Figure S1. Figure S1 discusses distributional assumptions and estimation procedures for Figure 1

$$L = \sum_{i=1}^{k} \ln[\lambda \beta(1,1)(p_i) + (1-\lambda)\beta(a,b)(p_i)] \quad (1)$$

Equation (1) is the log-likelihood of beta-uniform misture for a collection of *k p*-values, where p_i is the *p*-value for the i_{th} test. λ is the probability that a randomly chosen test is drawn from those that test true null hypotheses, and *(1- λ)* is the probabilty that a randomly chosen test is from those testing a true alternative or false null hypothesis. *β(a,b)* is the probability distribution function of the Beta distribution with the first shape parameter, *a*, and the second shape parameter, *b*, which lets us flexibly model *p*-values from testing false null hypotheses. *β(1,1)* is the Beta distribution with both shape parameters restricted to be 1, which is identical to a Uniform distribution of U[0, 1], and represents the theoretical *p*-value distribution from testing true null hypotheses.

We estimate the parameters of the beta distribution of *p*-values using Maximum Likelihood Estimation (MLE), we use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm with a simple box constraint (L-BFGS-B), $0 < \lambda < 1$. We constrain the maximum iteration to 500. The chosen starting point for (λ, *a, b)* is (0.3,0.15, 2). For each numerical optimation procedure, we check the plot of the fitted distribution with the histogram to check if the model falsely converges on a local maxima of the likelihood space. We use R's maximum likelihood estimation package *maxLik 1.2-4*.

First, we collect *p*-values from test and control variable. We do not assume censoring of p-values as in Jager and Leek (2014). We treat *p*-values reported as *, **, <0.001, <0.05, <0.01, as the rounded *p*-values from rounding-up. For example, we assume <0.05 is rounded up from *p*-values lower than 0.05. Next, We estimate the parameters of the beta-uniform distribution in

Equation (1) for test and control varaibles respectively. Third, we plot the estimated expected distribution of *p*-value for test and control variables. Forth, we plot histograms of *p*-values for test and control variables in density using bin size of 0.005. In this histogram, *p*-value from rounded *p*-value such as <0.05 belong to a bar just below 0.05 consistent with the rounding-up assumption.

Table S.1.1 Table S.1.1 discusses estimation procedures for Table 1

For Table 1, we first estimate the expected frequency of *p*-values. We estimate parameters of the beta-uniform mixture in Equation (1) over the range between 0 and 0.105 for each Panel. Then, we numerically integrate the estimated probability distribution function for each interval below and above the critical values. The interval size is 0.005. We obtain a pair of observed frequency and expected frequency for 20 cells. For each cell, we formally test relative excess or shortage using standardized Pearson residual which asymptotically follows a standard normal distribution (Harberman 1973; Agresti 2002). The z-statistic is as follows.

$$z = \frac{O_{ij} - E_{ij}}{[E_{ij}(1 - E_{i+})(1 - E_{+j})]^{1/2}} \tag{2}$$

, where O is observed frequency and E is expected frequency, and i and j indicate row and column of the contingency table respectively, and i+ indicates summation over all columns in i^{th} row, and +j indicates summation over all rows in j^{th} column.

To estimate the Diff (a) – (b) column, we bootstrap the difference statistic. We resample the *p*-values with replacement for 10,000 times. For each resampled *p*-values, we estimate the difference statistic. After 10,000 iteration, we take the simple average and the standard deviation of the 10,000 difference estimates as the bootstrap estimate and the standard error. t-statistic is the bootstrap estimate over the standard error.

Table S.1.2 Table S.1.2 discusses estimation procedures for Table 2

In Table 2, we estimate the false discovery rate for each category of *p*-values. All of papers in our study did not explicitly indicate a threshold for statistical significance. We estimate the false discovery rate, λ in Equation (1), for *p*-values from 0 to 0.05. We include 0.05 as we code *p*-values reported with *, ** or *p*-value <0.05 as 0.05. Excluding these p-value at 0.05 does not alter the magnitude of false discovery rate to a great extent. For example, the false discovery rate of the all *p*-values less than 0.05 is 0.294, but including 0.05 gives 0.305. Our results are robust to inclusion of the borderline value. Benjamini and Hechtlinger (2014) express concerns for estimates of Jager and Leek (2014) that are sensitive to inclusion of 0.05, which is likely to be from sample selection bias(Ioannidis 2014). We do not find sensitivity of including 0.05 and our sample selection of including all papers mitiagte selection bias.

In addition to false discvoery rate, we present the average of log(sample size) for each category of *p*-value. Each *p*-value is associated with the sample size of the test that generated *p*-value. To illustrate the varying size of sample size in each category, we provide simple average of sample size in *p*-values by category.

Our estimate and interpretation addresses concerns for the model-based estimate of false discovery rate. The false discovery rate relies on the identification of uniform distribution in the mixture of *p*-value distribution. Given the risk of confounding due to omitted variable or research design flaws, *p*-value distribution deviates from the uniform and beta component captures such *p*-values. The proportion of uniform distribution, which represents false discovery rate would be under-estimated. When *p*-hacking is intensive, the distribution of *p*-values from testing treu nulls may represent right-skewed distribution, making the detection of uniform compoent less likely. Hence, our measure of false discovery rate is a lower bound estimate. Jager and Leek (2014) show a simulation *p*-hacking that reports the lowest *p*-value from 20 tests and find the false discovery measures are underestimated. The false discovery rate esimtate does not represent an unbiased estimate (Gelman and O’Rourke 2014), and should be interpreted as a lower bound estimate. Also, we hypothesize the source of heterogneity in *p*-value distribution. This mitigates a potential bias of estimating false discvoery rate over the entire sample (Goodman 2014).

Table S.1.3 Table S.1.3 discusses estimation procedures for Table 3

Table 3 regresses false discovery estimates on statistical power estimates and average Log(sample size), and characteristics of bootstrap samples by categories of *p*-values. For each category, we generate 100 bootstrap sample and estimate false discovery, power, and mean of log(sample size) for each sample. The seed number is set to 4,912,392 for the reported results. We use bootstrap samples to identify the variations in false discovery rate holding one category at a time. For example, *p*-values from Test variables consists of *p*-values from financial, managerial accounting, auditing, and tax, rounded and non-rounded *p*-values. Variation in Test *p*-values is a function of being a Test variable and all other characteristics. Bootstrap samples from Test alone do not identify its effect. Taking bootstrap samples from the entire sample and Control allows the identification of Test effect as well as Control effect. To illustrate the point, assume that we estimate the effects of Test, Control using the entire sample comprised by Test and Control *p*-values. The regression model with the design matrix is

$$\begin{bmatrix} \mathbf{y}_{all} \\ \mathbf{y}_{test} \\ \mathbf{y}_{control} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \beta_{int} \\ \beta_{test} \\ \beta_{control} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{all} \\ \boldsymbol{\varepsilon}_{test} \\ \boldsymbol{\varepsilon}_{control} \end{bmatrix}. \quad (3)$$

, where each $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are column vectors of size 100. **0** and **1** in the explanatory value matrix are a column vector of size 100. β_{int}, β_{test}, and $\beta_{control}$ are the intercept, the test effect, and the control effect parameter in the regression. As the design matrix is non-singular, we can recover the regression coefficient.

In addition to All *p*-values and Test and Control categories, we construct bootstrap samples for following categories: TAR, JAE, JAR, CAR, Financial, Managerial, Tax, Audit, AIS, Archival, Experimental, Rounded, Non-Rounded, Economic Significance, No Economic Significance, Practitioner/Regulator, and Academic. Categories other than Test/Control, Rounding/Non-Rounding are not-mutually exclusive. The actual design matrix and the regression model in Equation (3) expand to include the above categories.

Because the results in Table 1 suggest that *p*-hacking from non-rejection region to rejection region, there is a discontinuity in the *p*-value distribution that is most conspicuous at 0.05. Hence, when we resample to construct bootstrap sample, we separately resample *p*-values in the rejection and non-rejection regions. We then estimate false discovery rate using *p*-values in the rejection region and statistical power using *p*-values from beta distributions in the both regions for each bootstrap sample.

The false discovery estimate from each bootstrapped sample is measured with errors. Each estimate from a resampled *p*-values represents the average estimate of the resampled *p*-values. The precision of the estimate is inversely related to the variance of the estimates from a category. Thus, we apply precision weight for the regression model, estimating a weighted least square. The weighted least square mitigates potential bias from the imprecision of false discovery for small categories such as AIS. The precision weight for false discovery rate is not exactly equivalent to the appropriate precision weight for statistical power, yet the results are robust to alternative weight schemes. Any precision weight such as precision weights for statistical power, average precision weights for false discovery and statistical power seems appropriate as long as the weighting scheme mitigates the influence from imprecise estimates of AIS sample. The results are also robust to estimation excluding AIS category. The associations of the false discovery rate and the rounding choices are not robust without the weighting scheme. The rounding effect is robust, however, when we exclude *p*-values from CAR.

Additional References cited in the online information

Agresti, A. 2002. *Categorical Data Analysis*. 2nd edition. New York: Wiley-Interscience.

Benjamini, Y., and Y. Hechtlinger. 2014. Discussion: An estimate of the science-wise false discovery rate and applications to top medical journals by Jager and Leek. *Biostatistics* 15 (1): 13–16.

Gelman, A., and K. O'Rourke. 2014. Discussion: Difficulties in making inferences about scientific truth from distributions of published p-values. *Biostatistics* 15 (1): 18–23.

Goodman, S. N. 2014. Discussion: An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics* 15 (1): 23–27.