**Journal of Modern Applied Statistical Methods**

Volume 2 | Issue 1 Article 7

5-1-2003

# A Parametric Bootstrap Version of Hedges' Homogeneity Test

Wim Van den Noortgate
*Katholieke Universiteit*, Wim.VandenNoortgate@ped.kuleuven.ac.be

Patrick Onghena
*Katholieke Universiteit*, Patrick.Onghena@ped.kuleuven.ac.be

Journal of Modern Applied Statistical Methods
May 2003, Vol. 2, No.1, 73-79

# A Parametric Bootstrap Version of Hedges' Homogeneity Test

Wim Van den Noortgate Patrick Onghena
Katholieke Universiteit Leuven, Belgium

Hedges' *Q*-test is frequently used in meta-analyses to evaluate the homogeneity of effect sizes, but for several kinds of effect size measures it does not always appropriately control the Type 1 error probability. Therefore we propose a parametric bootstrap version, which shows Type 1 error control under a broad set of circumstances. This is confirmed in a small simulation study.

Key words: *Q*-test, homogeneity, effect sizes, parametric bootstrap, Type 1 error

## Introduction

A meta-analysis cumulates the findings of previous research. Often fixed effects techniques are used to summarize the findings of several studies into one single result. The individual effect size estimates are averaged (usually with each effect size weighted by the size of the study or by the inverse of its sampling variance), to obtain an estimate of the overall effect size. These techniques of course are only appropriate if studies can be assumed to be sharing a common population effect size or if in the meta-analysis no inference to a broader population of effect sizes is aimed at (Hedges & Vevea, 1998).

The suitability of the fixed effects techniques therefore is usually statistically tested by means of a homogeneity test. If effect sizes are found heterogeneous, study characteristics are included in the model as covariates to investigate their moderating effect, resulting in a fixed effects regression model. Alternatively, or in addition to the inclusion of moderator variables, the heterogeneity may be explicitly modeled, by defining random study effects. This results in a random effects model or a random effects regression model (see Raudenbush, 1994, for more details). The homogeneity test thus often plays a crucial role in a meta-analysis, since its results are often used to decide if the simple fixed effects model is to be extended with moderator variables and/or random effects, and fixed effects and random effects meta-analytic models often give dissimilar results (Van den Noortgate & Onghena, in press).

Probably the most frequently used statistical test of the homogeneity of a set of effect sizes is the *Q*-test, which was described by Hedges (1982) and by DerSimonian and Laird (1986) and therefore is often referred to as the Hedges' or the DerSimonian and Laird's homogeneity test, although it was proposed before by Cochran (1954).

The test statistic for this test is calculated as

$$Q = \sum_{i=1}^{k} \frac{\left(t_i - \overline{t}\right)^2}{\hat{\sigma}^2_{\left(t_i\right)}} \qquad (1),$$

Wim Van den Noortgate is a postdoctoral researcher at the Department of Education at the Katholieke Universiteit Leuven (Belgium). His research interests include multilevel analysis, meta-analysis, resampling methods, single-case designs and item response theory. Email: Wim.VandenNoortgate@ped.kuleuven.ac.be.
Patrick Onghena is professor of Educational Methodology and Statistics at the Katholieke Universiteit Leuven (Belgium). His research interests include resampling inference, exact nonparametric inference, multilevel analysis, meta-analysis, and single-case designs. Email: Patrick.Onghena@ped.kuleuven.ac.be.

with $k$ the number of studies, $t_i$ the observed effect size in study $i$, $\bar{t}$ the precision weighted mean of the observed effect sizes, with the (estimated) precision of study $i$ defined as $1/\hat{\sigma}^2_{(t_i)}$, and $\hat{\sigma}^2_{(t_i)}$ the estimate of $\sigma^2_{(t_i)}$, the sampling variance of the observed effect size given the 'true' effect size in study $i$.

Under the null hypothesis of homogeneous effect sizes, $Q$ follows a $\chi^2$ distribution with $k$-1 degrees of freedom, given relatively large study sizes, and given that $\hat{\sigma}^2(t_i)$ is independent of $t_i$ (DerSimonian & Laird, 1986; Takkouche, Cadarso-Suárez, & Spiegelman, 1999).

Although several simulation studies showed the advantages of the $Q$-test compared to other kinds of homogeneity tests (e.g., Baydoun, 1995; Sanchez-Meca & Marin-Martinez, 1997; Takkouche, et al., 1999), using the $Q$-test is not without problems. Besides the problem that the $Q$-test, like other homogeneity tests, suffers from a lack of power (Harwell, 1997; Sanchez-Meca & Marin-Martinez, 1997; Takkouche, et al., 1999), the Type 1 error rate of the $Q$-test is not always under control, since the underlying assumptions are usually only approximately met. The degree of the violation of the assumptions, and therefore the behavior of the homogeneity test, depends on the kind of effect size measure that is used and on the conditions under which it is applied.

The proportion of Type 1 errors for instance was found inflated if the $Q$-test is used for evaluating the homogeneity of correlation coefficients, but close to the nominal level if the correlation coefficients are first transformed to Fisher's z-values (Alexander, Scozzaro, & Borodkin, 1989; Sagie & Koslowsky, 1993; Spector & Levine, 1987). Gavaghan, Moore and McQuay (2000) found a slightly inflated number of Type 1 errors when using the risk difference as a measure of effect size. The results of the $Q$-test for Hedges' $d$ are found highly liberal if used to test the homogeneity of a sample of Hedges' standardized mean differences ($d$), in case within studies the group sizes and population variances are unequal and the smaller group size is associated with the largest population variance (Harwell, 1997). If under both conditions scores are normally distributed with a common variance, the $Q$-test has been shown slightly conservative, especially if the study sizes are relatively small compared to the number of studies (Hedges & Olkin, 1985; Harwell, 1997).

In the following, we present a parametric bootstrap version of the $Q$-test, intended to estimate more closely the reference null distribution of $Q$ in case the $\chi^2$-distribution is inappropriate due to a violation of the underlying assumptions. In a small simulation study, we evaluate the performance of the bootstrap $Q$-test for different conditions and different effect size measures.

## Methodology

### A Parametric Bootstrap Version of the Q-test

In the bootstrap, the empirical data are used to estimate the population distribution(s), and samples are simulated from the estimated distribution(s) in order to approximate the sampling distribution of a certain quantity. For the application of the bootstrap procedure to the $Q$-test we propose the following procedure:

1. Perform a meta-analysis using techniques for fixed effects models (Hedges & Olkin, 1985), calculate and store the $Q$-statistic.
2. Simulate new raw data that could have been observed under the null hypothesis of homogeneity (see below).
3. Calculate for the simulated data of each study the measure of effect size that was used in the initial meta-analytic data set.
4. Perform a meta-analysis on those new effect sizes, calculate and store the $Q$-statistic.
5. Repeat step 2-4 a large number of times $B$, for instance 1000.
6. Compare the initial $Q$-value with the empirical distribution of $Q$-values from the $B$ bootstrap samples. The bootstrap $p$-value is the proportion of the $Q$-values that is larger than or equal to the initial $Q$-value.

In step 2, new raw data are sampled from the estimated population distributions, holding constant the study sizes and the number of studies. A general principle for estimating the population distributions is that for each study the population distributions must show the same effect size (fulfilling the null hypothesis of homogeneity). Furthermore, the population distributions are estimated based on the initial data and additional assumptions. The estimation of the distributions can easily be adapted according to the measure of effect size that is used and to the assumptions one is willing to make.

We give some examples. First, suppose the correlation coefficient is used as the measure of effect size, and data can be assumed bivariate normal. In this case, we can draw new raw data for each study from a bivariate normal distribution. Since the data are used only to calculate the correlation coefficient, means and variances of the distributions can be chosen freely. The population correlation for each bivariate normal distribution is set equal to the overall estimated correlation coefficient. One could for instance draw new data from bivariate normal distributions with zero mean, variances equal to 1 and a covariance equal to the estimated overall correlation coefficient.

As another example, suppose the risk difference or the difference between proportions is used as the effect size. If for each study the proportions for both groups can be retrieved (as is often the case), we can estimate the population proportions under both conditions by means of a precision weighted mean of the observed proportions, assuming equal population proportions in each study. For the bootstrap samples, new data are sampled for each study from two Bernoulli distributions, defined by the estimated population proportions.

Third, if the standardized mean difference is used as a measure of effect size, and raw data under both conditions can be assumed normally distributed with a common variance, for each study data are drawn from two normal distributions with the same variance, and with standardized mean difference that is the same for each study. This standardized mean difference is estimated by the precision-weighted average of the observed effect sizes. One could for instance draw data from $N(\bar{\mathrm{d}}, 1)$ and $N(0,1)$ for both groups respectively. Note that drawing data from normal distributions with other variances and means will not alter the results, as long as the variances are equal and the effect size is unchanged, since the raw data are used only to calculate the standardized mean difference.

The situation is somewhat more complicated if the population variances under both conditions cannot be assumed equal. If in the studies the observed within group variance estimates are reported, for study $i$ these are $\hat{s}^2_{\mathrm{Ai}}$ and $\hat{s}^2_{\mathrm{Bi}}$, one can calculate the pooled within group variance estimate for each study (Hedges, 1981). Multiplying the square root of this pooled variance with the estimated mean standardized mean difference estimate, results for study *i* in the estimated study-specific unstandardized mean difference, $Est(\mu_{\mathrm{Ai}} - \mu_{\mathrm{Bi}})$. Raw data can subsequently be drawn from $N(Est(\mu_{\mathrm{Ai}} - \mu_{\mathrm{Bi}}), \hat{s}^2_{\mathrm{Ai}})$ and $N(0, \hat{s}^2_{\mathrm{Bi}})$.

### A Simulation Study

In order to evaluate the parametric bootstrap version of the *Q*-test, we compared its results with the results of the ordinary *Q*-test, by means of a small simulation study. Here we show the results of both homogeneity tests for relatively extreme situations, in which (as described above) the ordinary *Q*-test has been shown in previous research failing to keep the proportion of Type 1 errors under control. More specifically, we simulated:

- sets of correlation coefficients,
- sets of risks differences,
- sets of standardized mean differences with small group sizes paired with large population variances (called negative paired variances and group sizes by Harwell, 1997),
- large sets of standardized mean differences stemming from small studies, and
- sets of values ("effect sizes") sampled from a normal distribution, with sampling variances independent of the effect sizes, intended as a control condition (see below).

The characteristics of the simulated data sets are summarized in Table 1. The values are chosen such that the situations are comparable with those discussed in previous research. For each of the five situations, we simulated 1000 homogeneous as well as 1000 heterogeneous data sets, 10 000 in total, making possible the assessment of both the proportion of Type 1 and Type 2 errors. The bootstrap as well as the ordinary *Q*-test was used for each set to evaluate its heterogeneity. For each data set, we drew 1000 bootstrap samples and calculated *Q* for each sample in order to approximate its null distribution. Bootstrap samples were drawn as described above. (Table 1 appears on following page.)

Based on the results of previous research described above, we expect that the proportion of Type 1 errors when using the ordinary *Q*-test will be too high in the first three situations, while it will be lower than the nominal level in the fourth situation. When sampling effect sizes from a normal distribution (with a variance that is independent of the effect size), we expect that the proportion of Type 1 errors will be close to the nominal level.

In Figure 1 (following page), histograms present the distributions of the *p*-values resulting from the ordinary *Q*-test and the bootstrap *Q*-test in case of homogeneous data. If the reference distribution is close to the true null distribution, we expect an approximately uniform distribution of the *p*-values. This means that under the null hypothesis, we expect that 1% of the *p*-values will be smaller than .01, 5% smaller than .05, 10 % smaller than .10 and so on, or otherwise stated, that regardless of the nominal $\alpha$-level chosen, the nominal and the actual $\alpha$-level correspond.

As expected, the distribution of the *p*-values for the ordinary *Q*-test is skewed in the first four situations. The ordinary *Q*-test gives too much relatively small *p*-values when using *r*, when using risk differences, or when using *d* in case *n* and the within group variance are negatively paired, while it yields too much relatively large *p*-values when using *d* with a small N/k ratio. This means that for a homogeneous set of effect sizes, the null hypothesis of homogeneity is too often rejected in the first three situations, but less than optimal in the fourth situation. As an example, in Table 2 the proportion of Type 1 errors is presented for a nominal level of .05. Note that in case the sampling variance of the effect sizes is independent of the effect sizes, the distribution of the *p*-values is approximately uniform, and the proportion of Type 1 errors is near to the nominal level.

Figure 1 and Table 2 (following page) furthermore reveal that the *p*-values of the bootstrap procedure are approximately uniformly distributed in all situations, yielding a relatively accurate proportion of Type 1 errors, although there seems to be a slightly liberal tendency.

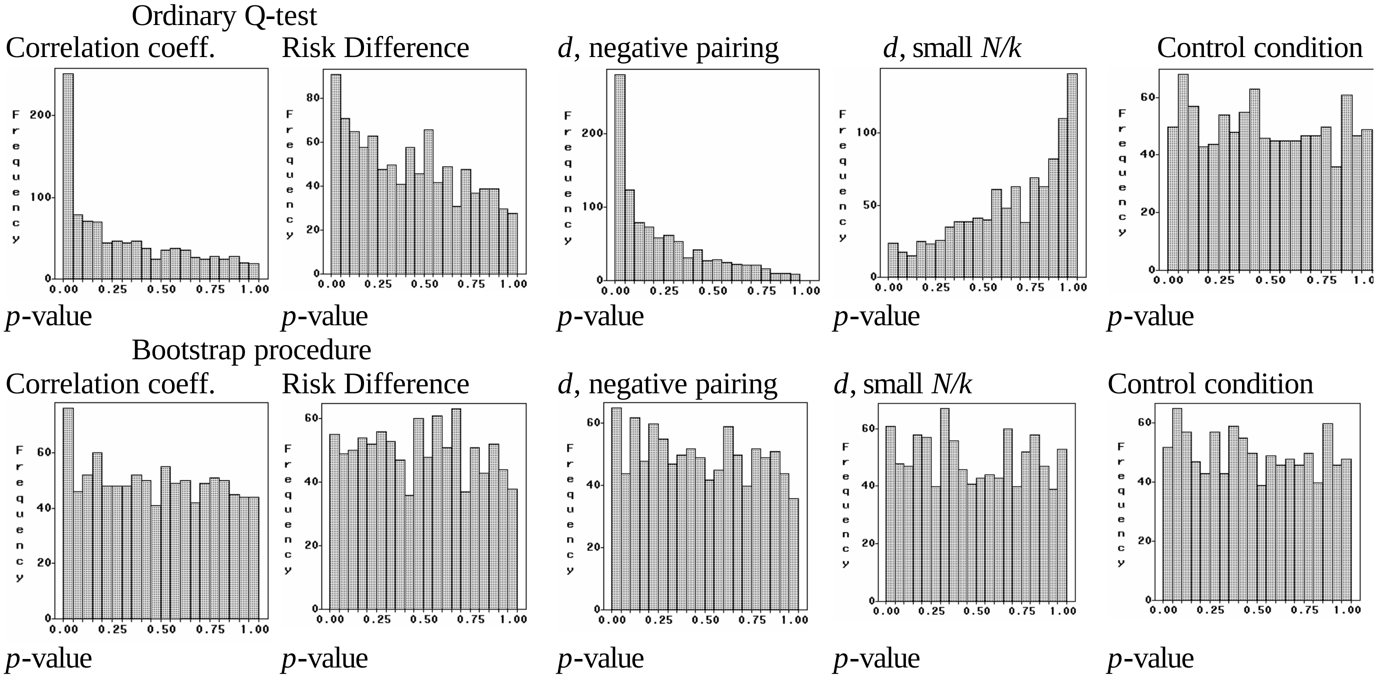
In Table 3, we see that both procedures are equally powerful when testing a set of normally distributed effect sizes with sampling variances that are independent of the effect sizes. In other situations, it is difficult to compare the power of both procedures, because for the ordinary *Q*-test the rejection rates are biased since the proportion of Type 1 errors is not under control. Anyway, we see that using the bootstrap procedure instead of the ordinary procedure affects the proportion of rejections in the same way in the homogeneous and the heterogeneous case. In case the *Q*-test is used for testing the homogeneity of a set of correlation coefficients, of a set of risk differences, or of a set of standardized mean differences with small group sizes paired with large variances, the proportion of rejections is lower if the bootstrap version is used. In contrast, the bootstrap version of the *Q*-test rejects the null hypothesis more often if the homogeneity of a large set of standardized mean differences stemming from small studies is tested.

## Conclusion

Although the *Q*-test is very often used in meta-analysis to test the homogeneity of effect sizes, it has been shown in previous research that in several situations the test fails to keep the proportion of Type 1 errors under control. In this article, we therefore present a parametric bootstrap version of the test, which allows freeing one or more assumptions underlying the *Q*-test or the calculation of the effect size measures and their sampling distribution. The results of a small simulation study suggest that even in situations where the ordinary *Q*-test does not succeed controlling the proportion of Type 1 errors, the Type 1 error rate for the bootstrap version is still close to the nominal level.

Table 1. Characteristics of the simulated data sets.

| | | | Population distribution | | |
|---|---|---|---|---|---|
| | | | Homogeneous case | Heterogeneous case | |
| | $K$ | N | | 80 % | 20 % |
| Correlation coefficient | 50 | $N$= 20 | Raw data<br>$\approx N(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & \\ .50 & 1\end{bmatrix})$ | Raw data<br>$\approx N(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & \\ .45 & 1\end{bmatrix})$ | Raw data<br>$\approx N(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & \\ .55 & 1\end{bmatrix})$ |
| Risk difference | 50 | $n_A = n_B = n = 50$ | Data group A<br>≈ Bin(.2, 1)<br>Data group B<br>≈ Bin(.5, 1) | Data group A<br>≈ Bin(.2, 1)<br>Data group B<br>≈ Bin(.45, 1) | Data group A<br>≈ Bin(.2, 1)<br>Data group B<br>≈ Bin(.55, 1) |
| Hedges' *d*, negative pairing | 50 | $n_A = 10$<br>$n_B = 20$ | Data group A<br>$\approx N(0.6,\ 2)$<br>Data group B<br>$\approx N(0,\ 1)$ | Data group A<br>$\approx N(0.3,\ 2)$<br>Data group B<br>$\approx N(0,\ 1)$ | Data group A<br>$\approx N(1,\ 2)$<br>Data group B<br>$\approx N(0,\ 1)$ |
| Hedges' *d*, small N/k | 100 | $n_A = n_B = n = 5$ | Data group A<br>$\approx N(0.5,\ 1)$<br>Data group B<br>$\approx N(0,\ 1)$ | Data group A<br>$\approx N(0.1,\ 1)$<br>Data group B<br>$\approx N(0,\ 1)$ | Data group A<br>$\approx N(0.8,\ 1)$<br>Data group B<br>$\approx N(0,\ 1)$ |
| Control condition | 50 | $n_A = n_B = n = 10$ | Effect size<br>$\approx N(0.5,\ 2/n)$ | Effect size<br>$\approx N(0.3,\ 2/n)$ | Effect size<br>$\approx N(0.8,\ 2/n)$ |


Ordinary Q-test

Correlation coeff. Risk Difference *d*, negative pairing *d*, small *N/k* Control condition

Bootstrap procedure

Correlation coeff. Risk Difference *d*, negative pairing *d*, small *N/k* Control condition

*Figure 1.* Distribution of the *p*-values in case of true homogeneity

Table 2. Rejection rates of the null hypothesis (with a nominal $\alpha$ of .05) in the homogeneous case (proportion Type 1 errors).

| | Correlation coefficient | Risk Difference | *d*, negative pairing | *d*, small *N/k* | Control condition |
|---|---|---|---|---|---|
| Ordinary | .251 | .091 | .280 | .024 | .050 |
| Bootstrap | .076 | .055 | .065 | .061 | .052 |

Table 3. Rejection rates of the null hypothesis (with a nominal $\alpha$ of .05) in the heterogeneous case (power).

| | Correlation coefficient | Risk Difference | *d*, negative pairing | *d*, small *N/k* | Control condition |
|---|---|---|---|---|---|
| Ordinary | .720 | .349 | .731 | .116 | .247 |
| Bootstrap | .347 | .258 | .367 | .302 | .252 |

Moreover, in case the assumptions of the ordinary *Q*-test are met, and the test yields appropriate Type 1 error rates, the bootstrap version seems to be equally powerful. A disadvantage of the bootstrap version of the test is that for some situations additional data are required, that may not always be available. E.g., for testing the homogeneity of a set of risk differences, the proportions for each of the groups must be available.

Based on the encouraging results of our simulation study, we suggest comparing the *Q*-statistic to the approximate null distribution based on the bootstrap, rather than to a $\chi^2$-distribution, whenever possible. Meanwhile however, we note that the power of both versions of the homogeneity test is low and recommend a prudent use of the tests in both modeling and evaluating the heterogeneity.

## References

Alexander, R.A., Scozzaro, M.J., & Borodkin, L.J. (1989). Statistical and empirical examination of the chi-square test for homogeneity of correlations in meta-analysis. *Psychological Bulletin, 106*, 329-331.

Baydoun, R.B. (1995). A Monte Carlo investigation of the Type 1 error rate and power of the Hedges and Olkin moderator search method. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 55,* 4152.

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

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

Gavaghan, D.J., Moore, R.A., & McQuay, H.J. (2000). An evaluation of homogeneity tests in meta-analyses in pain using simulations of individual patient data. *Pain, 85*, 415-424.

Harwell, M. (1997). An empirical study of Hedges's homogeneity test. *Psychological Methods, 2,* 219-231.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107-128.

Hedges, L. V. (1982). Fitting categorical models to effect sizes from a series of experiments. *Journal of Educational Statistics, 7,* 119-137.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando : Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3,* 486-504.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301-321). New York: Russell Sage Foundation.

Sagie, A., & Koslowsky, M. (1993). Detecting moderators with meta-analysis: an evaluation and comparison of techniques. *Personnel Psychology, 46*, 629-640.

Sanchez-Meca, J., & Marin-Martinez, F. (1997). Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type 1 error. *Quality and Quantity, 31*, 385-399.

Spector, P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: a Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology, 72*, 3-9.

Takkouche, B., Cadarso-Suarez, C., & Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology, 150*, 206-215.

Van den Noortgate, W., & Onghena, P. (In press). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement*.