 107.1061

# use anova() to conduct the likelihood ratio test
anova(res.ee, res.sat)

         df     AIC      BIC      AICc   logLik      LRT   pval
Full     48 41.4508 131.2684 4745.4508  27.2746
Reduced   1 54.5568  56.4280   54.6438 -26.2784 107.1061 <.0001
```


Wald-type Test

- for the saturated model: $b_i = y_i$ and $\mathrm{Var}[b_i] = v_i$
- define the contrasts $b_{i'} - b_i$ for $i = 1, \ldots, k; i \neq i'$
- the test of the $k - 1$ contrasts is identical to the Q-test
- this is invariant to the choice of i'

Wald-type Test

```
# define the contrast matrix C with i' = 1
k <- nrow(dat)
C <- cbind(1, -diag(k-1))

# conduct the Wald-type test of the linear contrasts defined by C
anova(res.sat, X=C)
```

```
Hypotheses:
1:   factor(id)1 - factor(id)2 = 0
2:   factor(id)1 - factor(id)3 = 0
...

Results:
    estimate     se    zval   pval
1:    1.4000 0.4427  3.1623 0.0016
2:    0.8600 0.3347  2.5697 0.0102
...

Omnibus Test of Hypotheses:
QM(df = 47) = 107.1061, p-val < .0001
```


Other Tests

- can also derive a score test but it is different from the Q-test and doesn't perform as well [4]
- can also compare the log likelihood of the equal-effects model (ll_{ee}) with that of the random-effects model (ll_{re})
- then

$$X^2 = -2(ll_{ee} - ll_{re})$$

  provides a likelihood ratio test of H_0: $\tau^2 = 0$, which is also a test for heterogeneity
- can either do this with ML or REML estimation

Other Tests

```
# LRT of H0: tau^2 = 0 using ML estimation
res.re <- rma(yi, vi, data=dat, method="ML")
anova(res.ee, res.re)
```

```
        df     AIC     BIC    AICc   logLik     LRT   pval
Full     2 40.5243 44.2667 40.7910 -18.2622
Reduced  1 54.5568 56.4280 54.6438 -26.2784 16.0325 <.0001
```

```
# LRT of H0: tau^2 = 0 using REML estimation
res.re <- rma(yi, vi, data=dat)
anova(res.ee, res.re)
```

```
        df     AIC     BIC    AICc   logLik     LRT   pval
Full     2 40.9886 44.6889 41.2613 -18.4943
Reduced  1 56.0767 57.9268 56.1656 -27.0383 17.0881 <.0001
```


Heteroscedastic Heterogeneity

- the **standard random-effects model** assumes that the amount of heterogeneity (τ^2) is **homoscedastic**
- this assumption may not be true
- consider the **saturated random-effects model**:

$$y_i = \mu + u_i + \varepsilon_i \quad \text{with} \quad u_i \sim N(0, \tau_i^2)$$

- model that allows each study to have its own study-specific τ_i^2 value (without imposing any constraints or distributional assumptions)

Saturated Random-Effects Model

- ML estimate of μ is

$$\hat{\mu} = \frac{\sum w_i y_i}{\sum w_i}$$

  where $w_i = 1/(\tau_i^2 + v_i)$
- ML estimate of τ_i^2 is

$$\hat{\tau}_i^2 = (y_i - \mu)^2 - v_i$$

- REML estimate of τ_i^2 is

$$\hat{\tau}_i^2 = (y_i - \hat{\mu})^2 - v_i + \frac{1}{\sum w_i}$$

- these can be obtained iteratively

Study-Specific τ_i^2 Estimates

study	ML	REML
Ashworth, 1992	0.2219	0.2216
Ayers, 1993	0.6131	0.6161
Baisch, 1990	0.0602	0.0619
Baker, 1994	0.0034	0.0047
Davis, JJ, 1996	0.3290	0.3284
Dipillo, 1994	0.1313	0.1314
Ganguli, 1989	0.1022	0.1021
Guckin, 1992	0.0443	0.0445
Guckin, 1992	0.1494	0.1492
Horton et al., 1985	0.0952	0.0953
Hyser, 1992	0.1482	0.1480
Johnson, LA, 1991	0.1242	0.1241
Kasparek, 1993	0.0078	0.0081
Langer & Applebee, 1987	0.0205	0.0220

study	ML	REML
Licata, 1993	0.0077	0.0083
Lodholz, 1980	0.0078	0.0091
Millican, 1994	0.2137	0.2135
Moynihan, 1994	0.4213	0.4206
Mulvaney, 1991	0.1246	0.1266
Radmacher, 1995	0.8917	0.8903
Reaves, 1991	0.0298	0.0313
Rivard, 1996	0.0972	0.0994
Rodgers, 1996	0.0000	0.0007
Ross & Faucette, 1994	0.0834	0.0830
Sharp, 1987	0.1056	0.1057
Stewart, 1992	0.1542	0.1540
Weiss & Walters, 1980	0.1027	0.1025
Willey, 1988	1.7243	1.7221

note: studies where $\hat{\tau}_i^2 = 0$ left out

Testing for Heteroscedastic Heterogeneity

- want to test:

$$H_0\colon \tau_1^2 = \tau_2^2 = \ldots = \tau_k^2 \equiv \tau^2$$

- can now derive various tests for **heteroscedastic heterogeneity** based on the same principles underlying the Q-test

Likelihood Ratio Test

- let ll_{re} and ll_{re}^{sat} denote the log likelihood for the standard and the saturated random-effects model
- then

$$X^2 = -2(ll_{re} - ll_{re}^{sat})$$

provides a **likelihood ratio test** of H_0

- can do the this with ML or REML estimation

Likelihood Ratio Test

```
hettest(yi, vi, data=dat, method="ML", test="lrt", boot=FALSE)

Test for Heteroscedastic Heterogeneity

Estimation method: Maximum likelihood
Test type:         Likelihood ratio test
Bootstrapping:     No

X^2(df = 47) = 28.6313, p = 0.9842
```

```
hettest(yi, vi, data=dat, method="REML", test="lrt", boot=FALSE)

Test for Heteroscedastic Heterogeneity

Estimation method: Restricted maximum likelihood
Test type:         Likelihood ratio test
Bootstrapping:     No

X^2(df = 47) = 27.8697, p = 0.9881
```


Wald-type Test

- let $\hat{\boldsymbol{\tau}}^2$ denote the vector with the MLEs of $\hat{\tau}_i^2$
- can show $\text{Var}[\hat{\boldsymbol{\tau}}^2] = \text{diag}\left(2(\hat{\tau}_1^2 + v_1)^2, \ldots, 2(\hat{\tau}_k^2 + v_k)^2\right)$
- define the contrasts $\hat{\tau}_{i'}^2 - \hat{\tau}_i^2$ for $i = 1, \ldots, k; i \neq i'$
- conduct an omnibus **Wald-type test** of the $k - 1$ contrasts
- can do an analogous test for REML estimation

Wald-type Test

```
hettest(yi, vi, data=dat, method="ML", test="wald", boot=FALSE)

Test for Heteroscedastic Heterogeneity

Estimation method: Maximum likelihood
Test type:         Wald-type test
Bootstrapping:     No

X^2(df = 47) = 5.2011, p = 1.0000
```

```
hettest(yi, vi, data=dat, method="REML", test="wald", boot=FALSE)

Test for Heteroscedastic Heterogeneity

Estimation method: Restricted maximum likelihood
Test type:         Wald-type test
Bootstrapping:     No

X^2(df = 47) = 5.1666, p = 1.0000
```


Score Test

- does not require that we fit the saturated random-effects model
- the **score test** of H_0 is given by

$$X^2 = \frac{1}{2}\sum(r_i^2 - 1)^2$$

where $r_i = (y_i - \hat{\mu})/\sqrt{\hat{\tau}^2 + v_i}$ is the ith Pearson residual under the random-effects model

- since the Pearson residuals are asymptotically normal, this is a test of over/underdispersion indicative of heteroscedasticity
- **interesting parallel**: Q-statistic is given by $Q = \sum r_i^2$ where $r_i = (y_i - \hat{\theta})/\sqrt{v_i}$ is the ith Pearson residual under the equal-effects model
- can do an analogous test for REML estimation

Score Test

```
hettest(yi, vi, data=dat, method="ML", test="score", boot=FALSE)

Test for Heteroscedastic Heterogeneity

Estimation method: Maximum likelihood
Test type:         Score test
Bootstrapping:     No

X^2(df = 47) = 74.2189, p = 0.0069
```

```
hettest(yi, vi, data=dat, method="REML", test="score", boot=FALSE)

Test for Heteroscedastic Heterogeneity

Estimation method: Restricted maximum likelihood
Test type:         Score test
Bootstrapping:     No

X^2(df = 47) = 73.2355, p = 0.0085
```


Parametric Bootstrap

- regularity conditions underlying these tests are violated
- the assumed null distribution (χ^2 with $k-1$ df) is not correct
- can consider **parametric bootstrapping** as a potential fix:
  - fit the random-effects model
  - simulate k new effect size estimates from $N(\mu, \hat{\tau}^2 + v_i)$
  - obtain the test statistic of interest based on the simulated data
  - repeat many (e.g., $B = 1000$) times
  - let X_b^2 denote the statistic for the bth bootstrap sample
  - then $\sum_{b=1}^{B} (X_b^2 \geq X^2)/B$ is the bootstrap p-value
- use `hettest(..., boot=TRUE)` to use bootstrapping with a test (which is the default)

Results

Test	Estimator	X^2	Regular p-value	Bootstrap p-value
LRT	ML	28.63	0.98	0.21
LRT	REML	27.87	0.99	0.23
Wald	ML	5.20	1.00	0.24
Wald	REML	5.17	1.00	0.27
Score	ML	74.22	0.01	0.06
Score	REML	73.24	0.01	0.06

Random Heteroscedasticity Random-Effects Model

- location-scale models can be used to model the amount of heterogeneity as a function of one or more predictors
- **meta-analytic location-scale model**: [5]

$$y_i = \beta_0 + \beta_1 x_i + u_i + \varepsilon_i$$

where $u_i \sim N(0, \tau_i^2)$ and $\varepsilon_i \sim N(0, v_i)$ and where

$$\ln(\tau_i^2) = \alpha_0 + \alpha_1 z_i$$

- can extend this model to allow for 'residual heteroscedasticity':

$$\ln(\tau_i^2) = \alpha_0 + \alpha_1 z_i + h_i$$

where $h_i \sim N(0, \omega^2)$

Random Heteroscedasticity Random-Effects Model

- now let's drop all predictors for the location and scale part
- yields the **random heteroscedasticity random-effects model**:

$$y_i = \mu + u_i + \varepsilon_i \text{ with } u_i \sim N(0, \tau_i^2)$$
$$\ln(\tau_i^2) = \alpha_0 + h_i \text{ with } h_i \sim N(0, \omega^2)$$

- all heteroscedasticity in the τ_i^2 values will be subsumed by ω^2
- much more parsimonious than the saturated random-effects model (at the cost of a distributional assumption)
- can be fitted with ML estimation
- can conduct a likelihood ratio test comparing this model with the standard random-effects model to test

$$H_0\colon \omega^2 = 0$$

Random Heteroscedasticity Random-Effects Model

```
# fit the random heteroscedasticity random-effects model
res.rh <- rma(yi, vi, data=dat, scale = ~ 1, randhet=TRUE)
res.rh

Location-Scale Model (k = 48; tau^2 estimator: ML)

Test for Heterogeneity:
Q(df = 47) = 107.1061, p-val < .0001

Model Results (Location):

estimate      se    zval    pval   ci.lb   ci.ub
  0.2207  0.0463  4.7691  <.0001  0.1300  0.3114

Model Results (Scale):

          estimate      se     zval    pval    ci.lb    ci.ub
intrcpt    -3.0567  0.4674  -6.5399  <.0001  -3.9728  -2.1406
omega^2     0.0000      NA      ---     ---      NA       NA
```

Random Heteroscedasticity Random-Effects Model

```
# conduct a likelihood ratio test comparing the standard random-effects
# model with the random heteroscedasticity random-effects model
res.re <- rma(yi, vi, data=dat, method="ML")
X2 <- -2 * c(logLik(res.re) - logLik(res.rh))
X2
```

```
[1] 0.007980385
```

```
pchisq(X2, df=1, lower.tail=FALSE)
```

```
[1] 0.9288173
```


Simulation Study

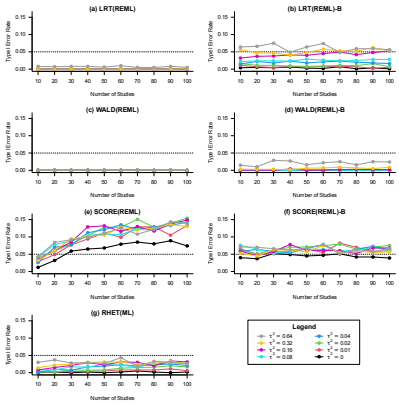
- conducted an extensive simulation study to examine the Type I error and power of these tests
- simulated standardized mean differences
- conditions:
  - 10 values of k $(10, 20, \ldots, 100)$
  - 4 values of $\bar{n}$ $(30, 50, 80, 100)$
  - 8 values of τ^2 $(0, 0.01, 0.02, 0.04, 0.08, 0.16, 0.32, 0.64)$
  - for power, assumed the presence of two groups of studies with $\tau_1^2 \neq \tau_2^2$ (all 28 combinations of the τ^2 values)
- tests: LRT(ML), LRT(REML), WALD(ML), WALD(REML), SCORE(ML), SCORE(REML), LRT(ML)-B, LRT(REML)-B, WALD(ML)-B, WALD(REML)-B, SCORE(ML)-B, SCORE(REML)-B, and RHET(ML)
- 1000 iterations per condition

Results

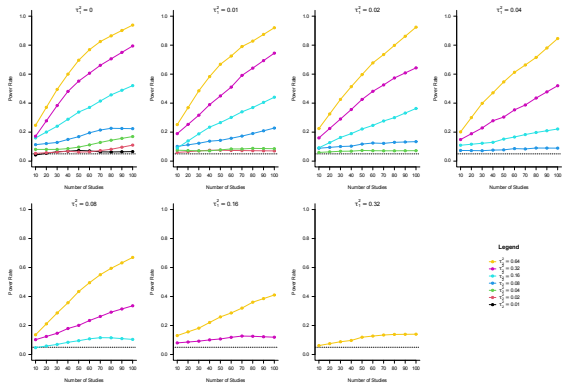
- $\bar{n}$ had relatively little influence on the results[1]
- ML and REML estimation yielded very similar results
- likelihood ratio and Wald-type tests were conservative
- score test was liberal, esp. when k and/or τ^2 increased
- bootstrapping helped for all tests, especially for the score test
- likelihood ratio test of the random heteroscedasticity random-effects model was conservative
- power increased with k and the τ_2^2/τ_1^2 ratio

[1] This may not be entirely accurate. Due to lack of time, some of the conditions were run with a smaller number of iterations for this presentation. The results were therefore more noisy. The final simulation study will indicate how much $\bar{n}$ matters.

Type I Error Rates



Power of the SCORE(REML)-B Test



Conclusions

- the bootstrapped score test adequately controls the Type I error rate (range: 0.037 to 0.082, mean 0.06)
- has adequate power to detect heteroscedasticity when k is large (and there is sufficient heteroscedasticity)
- test is implemented in `hettest()` function in the `metafor` package (development version!)
- if heteroscedasticity in τ^2 is detected, may be able to account for it by means of location-scale models (see my talk from the previous meeting [6])
- can also report the results from the saturated random-effects model (note that $\text{Var}[\hat{\mu}] = \frac{1}{\sum w_i}$ for both models, with $w_i = 1/(\hat{\tau}^2 + v_i)$ for the standard and $w_i = 1/(\hat{\tau}_i^2 + v_i)$ for the saturated model)[1]

[1] Not sure if this is a good idea though (needs to be examined).

Conclusions

```
# conduct the score test (with REML estimation and bootstrapping)
res <- hettest(yi, vi, data=dat)
res

Test for Heteroscedastic Heterogeneity

Estimation method: Restricted maximum likelihood
Test type:         Score test
Bootstrapping:     Yes (1000/1000 iterations)

X^2(df = 47) = 73.2355, p = 0.0630
```


Conclusions

```
# fit the saturated random-effects model with rma.mv()
dat$id2 <- dat$id
res.re.sat <- rma.mv(yi, vi, random = ~ id2 | id, struct="DIAG",
                     data=dat, tau2=res$tau2i)
res.re.sat

Multivariate Meta-Analysis Model (k = 48; method: REML)

Variance Components:

             estim    sqrt  k.lvl  fixed  level
tau^2.1     0.2216  0.4708      1    yes      1
tau^2.2     0.6161  0.7849      1    yes      2
...

Model Results:

estimate      se    zval    pval   ci.lb   ci.ub
  0.1109  0.0313  3.5458  0.0004  0.0496  0.1722
```


References [1]

1. Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, *74*(1), 29–58. doi:10.3102/00346543074001029
2. Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Supplement to the Journal of the Royal Statistical Society*, *4*(1), 102–118. doi:10.2307/2984123
3. Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*(1), 101–129. doi:10.2307/3001666
4. Viechtbauer, W. (2007). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *60*(1), 29–60. doi:10.1348/000711005X64042

References [2]

5. Viechtbauer, W., & López-López, J. A. (2022). Location-scale models for meta-analysis. *Research Synthesis Methods*, *13*(6), 697–715. doi:10.1002/jrsm.1562
6. Viechtbauer, W. (2025). *Location-scale models for meta-analysis*. Symposium on 'Recent Advances in Meta-Analysis', Department of Mathematical Statistics, TU Dortmund University, Germany.

Thank You for Your Attention!

Questions, Comments, Suggestions?

wolfgang.viechtbauer@maastrichtuniversity.nl

https://www.wvbauer.com

https://www.metafor-project.org

@wviechtb @wviechtb