Refining the Test of Excess Significance for detecting selective publication in meta-analysis

James E. Pustejovsky[1]

[1] University of Texas at Austin

Author Note

James E. Pustejovsky, Educational Psychology Department, University of Texas at Austin.

Correspondence concerning this article should be addressed to James E. Pustejovsky, 1912 Speedway, MS D5800, Austin, TX 78712. E-mail: pusto@austin.utexas.edu

Abstract

Publication bias and other forms of selective outcome reporting are important threats to the validity of findings from research syntheses—even undermining their special status for informing evidence-based practice and policy guidance. An array of methods have been proposed for detecting selective publication. In particular, Ioannidis and Trikalinos (2007) proposed the Test of Excess Significance (TES), which diagnoses publication bias by comparing the observed number of statistically significant effect sizes to the number expected based on the power of included studies to detect the estimated average effect. Another approach is based on explicit modeling of the selective publication process, as in the weight function model developed by Hedges (1992) and Vevea and Hedges (1995). There is a close connection between these two methods: TES is based on the score function of a simple form of the weight function model. This connection motivates some refinements to TES that improve its operating characteristics and allow for between-study heterogeneity through random effects and regression on study characteristics. After describing the refined tests, I report a small simulation evaluating their calibration and power compared to conventional TES, a likelihood ratio test based on the weight function model, and p-uniform.

Keywords: keywords

Word count: X

Refining the Test of Excess Significance for detecting selective publication in meta-analysis

Systematic reviews and quantitative research syntheses now lie at the heart of debates about scientific theories and guidance about evidence-based policy.

It has been argued that tests of publication bias are irrelevant and unnecessary because there is overwhelming evidence that selective publication is at work across multiple scientific fields (Morey, 2013). Need for powerful tests of selective publication.

The remainder of the article proceeds as follows. In the next section, I briefly review the TES and Vevea-Hedges selection model before developing the connection between the two methods and proposing a refinement to TES. The following section reports a small Monte Carlo simulation evaluating the size and power of the proposed method. A brief discussion section highlights limitations and future directions.

A selection of tests for selective publication

Consider a meta-analysis of k studies, where a conventional random effects model might be applied. Let T_i denote the effect size estimate from study i, with standard error σ_i, each for $i = 1, ..., k$. Let θ_i denote the true effect size parameter from study i. I shall assume that the included studies are large enough that it is reasonable to treat T_i as following a normal distribution: $T_i \sim N(\theta_i, \sigma_i^2)$. Under a random effects model, and in the absence of selective publication, the true effects are assumed to follow a normal distribution with mean μ and standard deviation τ.

Let α denote the conventional type-I error level used for one-sided hypothesis tests; in areas of research that typically use two-sided tests and the .05 level to determine significance, the one-sided level would be $\alpha = .025$. Let $\Phi(x)$ and $\phi(x)$ denote the standard normal cumulative distribution and density function, respectively. Let z_α denote the standard normal α-level critical value, that is, $z_\alpha = \Phi^{-1}(1 - \alpha)$. Let O_i be an indicator for the statistical significance of study i, so that $O_i = 1$ when $T_i/\sigma_i > z_\alpha$ and $O_i = 0$ otherwise.

Test of Excess Significance

As a test of selective publication, Ioannidis and Trikalinos (2007) proposed to compare the number of statistically significant effects among the set of k studies to the number of significant effects expected if there were no selection. The observed number of significant effects is $O = \sum_{i=1}^{K} O_i$. In the absence of selection, the expected number of effects is the sum of the power of each study to detect a true effect of a given size. Letting $P_i(\mu, \tau^2)$ denote the power of study i under a random effects model, which is equal to

$$P_i(\mu, \tau^2) = 1 - \Phi\left(\frac{\sigma_i z_\alpha - \mu}{\sqrt{\tau^2 + \sigma_i^2}}\right). \tag{1}$$

The expected number of significant effects is then $E(\mu, \tau^2) = \sum_{i=1}^{k} P_i(\mu, \tau^2)$, a quantity that depends on the unknown average effect μ and between-study heterogeneity τ. Ioannidis and Trikalinos (2007) suggested estimating expected power based on a fixed effect meta-analysis, taking $\hat{E} = E(\hat{\mu}_F, 0)$, where $\hat{\mu}_F$ is the usual fixed effect average. They justify this approach by arguing that heterogeneity is often negligible and that, if there is selective publication, the fixed effect average is less biased than the random effects average. Subsequent applications of TES have typically followed this approach. Is this true?

Ioannidis and Trikalinos (2007) proposed two approximate tests for drawing an inference about whether the set of included studies has been selected for statistical significance. First, they suggest using the test statistic

$$A = \frac{(O - \hat{E})^2}{\hat{E}(k - \hat{E})/k}, \tag{2}$$

compared to a χ_1^2 reference distribution. Alternately, they suggest using a binomial test, comparing O to a binomial reference distribution with size k and probability $\hat{E}/k$. Applications of TES have typically followed the latter approach. Is this true?

Both variants of TES involve approximations. One approximation arises from treating the expected number of studies as known with certainty, whereas in practice it must be

estimated. Further, using a binomial reference distribution amounts to assuming that power is constant across studies, which will not be the case unless all included studies are equally precise. Calculating $\hat{E}$ under a fixed effect model entails the further assumption that between-study heterogeneity is negligible (Johnson & Yuan, 2007). Thus, one might expect that the type I error rate of the tests may be distorted when studies vary in precision or are truly heterogeneous. Indeed, van Assen, van Aert, and Wicherts (2015) reported simulation results in which TES had below-nominal type I error, even when all studies are equally precise.

TES is an exploratory test for selective publication, intended to be used as a signal that a body of evidence may be unrepresentative (Ioannidis, 2013). It does not, however, invoke any particular model of the selection process. In contrast, other approaches are based on specific models of selective publication.

Weight function selection models

Many meta-analytic models have been developed that make specific assumptions about the process of selective publication. One such class of models, often called "weight function" models, assume that the probability of publication depends on a piece-wise constant function of the statistical significance of the effect size estimate. Building on earlier work by Iyengar and Greenhouse (1988), Hedges (1992) and Dear and Begg (1992) proposed weight function models that allow for heterogeneity in true effect sizes. Vevea and Hedges (1995) further developed the approach to allow for moderators of effect size through a meta-regression model. Based on several extensive Monte Carlo simulations, a very simple, three-parameter version of the weight function model has recently been highlighted as a promising technique for dealing with selective publication (Carter, Schönbrodt, Gervais, & Hilgard, 2018; McShane, Böckenholt, & Hansen, 2016). I limit consideration to this three-parameter model because it is mostly directly connected to TES. I discuss a more general form of the weight function model in the Appendix.

The weight function model involves two components: a sampling model and a selection model. Following Hedges (1992), the sampling model assumes that effect size estimates follow a basic random effects model as outlined previously. However, not all effect sizes are published (or more generally, not all effect sizes are available for inclusion in the meta-analysis). Rather, the probability that an effect size estimate is included is a multiple of the weight function

$$w(T_i, \sigma_i) = \begin{cases} 1 & \text{if} \quad T_i > \sigma_i z_\alpha \\ \\ \pi & \text{if} \quad T_i \leq \sigma_i z_\alpha \end{cases} \tag{3}$$

where $\pi \geq 0$ is the probability that a statistically insignificant effect size is included, relative to the inclusion probability for an equally precise, statistically significant effect size. Note that if $\pi = 1$, then all studies appear with equal probability and there is no selective publication.

The weight function model and tests associated with it are usually based on maximum likelihood estimation models. Assuming that studies are mutually independent, the joint likelihood of the weight function model is

$$\mathcal{L}(\mu, \tau^2, \pi) = \prod_{i=1}^{k} \frac{w(T_i, \sigma_i)\phi\left(\frac{T_i - \mu}{\sqrt{\tau^2 + \sigma_i^2}}\right)}{\sqrt{\tau^2 + \sigma_i^2} A_i(\mu, \tau^2, \pi)}, \tag{4}$$

where A_i is a normalizing constant given by

$$A_i(\mu, \tau^2, \pi) = 1 - (1 - \pi)\Phi\left(\frac{\sigma_i z_\alpha - \mu}{\sqrt{\tau^2 + \sigma_i^2}}\right) = P_i(\mu, \tau^2) + \pi\left[1 - P_i(\mu, \tau^2)\right],$$

with $P_i(\mu, \tau^2)$ as given in (1). The log likelihood is thus (up to a constant):

$$l(\mu, \tau^2, \pi) = \sum_{i=1}^{k} \ln w(T_i, \sigma_i) - \frac{1}{2}\sum_{i=1}^{k} \frac{(T_i - \mu)^2}{\tau^2 + \sigma_i^2} - \frac{1}{2}\ln(\tau^2 + \sigma_i^2) - \sum_{i=1}^{k} \ln A_i(\mu, \tau^2, \pi). \tag{5}$$

Let $\hat{\mu}$, $\hat{\tau}^2$, and $\hat{\pi}$ denote the values that maximize (5)—that is, the maximum likelihood estimates of the model parameters.

Hedges (1992) proposed to a test for the null hypothesis that $\pi = 1$ (i.e., no selective publication) using a likelihood ratio criterion. Let $\hat{\mu}_R$ and $\hat{\tau}_R^2$ denote the values that

maximize (5) when π is set equal to 1. The likelihood ratio test statistic is then

$$G^2 = 2\left[l(\hat{\mu}, \hat{\tau}^2, \hat{\pi}) - l(\hat{\mu}_R, \hat{\tau}^2_R, 1)\right], \tag{6}$$

which is compared to a χ^2_1 reference distribution.

In practice, a difficulty with the three-parameter weight function model is that maximum likelihood estimates do not converge when all included studies are statistically significant or when no included studies are statistically significant at level α. If one of these conditions occurs, a researcher might choose to adjust the α level defining statistical significance so that at least one study is statistically significant and at least one is statistically insignificant. I implement this ad hoc modification when evaluating the operating characteristics of the likelihood ratio test in the Monte Carlo simulations.

Are both of these conditions correct

Refined excess significance tests

TES and the three-parameter weight function model are closely connected, in TES is based on the null score of the weight function model. The score function is the derivative of the log likelihood with respect to its parameters. Note that the derivative of (5) with respect to π is

$$S^{\pi}(\mu, \tau^2, \pi) = \frac{\partial l}{\partial \pi} = \frac{1}{\pi}\sum_{i=1}^{k}(1 - S_i) - \sum_{i=1}^{k}\frac{1}{A_i}\Phi\left(\frac{\sigma_i z_\alpha - \mu}{\sqrt{\tau^2 + \sigma_i^2}}\right). \tag{7}$$

Under the null hypothesis of $\pi = 1$, $A_i = 1$ for $i = 1, ..., k$ and the score simplifies to

$$\begin{aligned} S^{\pi}(\mu, \tau^2, 1) &= \sum_{i=1}^{k}(1 - O_i) - \sum_{i=1}^{k}\Phi\left(\frac{\sigma_i z_\alpha - \mu}{\sqrt{\tau^2 + \sigma_i^2}}\right) \\ &= (k - O) - \sum_{i=1}^{k}\left[1 - P_i(\mu, \tau^2)\right] \\ &= E(\mu, \tau^2) - O. \end{aligned}$$

Thus, the score of the three-parameter weight function model, evaluated under the null, is equivalent to the discrepancy between the expected and observed number of significant effects—the same statistic that is used to construct TES. TES is typically calculated under a

fixed effects model, in which case the discrepancy is $O - \hat{E} = -S^{\pi}(\hat{\mu}_{FE}, 0, 1)$. If expected power is instead calculated using the maximum likelihood estimates under a random effects model, then $O - E(\hat{\mu}_R, \hat{\tau}^2_R) = -S^{\pi}(\hat{\mu}_R, \hat{\tau}^2_R, 1)$.

This connection to the weight function model suggests that TES could be refined using score tests, a standard tool from mathematical statistics. Score tests (Rao, 1948) are asymptotically equivalent to likelihood ratio tests, but have the advantage that they do not require obtaining maximum likelihood estimates under the unrestricted model (Boos, 1992). This feature is advantageous in the present context because it allows one to circumvent potential convergence problems with the weight function model. Two forms of score tests are available—a parametric form and a robust form—which use different approaches to estimating the variance of the null score.

The parametric score test (Rao, 1948) uses the Fisher information matrix to estimate the variance of the null score. Denote the full parameter vector of the weight function model as $\boldsymbol{\theta} = (\mu, \tau^2, \pi)$. Let $\boldsymbol{S}$ be the full score vector, so $\boldsymbol{S}(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$, with $\tilde{\boldsymbol{S}}_R = \boldsymbol{S}(\hat{\mu}_R, \hat{\tau}^2_R, 1)$. Let $\mathcal{I}$ denote the Fisher information matrix:

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right),$$

where $\mathbb{E}$ denotes the expectation over $T_1, ..., T_k$, and denote $\tilde{\mathcal{I}}_R = \mathcal{I}(\hat{\mu}_R, \hat{\tau}^2_R, 1)$. The parametric score test statistic is then given by

$$X^2_P = \tilde{\boldsymbol{S}}'_R \tilde{\mathcal{I}}_R \tilde{\boldsymbol{S}}_R.$$

This quantity can be expressed more directly as:

$$X^2_P = \frac{\left[E(\hat{\mu}_R, \hat{\tau}^2_R) - O\right]^2}{V_{\pi} - U_{\mu} - U_{\tau^2}}, \tag{8}$$

where

$$V_\pi = \sum_{i=1}^{k} P_i(\hat{\mu}_R, \hat{\tau}_R^2)\left[1 - P_i(\hat{\mu}_R, \hat{\tau}_R^2)\right],$$
$$U_\mu = \left[\sum_{i=1}^{k} \frac{1}{\hat{\tau}_R^2 + \sigma_i^2}\right]^{-1} \left[\sum_{i=1}^{k} \frac{\phi(c_i)}{\sqrt{\hat{\tau}_R^2 + \sigma^2}}\right]^2,$$
$$U_{\tau^2} = \left[\sum_{i=1}^{k} \frac{1}{(\hat{\tau}_R^2 + \sigma_i^2)^2}\right]^{-1} \left[\sum_{i=1}^{k} \frac{c_i\phi(c_i)}{\hat{\tau}_R^2 + \sigma^2}\right]^2,$$

and $c_i = (\sigma_i z_\alpha - \hat{\mu}_R)/\sqrt{\hat{\tau}_R^2 + \sigma_i^2}$. The null hypothesis of $\pi = 1$ is rejected if X_P^2 exceeds a χ_1^2 critical value. Alternately, a one-sided, level-α_S test can be calculated by taking

$$Z_P = \frac{E(\hat{\mu}_R, \hat{\tau}_R^2) - O}{\sqrt{V_\pi - U_\mu - U_{\tau^2}}} \tag{9}$$

and rejecting if $Z_P < \Phi(\alpha_S)$.

This parametric score test requires using the maximum likelihood estimates (under the restriction of the null hypothesis). Alternative forms of the score test have been described by Kent (1982), White (1982), and Engle (1984), among others (see Boos, 1992 for a review). These generalized, or robust, forms provide more flexibility in how μ and θ may be estimated, assuming that there is not selective publication.

Consider estimators $\acute{\mu}$ and $\acute{\tau}^2$ that are defined as the solutions of estimating equations

$$S^\mu\left(\mu, \tau^2\right) = \sum_{i=1}^{k} S_i^\mu\left(\mu, \tau^2\right) = 0,$$
$$S^{\tau^2}\left(\mu, \tau^2\right) = \sum_{i=1}^{k} S_i^{\tau^2}\left(\mu, \tau^2\right) = 0,$$

where the estimating equation for μ has the form

$$S_i^\mu(\mu, \tau^2) = w_i(T_i - \mu)$$

for some set of weights $w_1, ..., w_k$, which may depend on τ^2. The usual fixed effect estimator uses $w_i = \sigma_i^{-2}$; the random effects estimator uses $w_i = (\tau^2 + \sigma_i^2)^{-1}$. For the between-study heterogeneity, many of the available estimators of τ^2 can be expressed as solutions to

estimating equations. For example, the maximum likelihood estimator uses

$$S_i^{\tau^2}(\mu, \tau^2) = \frac{1}{\tau^2 + \sigma_i^2}\left[\frac{(T_i - \mu)^2}{\tau^2 + \sigma_i^2} - 1\right].$$

and the restricted maximum likelihood estimator uses

$$S_i^{\tau^2}(\mu, \tau^2) = \frac{1}{\tau^2 + \sigma_i^2}\left[\frac{(T_i - \mu)^2}{\tau^2 + \sigma_i^2} + \frac{1}{(\tau^2 + \sigma_i^2)\sum_{i=1}^k (\tau^2 + \sigma_i^2)^{-1}} - 1\right].$$

In conjunction with estimating equations for μ and τ^2, a robust score test can be defined based on the null score function from the three-parameter weight function model:

$$S^\pi\left(\mu, \tau^2\right) = E(\mu, \tau^2) - O = \sum_{i=1}^k S_i^\pi\left(\mu, \tau^2\right)$$

where $S_i^\pi(\mu, \tau^2) = P_i(\mu, \tau^2) - O_i$. Let $\mathcal{I}_\mu^\mu(\mu, \tau^2) = -\mathbb{E}(\partial S^\mu / \partial \mu)$, $\mathcal{I}_\mu^\pi(\mu, \tau^2) = -\mathbb{E}(\partial S^\pi / \partial \mu)$, etc., where all expectations are taken under the null with $\pi = 1$. Let F_i denote the efficient score function (Tsiatis, 2006), given by

$$F_i = S_i^\pi - \left(\frac{\mathcal{I}_\mu^\pi}{\mathcal{I}_\mu^\mu}\right) S_i^\mu - \left(\frac{\mathcal{I}_\mu^\pi}{\mathcal{I}_\mu^\mu}\right) S_i^\mu,$$

with all quantities evaluated at the parameter estimates $\acute{\mu}$ and $\acute{\tau}^2$. A robust score test statistic is then

$$X_R^2 = \frac{\left[E(\acute{\mu}, \acute{\tau}^2) - O\right]^2}{\sum_{i=1}^k F_i^2}, \tag{10}$$

compared to a χ_1^2 reference distribution. Alternately, a one-sided test can be obtained by taking

$$Z_R = \frac{E(\acute{\mu}, \acute{\tau}^2) - O}{\sqrt{\sum_{i=1}^k F_i^2}} \tag{11}$$

and rejecting if $Z_R < \Phi(\alpha_S)$.

The robust form of the score test is attractive because it can be applied with any of an array of estimators for μ and τ^2. For instance, an analyst might prefer to use the fixed effect estimator $\hat{\mu}_F$ for the average effect size because it is less biased than the random effects estimator under selective publication, along with the restricted maximum likelihood

estimator for τ^2 because it is less biased than the maximum likelihood estimator (cf. Henmi & Copas, 2010; Stanley & Doucouliagos, 2015). It is known that the parametric score test and likelihood ratio test have equivalent power, asymptotically, under a fixed alternative model (Rao, 2005). Beyond that, however, it is difficult to determine which test is optimal. To investigate whether either the parametric or robust test offers advantages over the existing TES or the likelihood ratio test from the three-parameter weight function model, and whether $\hat{\mu}_F$ offers any advantage over the use of random effects maximum likelihood estimation, I turn to Monte Carlo simulations.

Size and power comparisons

In this section, I report simulations examining the size (Type-I error rates) and power properties of these tests under the random effects models and a basic form of selective publication—consistent

- TES (FE, chi-sq)
- TES (FE, binom)
- TES (WLS, chi-sq)
- TES (WLS, binom)
- LRT (2-sided)
- LRT (restricting to $\pi \leq 1$)
- GEST (model-based)
- GEST (robust, ML)
- GEST (robust, WLS)

Discussion

Limitations

- Independent effects

- Simulations assume 3PSM. Further worked needed on other selection processes, models with precision-effect confounding, etc.

References

Boos, D. D. (1992). On Generalized Score Tests. *The American Statistician*, *46*(4), 327. doi:10.2307/2685328

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2018). Correcting for bias in psychology: A comparison of meta-analytic methods. doi:10.31234/osf.io/9h3nu

Dear, K. B. G., & Begg, C. B. (1992). An Approach for Assessing Publication Bias Prior to Performing a Meta-Analysis. *Statistical Science*, *7*(2), 237–245.

Hedges, L. V. (1992). Modeling Publication Selection Effects in Meta-Analysis. *Statistical Science*, *7*(2), 246–255. doi:10.1214/ss/1177011364

Henmi, M., & Copas, J. B. (2010). Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine*, *29*(29), 2969–2983. doi:10.1002/sim.4029

Ioannidis, J. P. A. (2013). Clarifications on the application and interpretation of the test for excess significance and its extensions. *Journal of Mathematical Psychology*, *57*(5), 184–187. doi:10.1016/j.jmp.2013.03.002

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials: Journal of the Society for Clinical Trials*, *4*(3), 245–253. doi:10.1177/1740774507079441

Iyengar, S., & Greenhouse, J. B. (1988). Selection Models and the File Drawer Problem. *Statistical Science*, *3*(1), 109–117. doi:10.1214/ss/1177013012

Johnson, V., & Yuan, Y. (2007). Comments on "An exploratory test for an excess of significant findings" by JPA loannidis and TA Trikalinos. *Clinical Trials: Journal of the Society for Clinical Trials*, *4*(3), 254–255. doi:10.1177/1740774507079437

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspectives on Psychological Science*, *11*(5), 730–749. doi:10.1177/1745691616662243

Morey, R. D. (2013). The consistency test does notand cannotDeliver what is advertised: A comment on Francis (2013). *Journal of Mathematical Psychology*, *57*(5), 180–183. doi:10.1016/j.jmp.2013.03.004

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *44*(01), 50. doi:10.1017/S0305004100023987

Rao, C. R. (2005). Score Test: Historical Review and Recent Developments. In N. Balakrishnan, H. N. Nagaraja, & N. Kannan (Eds.), *Advances in Ranking and Selection, Multiple Comparisons, and Reliability: Methodology and Applications* (pp. 3–20). Boston, MA: Birkhäuser Boston. doi:10.1007/0-8176-4422-9_1

Stanley, T. D., & Doucouliagos, H. (2015). Neither fixed nor random: Weighted least squares meta-analysis. *Statistics in Medicine*, *34*(13), 2116–2127. doi:10.1002/sim.6481

Tsiatis, A. A. (2006). The Geometry of Influence Functions. In *Semiparametric Theory and Missing Data* (pp. 21–51). New York, NY: Springer New York. doi:10.1007/0-387-37345-4_3

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293–309. doi:10.1037/met0000025

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. doi:10.1007/BF02294384