

Testing for Heteroscedastic Heterogeneity in Meta-Analysis

Society for Research Synthesis Methodology
2026 Annual Conference
Chania, Crete, Greece
June 10-12, 2026

Wolfgang Viechtbauer
Maastricht University
2026-06-12

1

Heterogeneity and Heteroscedasticity

- **heterogeneity** is ubiquitous in meta-analyses [1, 2]
- increased interest in studying under which circumstances there is more/less heterogeneity in the true effects [3-7]
- if the amount of heterogeneity is not constant across studies, then this implies that the heterogeneity is **heteroscedastic**
- can we test for this?

2

Example: Writing-to-Learn Interventions

- Bangert-Drowns et al. (2004) [8] 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
```

3

Example: Equal-Effects Model

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

Equal-Effects Model (k = 48)

I² (total heterogeneity / total variability): 56.12%
H² (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

4

Example: Random-Effects Model

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

Random-Effects Model (k = 48; tau² estimator: REML)

tau² (estimated amount of total heterogeneity): 0.0499 (SE = 0.0197)
tau (square root of estimated tau² value): 0.2235
I² (total heterogeneity / total variability): 58.37%
H² (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

5

Testing for Heterogeneity

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

$$H_0: \theta_1 = \theta_2 = \dots = \theta_k \equiv \theta$$

- **Q-test**: standard test for heterogeneity [e.g., 9, 10]

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

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

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

6

Saturated Meta-Regression Model

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

$$y_i = \beta_1 s_{i1} + \dots + \beta_k s_{ik} + \varepsilon_i$$

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

- likelihood ratio test**:
 - let ll_{ee} and ll_{fe}^{sat} denote the log likelihoods of the equal- and the saturated meta-regression models; then

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

is identical to the Q-test

- Wald-type test**:
 - for the saturated model: $b_i = y_i$
 - define the contrasts $b_{i'} - b_i$ for $i = 1, \dots, k; i \neq i'$
 - the test of the $k - 1$ contrasts is identical to the Q-test

7

Other Tests

- score test**:
 - is different from the Q-test [11]
- likelihood ratio test**:
 - 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

8

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)

9

Testing for Heteroscedastic Heterogeneity

- want to test:

$$H_0: \tau_1^2 = \tau_2^2 = \dots = \tau_k^2 \equiv \tau^2$$

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

10

Likelihood Ratio Test

- let ll_{re} and ll_{re}^{sat} denote the log likelihood of 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

11

Wald-type Test

- define the contrasts $\hat{\tau}_{i'}^2 - \hat{\tau}_i^2$ for $i = 1, \dots, k; i \neq i'$
- conduct an omnibus **Wald-type test** of the $k - 1$ contrasts
- can do the this with ML or REML estimation

12

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 standard normal, this is a test of over/underdispersion (i.e., 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

13

Results (including parametric bootstrap versions)

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

14

Random Heteroscedasticity Random-Effects Model

- the **random heteroscedasticity random-effects model:**

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

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

- 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

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

- in the present example, $\hat{\omega}^2 \approx 0$, so $p \approx 1$

15

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, ..., 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 $\tau_1^2 < \tau_2^2$ (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

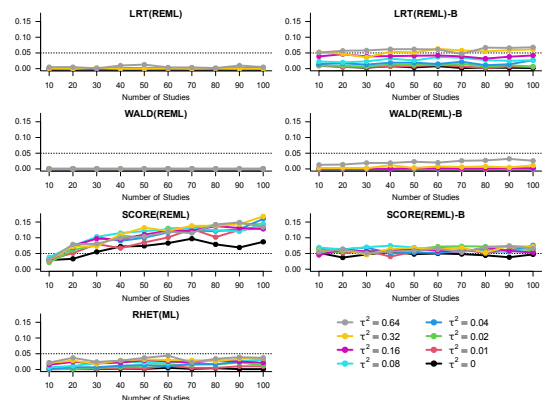
16

Results

- 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 difference between τ_2^2 and τ_1^2

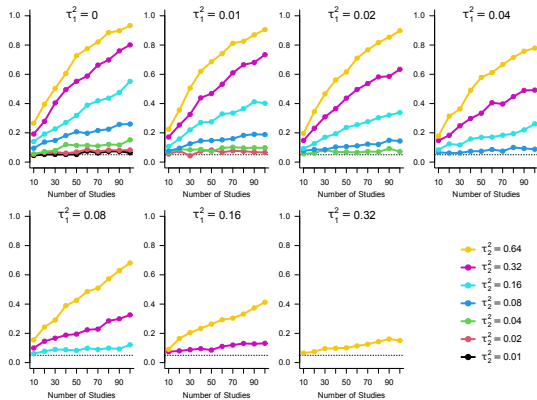
17

Type I Error Rates



18

Power of the SCORE(REML)-B Test



19

Conclusions

- the bootstrapped score test adequately controls the Type I error rate (range: 0.037 to 0.076, mean 0.059)
- has adequate power to detect heteroscedasticity when k is large (and there is sufficient heteroscedasticity)
- tests are implemented in the `hettest()` function in the `metafor` package (development version!): `hettest(yi, vi, data=dat)` for the bootstrap score test
- if heteroscedasticity in τ^2 is detected, may be able to account for it by means of location-scale models (see [4])

20

References [1]

1. Rhodes, K. M., Turner, R. M., & Higgins, J. P. (2015). Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology*, 68(1), 52–60. doi:10.1016/j.jclinepi.2014.08.012
2. Rhodes, K. M., Turner, R. M., & Higgins, J. P. (2016). Empirical evidence about inconsistency among studies in a pair-wise meta-analysis. *Research Synthesis Methods*, 7(4), 346–370. doi:10.1002/jrsm.1193
3. Int'Hout, J., Ioannidis, J. P. A., Borm, G. F., & Goeman, J. J. (2015). Small studies are more heterogeneous than large ones: A meta-meta-analysis. *Journal of Clinical Epidemiology*, 68(8), 860–869. doi:10.1016/j.jclinepi.2015.03.017
4. 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
5. Kuper, P., Miguel, C., Cuijpers, P., Apfelbacher, C., Buntrock, C., Karyotaki, E., ... Harter, M. (2025). Sample size and geographical region predict effect heterogeneity in psychotherapy research for depression: A meta-epidemiological study. *Journal of Clinical Epidemiology*, 183, 11179. doi:10.1016/j.jclinepi.2025.11179

21

References [2]

6. Blowes, S. A. (2026). Location-scale models and cross validation to advance quantitative evidence synthesis. *Ecology*, 107(1), e70303. doi:10.1002/ecy.70303
7. Jansen, K., & Nestler, S. (in press). Multivariate location-scale models for meta-analysis. *Multivariate Behavioral Research*. doi:10.1080/00273171.2026.2643868
8. Bangert-Drowns, R. L., Hurlley, 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
9. 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
10. Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, 10(1), 101–129. doi:10.2307/3001666
11. 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

22

Thank You for Your Attention!

Questions, Comments, Suggestions?

✉ wolfgang.viechtbauer@maastrichtuniversity.nl

🌐 <https://www.wvbauer.com>

🌐 <https://www.metafor-project.org>

🐦 [@wiechtb](https://twitter.com/wiechtb) 🐦 [@wiechtb](https://twitter.com/wiechtb)

23

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

24

Likelihood Ratio Test

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

[i] 107.1061

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

25

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:

- factor(id)1 - factor(id)2 = 0
- 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

26

Other Tests

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

27

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

28

Study-Specific τ_i^2 Estimates

study	ML	REML	study	ML	REML
Ashworth, 1992	0.2219	0.2216	Licata, 1993	0.0077	0.0083
Ayers, 1993	0.6131	0.6161	Lodholz, 1980	0.0078	0.0091
Baisch, 1990	0.0602	0.0619	Millican, 1994	0.2137	0.2135
Baker, 1994	0.0034	0.0047	Moynihan, 1994	0.4213	0.4206
Davis, JJ, 1996	0.3290	0.3284	Mulvaney, 1991	0.1246	0.1266
Dipillo, 1994	0.1313	0.1314	Radmacher, 1995	0.8917	0.8903
Ganguli, 1989	0.1022	0.1021	Reaves, 1991	0.0298	0.0313
Guckin, 1992	0.0443	0.0445	Rivard, 1996	0.0972	0.0994
Guckin, 1992	0.1494	0.1492	Rodgers, 1996	0.0000	0.0007
Horton et al., 1985	0.0952	0.0953	Ross & Faucette, 1994	0.0834	0.0830
Hyser, 1992	0.1482	0.1480	Sharp, 1987	0.1056	0.1057
Johnson, LA, 1991	0.1242	0.1241	Stewart, 1992	0.1542	0.1540
Kasperek, 1993	0.0078	0.0081	Weiss & Walters, 1980	0.1027	0.1025
Langer & Applebee, 1987	0.0205	0.0220	Willey, 1988	1.7243	1.7221

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

29

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

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

30

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
```

31

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
```

32

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 b th 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)

33

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; tau² 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

34

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
```

35