Supplement to "The p-value plot does not provide evidence against air pollution hazards"

Daniel J. Hicks

1 Supplemental methods

1.1 The three plots

Young and collaborators frequently cite two other graphical methods (Schweder and Spjøtvoll 1982; Simonsohn, Nelson, and Simmons 2014).

All three graphical methods take as input a set of N p-values $\mathbb{P} = \{p_1, p_2, \dots, p_N\}$. In Schweder and Spjøtvoll (1982) these are taken from separate tests of N different hypotheses. In Simonsohn, Nelson, and Simmons (2014) and the works by Young and collaborators, the p-values are (nominally) produced by applying a given statistical hypothesis test to N replications of a given study design, each replication drawing samples of size n from a given population. This corresponds to the simplest case of meta-analysis. Thus the p-values in $\mathbb{P}$ are nominally samples from a single underlying distribution $p_i \sim P$. Note that, if the real effect is zero $\delta = 0$, then P is the uniform distribution on $[0, 1]$.

1.2 Schweder and Spjøtvoll's p-value plot

Young and collaborators have frequently cited the "p-value plot" presented in Schweder and Spjøtvoll (1982). For this p-value plot, let $rank_{desc}(p_i)$ be the (1-indexed) *descending rank* of $p_i \in \mathbb{P}$, i.e., $rank_{desc}(p_i)$ is the number of p-values $p_j \in \mathbb{P}$ greater than or equal to p_i. The largest p-value has descending rank 1, and the smallest p-value has descending rank N. Then Schweder and Spjøtvoll's p-value plot plots the graph $(1 - p_i, rank_{desc}(p_i))$. See fig. 1.

Schweder and Spjøtvoll (1982) give a brief (and rather informal) argument that the relationship between $1 - p_i$ and $rank_{desc}(p_i)$ should be approximately linear "when [p_i] is not too small" and that, from left to right, "often, the plot will not show a clearcut break but rather a gradual bend" away from linearity. Their argument concludes that "the slope of that straight line is an estimate of … the number of true null hypotheses" in $\mathbb{P}$, and so the former can be used to estimate the latter (Schweder and Spjøtvoll 1982, 494). Schweder and Spjøtvoll's p-value plot is designed around different assumptions and to answer a different question than either of the other two plots. Also, Schweder and Spjøtvoll's p-value plot generally ignores "small" p-values, which roughly correspond to the statistically significant p-values. They illustrate their method with example data and a straight line "drawn by visual fit" rather than regression analysis.

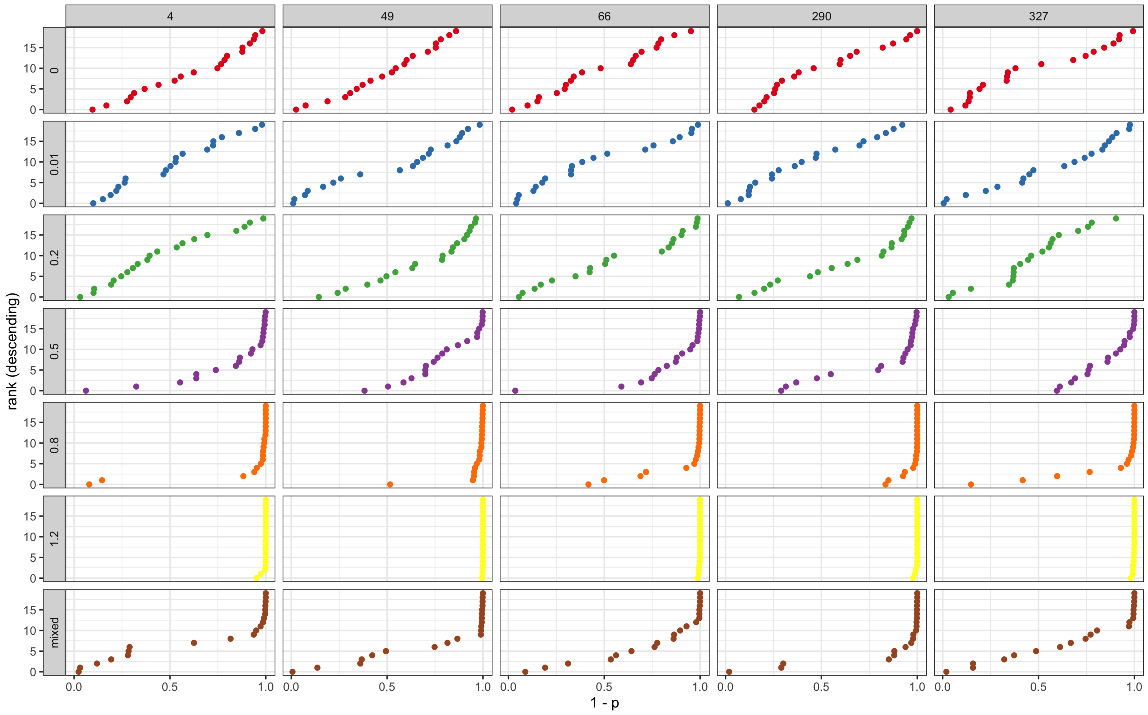

Figure 1: Examples of Schweder and Spjøtvoll's p-value plot, drawn at random from the simulation results (Schweder and Spjøtvoll 1982). Rows and colors correspond to conditions or real effects (δ), from zero (0) to moderate-strong (0.6) and a mixed condition $\delta = \{0.0, 0.6\}$. Columns correspond to indices for the simulation runs that produced these results, and are not meaningful. (In particular, there is no relationship between simulation run j in condition a and simulation run j in condition b.) In Schweder and Spjøtvoll's p-value plot, each point corresponds to a single p-value in the meta-analysis (simulation run); the y-axis is the p-value itself and the x-axis is the descending rank of the p-value.

1.3 Simonsohn, Nelson, and Simmons' p-curve

Young and collaborators have regularly echoed concerns about the replication crisis unfolding in social psychology and certain areas of biomedical research (for example, Young, Acharjee, and Das 2019, 50). In particular, they appeal to concerns about p-hacking (Simonsohn, Nelson, and Simmons 2014). Note that, other than Young's p-value plot, Young and collaborators have provided no specific[1] empirical[^counting] evidence of p-hacking in environmental epidemiology, and to my knowledge no such evidence has been published (Hicks 2021).

Young and collaborators have frequently associated Young's p-value plot with a method developed to detect p-hacking, called a "p-curve" (Simonsohn, Nelson, and Simmons 2014; for a comparison of several methods to detect p-hacking, see McShane, Böckenholt, and Hansen 2016). The intuition behind the p-curve is that p-hacking will tend to produce an excess number of p-values "just below" the conventional 0.05 threshold for statistical significance. Formally, Simonsohn et al.'s p-curve first divides the interval $[0, 0.05]$ into 5 bins at the thresholds $0.01, 0.02, 0.03, 0.04, 0.05$, then calculates N_b, the number of p-values in bin b. The p-curve is the graph (b_t, N_b), where b_t is the threshold for bin b. The method then formally tests for p-hacking by applying statistical tests of the null hypothesis that the restricted distribution $P|_{p<0.05}$ is uniform. See fig. 2.

Note that the p-curve is a histogram on the interval $[0, 0.5]$ with binwidth 0.01, and that it only includes statistically significant p-values. Thus the p-curve and Schweder and Spjøtvoll's p-value plot not only produce different kinds of plots, but actually direct their attention to different — typically disjoint — subsets of p-values. The two kinds of plots cannot be equivalent to each other. At no point have Young and collaborators acknowledged this fundamental difference between the two methods that they cite as support for their own method.

1.4 Young's p-value plot

For Young's p-value plot, let $rank_{asc}(p_i)$ be the (1-indexed) *ascending rank* of $p_i \in \mathbb{P}$, i.e., $rank_{asc}(p_i)$ is the number of p-values $p_j \in \mathbb{P}$ less than or equal to p_i. The smallest p-value has ascending rank 1, and the largest p-value has ascending rank N. Without loss of generality, if $\mathbb{P}$ is already in ascending order $p_1 < p_2 < \cdots < p_N$, then $rank_{asc}(p_i) = i$. And Young's p-value plot is the graph (i, p_i).

Statistically-minded readers might have already noted that Young's p-value plot is a rescaled QQ-plot of $\mathbb{P}$ against the uniform distribution, with the theoretical quantiles $q_i = \frac{i}{N} = \frac{rank_{asc}(p_i)}{N}$. It is not equivalent to the other two plots, and so cannot be validated by references to them. First, $rank_{desc}(p_i) = N - rank_{asc}(p_i) + 1$, and so for a fixed number

[1] Young and collaborators do cite Head et al. (2015), which used text mining methods to examine p-values reported in "all Open Access papers available in the PubMed database," classified at the journal level into 22 disciplines. Head et al. (2015) did not report the full size of their sample, but did include tens of thousands of p-values from "Medical and health sciences," which likely includes epidemiology but also other fields with very different methods, e.g., small-n animal model experiments and industry-funded clinical trials. While their statistical tests did find evidence of p-hacking in "Medical and health sciences," they did not consider subfields. Based on the methodological and funding diversity of the fields covered, and the lack of information about the distribution of subfields within their sample, we should be hesitant about drawing inferences to subfields.

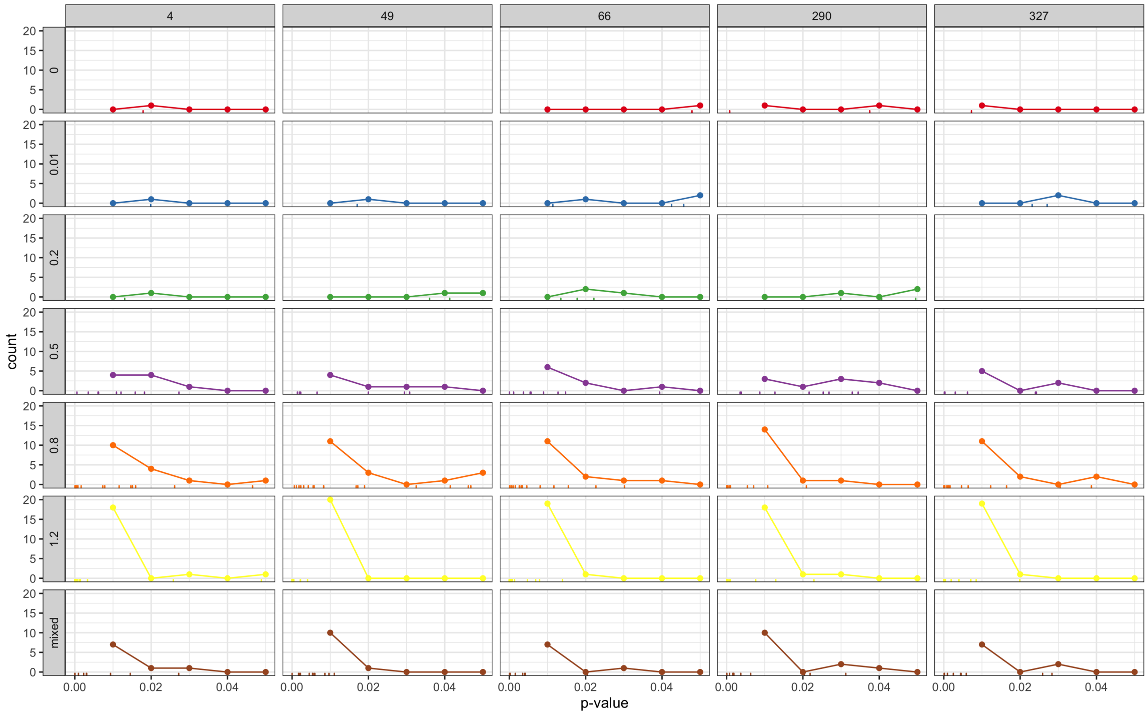

Figure 2: Examples of Simonsohn et al.'s p-curve, drawn at random from the simulation results (Simonsohn, Nelson, and Simmons 2014). Rows and colors correspond to conditions or real effects (δ), from zero (0) to moderate-strong (0.6) and a mixed condition $\delta = \{0.0, 0.6\}$. Columns correspond to indices for the simulation runs that produced these results, and are not meaningful. (In particular, there is no relationship between simulation run j in condition a and simulation run j in condition b.) Simonsohn et al.'s p-curve is restricted to p-values below the conventional 0.05 threshold; empty plots correspond to cases in which no p-values were below the threshold. These p-values are binned at the thresholds $0.01, 0.02, 0.03, 0.04, 0.05$, and each point corresponds to the number of p-values in the given bin. This plot is equivalent to a histogram.

of studies N Young's p-value does have a 1-1 mathematical relationship to Schweder and Spjøtvoll's p-value plot. In geometric terms, Young's p-value plot swaps the axes of Schweder and Spjøtvoll's p-value plot and reverses the direction of the ranking. However, a regression line fit to Schweder and Spjøtvoll's p-value plot will not deterministically correspond to a regression line fit to Young's p-value plot; see discussion in the supplemental results. In addition, the properties that Young and collaborators use in their analysis of their p-value plots do not correspond to the properties used by Schweder and Spjøtvoll (which, again, are justified with informal arguments rather than a formal analysis). So, even if Schweder and Spjøtvoll's p-value plot can be considered validated for certain purposes (it should be clear that I'm skeptical on this point), this does not validate Young's p-value plot as used by Young and collaborators. Second, Simonsohn, Nelson, and Simmons' p-curve constructs a histogram on a subset of p-values; this is a completely different construction from Young's p-value plot, and so citations to the former also do not validate the latter.

1.5 Likelihood

The likelihood conception of evidence is not strongly associated with any one statistician or philosopher of science, though it can be associated with one approach to Bayesian statistics (Kass and Raftery 1995; Romeijn 2017).

Formally, the likelihood conception of evidence compares two rival hypotheses H_1 and H_2 using some data d. The likelihood ratio is defined as

$$K(H_1, H_2; d) = \frac{L(H_1; d)}{L(H_2; d)} = \frac{pr(d|H_1)}{pr(d|H_2)}.$$

If $K > 1$, then the evidence favors H_1; and $K < 1$ then the evidence favors H_2. Sometimes $\log K$ is used to create symmetry between H_1 and H_2. On one common interpretive scale, $\left|\log_{10} K\right| < 0.5$ is "not worth more than a bare mention," not supporting either hypothesis; $0.5 < \left|\log_{10} K\right| < 1$ is "substantial"; $1 < \left|\log_{10} K\right| < 2$ is "strong"; and $2 < \left|\log_{10} K\right|$ is "decisive" (Kass and Raftery 1995).

To apply the likelihood conception of evidence to Young and collaborators' skeptical claims about air pollution, H_1 will be the zero or mixture hypothesis, the rival hypothesis H_2 will be the hypotheses (a-h), and the data d will be the analysis outputs (i-iv). (For simplicity, the same dichotomous frequentist test outputs are used, e.g., statistically significant or not, rather than continuous-valued likelihoodist or Bayesian alternatives.) In each case, insofar as $K < 0.5$, this implies that the p-value plot does not provide evidence to support the zero or mixture hypotheses.

2 Supplemental results

2.1 Gaps

Fig. 3 shows the distribution of sizes of the largest gap across effect sizes.

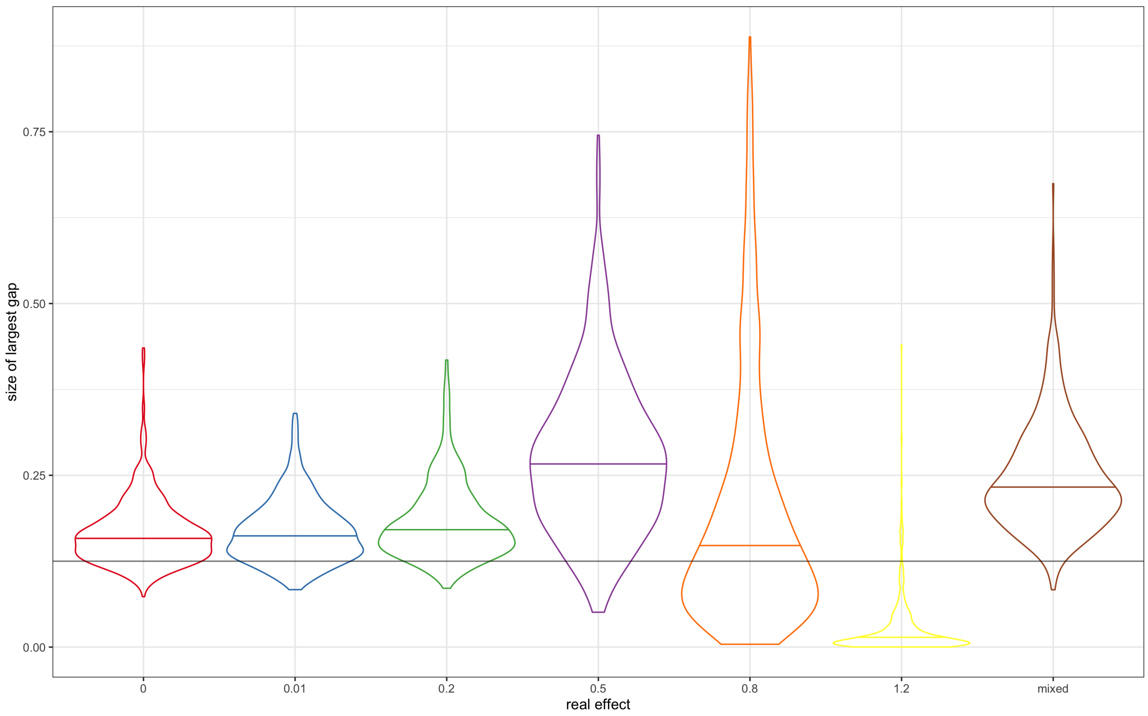

Figure 3: **Distribution of gaps**: Each violin plot shows the distribution of sizes of the largest gap for each real effect size. The horizontal line indicates the threshold .125 for a plot to be considered “gappy.” Except for the very large effect size, most plots across almost all effect sizes have gaps.

2.2 Slopes

Fig. 4 and fig. 5 show the distribution of slopes of regression lines fit to the QQ-plot, Schweder and Spjøtvoll's p-value plot, and Young's p-value plot across effect sizes. fig. 5 uses the same data as the left panel of fig. 4, with a wider aspect ratio for readability and dashed lines indicating the thresholds for a slope of "approximately 1."

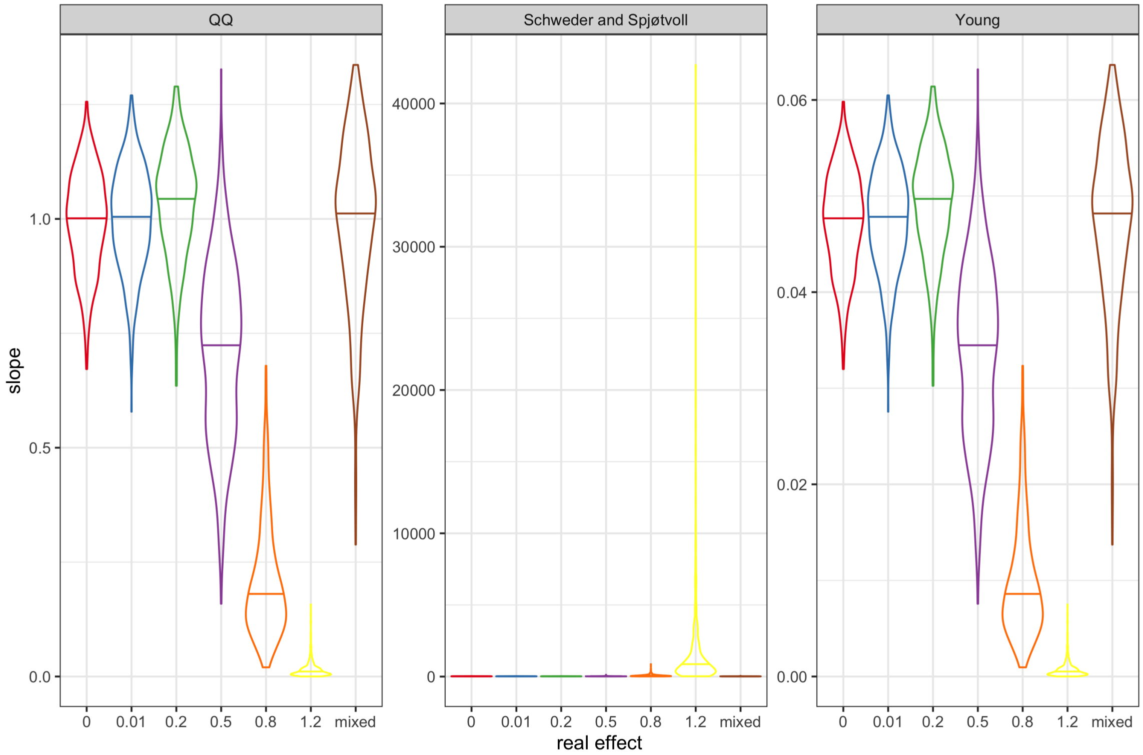

Figure 4: **Distribution of slopes**: Each violin plot shows the distribution of slopes for the linear regression fit to each plot across each real effect size. Note that the slopes for Young's p-value plot are rescaled from the QQ-plot.

Fig. 6 shows the relationship between the slopes of Young's and Schweder and Spjøtvoll's p-value plots. For an ordinary least-squares regression of y against x, the slope of the fitted regression line is rs_x/s_y, where r is the estimated correlation coefficient between x and y and the s are the estimated standard deviations. So the slope of x and y (swapping the axes) is rs_y/s_x. The ratio of the first slope to the reciprocal of the second slope is r^2. This might suggest a deterministic relationship between the slopes of Schweder and Spjøtvoll's and Young's p-value plots. However, the estimated correlation coefficient r is calculated from observations drawn from the random variables X and Y, and so r itself is an observation drawn from a random variable. That is, the observed value of r will vary between different iterations of the study. So the relationship between the slopes of the two p-value plots is noisy. See fig. 4.

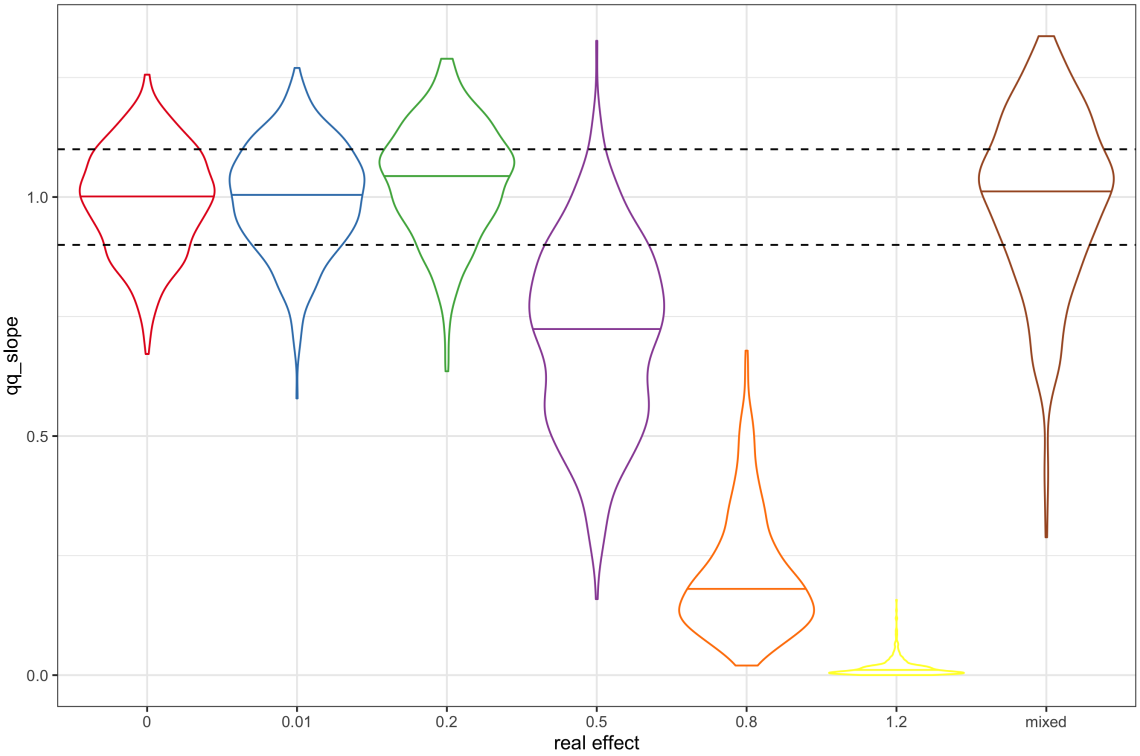

Figure 5: **Slopes for the QQ-plot**: Data are the same as in the left panel of fig. 4. These slopes are used for the slope analysis that supposedly gives evidence of zero effects. The dashed lines indicate the $1 \pm .1$ threshold used for "approximately 1."

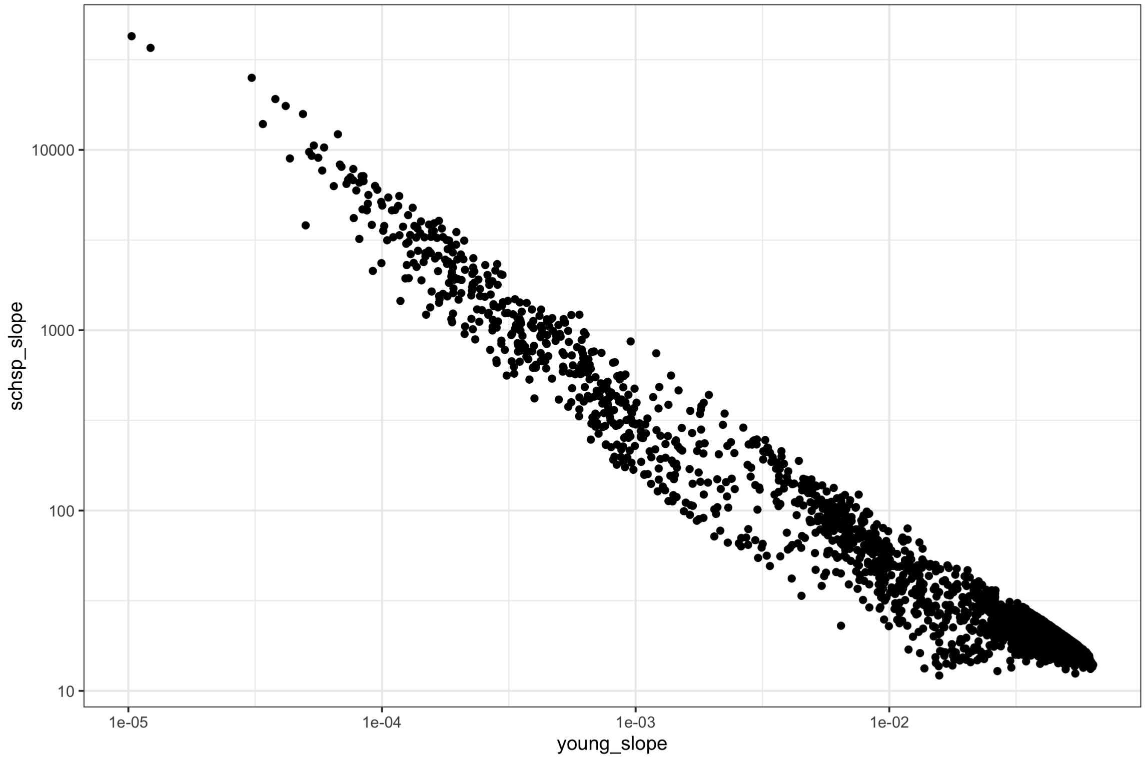

Figure 6: **Relationship between slopes of Young's and Schweder and Spjøtvoll's p-value plots**: Scatterplot of the slopes for the two p-value plots, log-log scale. The two plots are in a 1-1 relationship with each other, by reversing an axis and swapping the x and y axes. But the slopes of the regression lines fit to each plot are not in a 1-1 relationship.

2.3 Linearity

Fig. 7 shows the distribution of area under the curve (AUC) of the QQ-plot across each real effect size. AUC has its maximum (expected) value of 0.5 when the real effect size $\delta = 0$, and decreases as the curve of the QQ-plot bends away from linearity. So AUC can be used as a continuous measure of linearity.

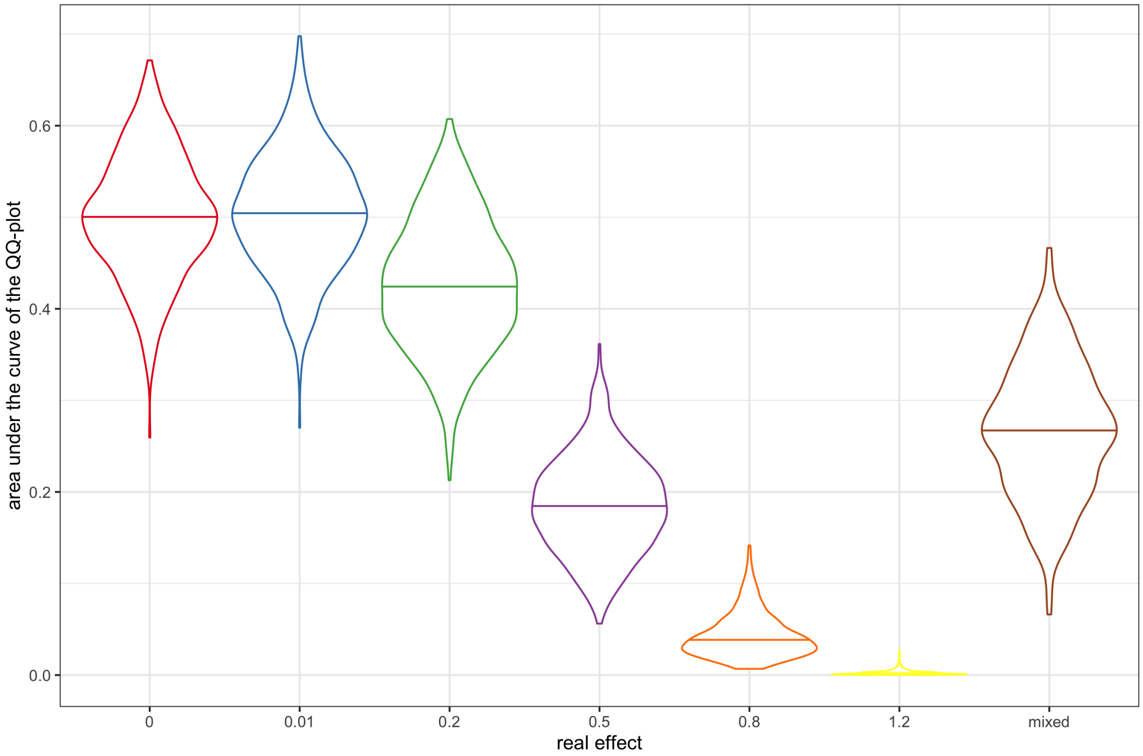

Figure 7: **Area under the curve (AUC) of the QQ-plot**: Each violin plot shows the distribution of area under the curve (AUC) across each real effect size. When the real effect size is $\delta = 0$, distribution of p-values is uniform and the expected QQ-plot is a straight line $y = x$. The area under this perfectly linear QQ-plot is .5, as indicated by the median in the corresponding violin plot. The AUC distribution for the very small effect is almost identical to the zero effect, and expected AUC decreases as real effect size increases. The distribution for the mixed or heterogenous effect overlaps with the distributions for almost every other effect size, except for very large effects.

2.4 Severity analysis

Table 1: Severity analysis results. H_0 indicates the null or rival hypothesis playing the role of $\neg H$. "False" and "true" indicate the number of simulation runs in which the test output is false and true, respectively, and p is calculated as the fraction of true runs.

H_0	output	p	false	true
$\delta = 0$	ii-gap	0.84	79	421
$\delta = 0$	iii-range	0.61	194	306
$\delta = 0$	iii-T	0.55	226	274
$\delta = 0$	iii-TOST	0.21	394	106
$\delta = 0$	iii-KS	1.00	1	499
$\delta = 0$	iv-AIC	0.68	158	342
$\delta = 0$	iv-F	0.54	232	268
$\delta = 0.01$	ii-gap	0.85	75	425
$\delta = 0.01$	iii-range	0.61	193	307
$\delta = 0.01$	iii-T	0.54	230	270
$\delta = 0.01$	iii-TOST	0.17	416	84
$\delta = 0.01$	iii-KS	1.00	0	500
$\delta = 0.01$	iv-AIC	0.71	144	356
$\delta = 0.01$	iv-F	0.55	225	275
$\delta = 0.2$	ii-gap	0.91	43	457
$\delta = 0.2$	iii-range	0.56	218	282
$\delta = 0.2$	iii-T	0.58	209	291
$\delta = 0.2$	iii-TOST	0.13	435	65
$\delta = 0.2$	iii-KS	0.98	9	491
$\delta = 0.2$	iv-AIC	0.78	109	391
$\delta = 0.2$	iv-F	0.67	164	336
$\delta = 0.5$	ii-gap	0.93	35	465
$\delta = 0.5$	iii-range	0.16	419	81
$\delta = 0.5$	iii-T	0.42	291	209
$\delta = 0.5$	iii-TOST	0.00	500	0
$\delta = 0.5$	iii-KS	0.05	474	26
$\delta = 0.5$	iv-AIC	1.00	0	500
$\delta = 0.5$	iv-F	1.00	0	500
$\delta = 0.8$	ii-gap	0.54	230	270
$\delta = 0.8$	iii-range	0.00	500	0
$\delta = 0.8$	iii-T	0.00	498	2
$\delta = 0.8$	iii-TOST	0.00	500	0
$\delta = 0.8$	iii-KS	0.00	500	0
$\delta = 0.8$	iv-AIC	1.00	0	500
$\delta = 0.8$	iv-F	1.00	0	500
$\delta = 1.2$	ii-gap	0.04	480	20
$\delta = 1.2$	iii-range	0.00	500	0

δ = 1.2	iii-T	0.00	500	0
δ = 1.2	iii-TOST	0.00	500	0
δ = 1.2	iii-KS	0.00	500	0
δ = 1.2	iv-AIC	1.00	0	500
δ = 1.2	iv-F	1.00	0	500
δ > 0	ii-gap	0.65	863	1637
δ > 0	iii-range	0.27	1830	670
δ > 0	iii-T	0.31	1728	772
δ > 0	iii-TOST	0.06	2351	149
δ > 0	iii-KS	0.41	1483	1017
δ > 0	iv-AIC	0.90	253	2247
δ > 0	iv-F	0.84	389	2111
δ is mixed	ii-gap	0.99	7	493
δ is mixed	iii-range	0.45	276	224
δ is mixed	iii-T	0.75	125	375
δ is mixed	iii-TOST	0.00	500	0
δ is mixed	iii-KS	0.32	338	162
δ is mixed	iv-AIC	0.98	8	492
δ is mixed	iv-F	0.96	18	482
δ is non-zero	ii-gap	0.71	870	2130
δ is non-zero	iii-range	0.30	2106	894
δ is non-zero	iii-T	0.38	1853	1147
δ is non-zero	iii-TOST	0.05	2851	149
δ is non-zero	iii-KS	0.39	1821	1179
δ is non-zero	iv-AIC	0.91	261	2739
δ is non-zero	iv-F	0.86	407	2593
δ is not mixed	ii-gap	0.69	942	2058
δ is not mixed	iii-range	0.33	2024	976
δ is not mixed	iii-T	0.35	1954	1046
δ is not mixed	iii-TOST	0.09	2745	255
δ is not mixed	iii-KS	0.51	1484	1516
δ is not mixed	iv-AIC	0.86	411	2589
δ is not mixed	iv-F	0.79	621	2379

2.5 Likelihood analysis

Fig. 8 and fig. 9 show the results of the likelihood analysis; see the supplemental materials for a tables and interactive versions of these results. Log likelihood ratios are reported, so results above 0.5 support H_1 and results below -0.5 support H_2. So, by the weak severity criterion, when the results of the likelihood analysis are < 0.5, the test output does not provide evidence supporting the target hypothesis of zero or mixed effect.

For "gaps" in the plot, calculating the likelihood ratio would require simulation conditions that included p-hacking and other questionable research practices. Because the simulation

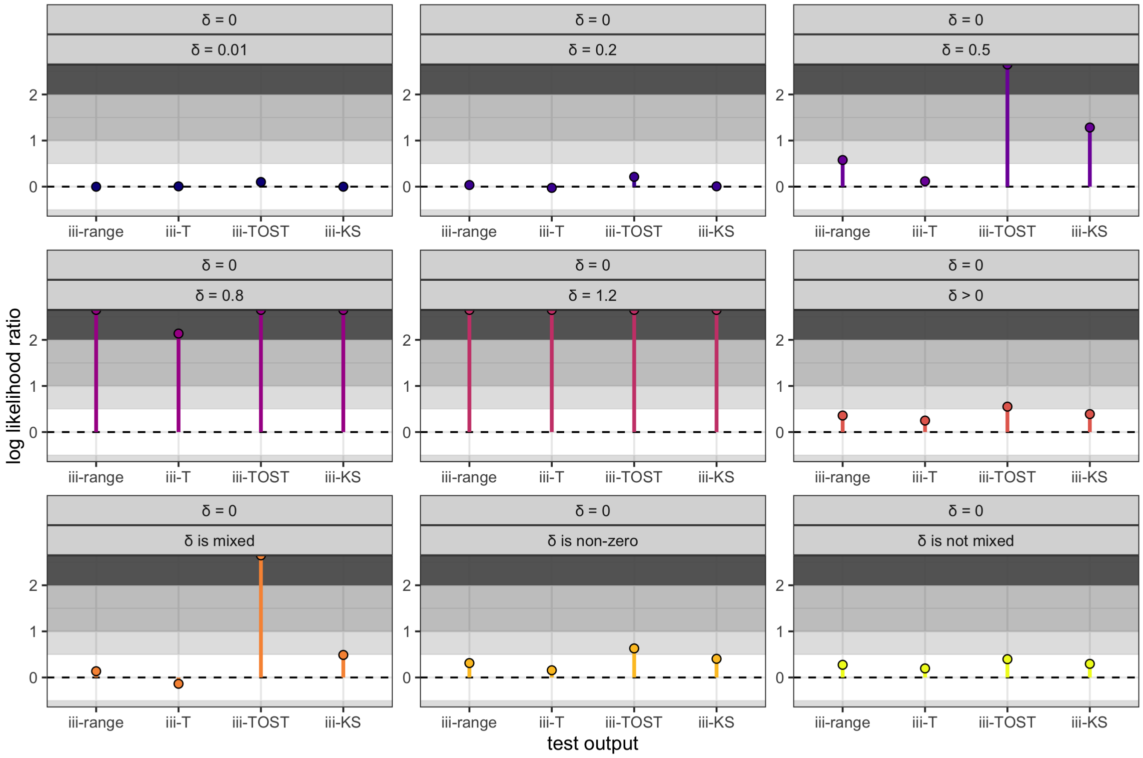

Figure 8: **Results of the likelihood analysis for $H_1 : \delta = 0$.** Each point gives the log likelihood ratio for H_1 vs. rival hypothesis, given an output. Each panel represents one comparison of H_1 against a rival hypothesis H_2. Position on the y-axis indicates the strength of the evidence that the output provides to the hypotheses: *greater values indicate more support for H_1 over H_2*. (Points at the plot margins have infinite value due to division by zero.) Shaded regions indicate the degree of support for one hypothesis against the other, in order from lightest to darkest: none, “substantial,” “strong,” “decisive.” An interactive version of this plot is included in the automatic reproduction of the analysis for this paper.

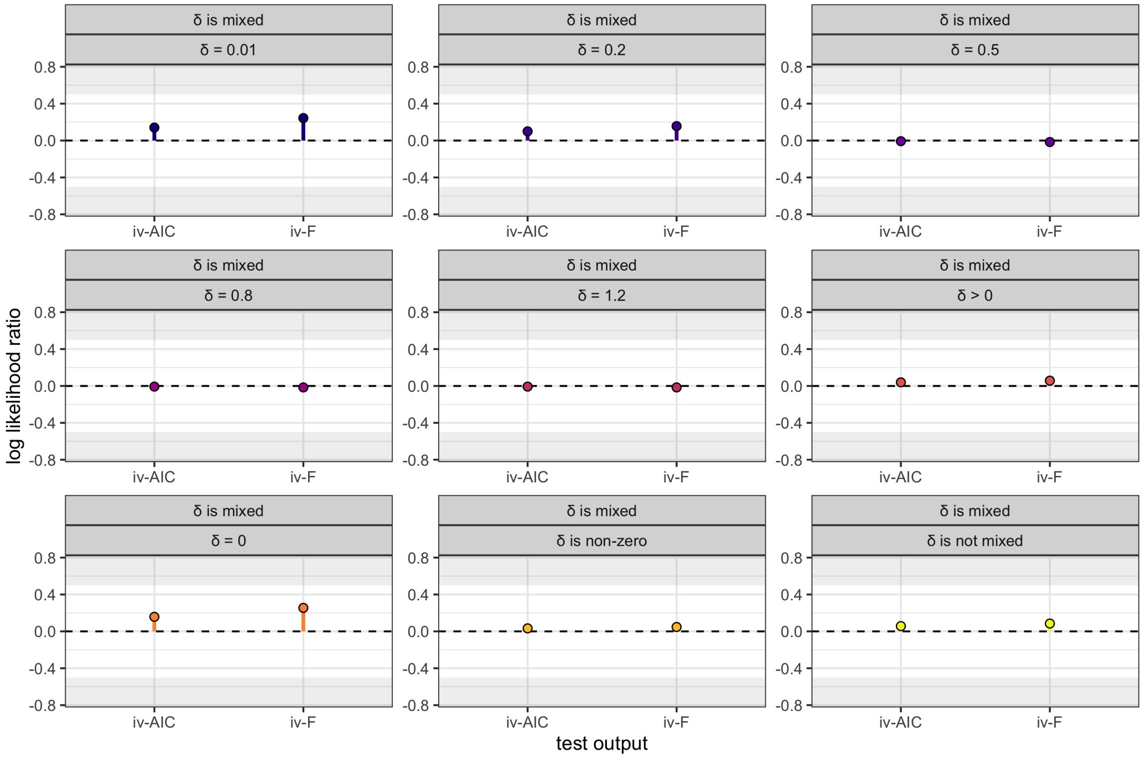

Figure 9: **Results of the likelihood analysis for $H_1 : \delta$ is mixed.** Interpretation is the same as fig. 8.

does not currently support these kinds of conditions, likelihood analysis cannot be used for this output.

The zero effect hypothesis is supposedly supported by a slope of approximately 1. All four methods provide "decisive" support for a zero effect against strong and very strong effects, and all except the T-test provide "substantial" or better support against moderate effects. The TOST approach provides stronger or equally strong evidence, compared to the other approaches, across all of the rival hypotheses. When the rival hypothesis includes small or very small effects, other approaches either do not provide evidence to support zero effects. So, as with the severity analysis, *whether and to what degree the "45-degree line" might provide evidence for zero effects depends on the choice of rival hypothesis and analytical approach used.* In addition, and again as in the severity analysis, the visual assessment apparently used by Young and collaborators will provide weaker evidence than the range test, which does not provide evidence against rivals that include small effects. *So, against rival hypotheses that include small effects, insofar as Young and collaborators are relying on visual judgment, the "45-degree line" does not provide evidence of a zero effect.*

For the mixed effect hypothesis, all of the points for both AIC and the F-test are in the "not worth mentioning" or no evidence range, and so *neither method provides evidence to support heterogeneity.* This is the same conclusion reached by the severity analysis.

Table 2: Likelihood analysis results. "llr" is the log likehood ratio. Values > 0.5 indicate support for H_1.

H_1	H_2	output	llr	$L(H_1)$	$L(H_2)$
δ = 0	δ = 0.01	iii-range	0.00	0.61	0.61
δ = 0	δ = 0.01	iii-T	0.01	0.55	0.54
δ = 0	δ = 0.01	iii-TOST	0.10	0.21	0.17
δ = 0	δ = 0.01	iii-KS	0.00	1.00	1.00
δ = 0	δ = 0.2	iii-range	0.04	0.61	0.56
δ = 0	δ = 0.2	iii-T	-0.03	0.55	0.58
δ = 0	δ = 0.2	iii-TOST	0.21	0.21	0.13
δ = 0	δ = 0.2	iii-KS	0.01	1.00	0.98
δ = 0	δ = 0.5	iii-range	0.58	0.61	0.16
δ = 0	δ = 0.5	iii-T	0.12	0.55	0.42
δ = 0	δ = 0.5	iii-TOST	Inf	0.21	0.00
δ = 0	δ = 0.5	iii-KS	1.28	1.00	0.05
δ = 0	δ = 0.8	iii-range	Inf	0.61	0.00
δ = 0	δ = 0.8	iii-T	2.14	0.55	0.00
δ = 0	δ = 0.8	iii-TOST	Inf	0.21	0.00
δ = 0	δ = 0.8	iii-KS	Inf	1.00	0.00
δ = 0	δ = 1.2	iii-range	Inf	0.61	0.00
δ = 0	δ = 1.2	iii-T	Inf	0.55	0.00
δ = 0	δ = 1.2	iii-TOST	Inf	0.21	0.00
δ = 0	δ = 1.2	iii-KS	Inf	1.00	0.00

$\delta = 0$	$\delta > 0$	iii-range	0.36	0.61	0.27
$\delta = 0$	$\delta > 0$	iii-T	0.25	0.55	0.31
$\delta = 0$	$\delta > 0$	iii-TOST	0.55	0.21	0.06
$\delta = 0$	$\delta > 0$	iii-KS	0.39	1.00	0.41
$\delta = 0$	δ is mixed	iii-range	0.14	0.61	0.45
$\delta = 0$	δ is mixed	iii-T	-0.14	0.55	0.75
$\delta = 0$	δ is mixed	iii-TOST	Inf	0.21	0.00
$\delta = 0$	δ is mixed	iii-KS	0.49	1.00	0.32
$\delta = 0$	δ is non-zero	iii-range	0.31	0.61	0.30
$\delta = 0$	δ is non-zero	iii-T	0.16	0.55	0.38
$\delta = 0$	δ is non-zero	iii-TOST	0.63	0.21	0.05
$\delta = 0$	δ is non-zero	iii-KS	0.40	1.00	0.39
$\delta = 0$	δ is not mixed	iii-range	0.27	0.61	0.33
$\delta = 0$	δ is not mixed	iii-T	0.20	0.55	0.35
$\delta = 0$	δ is not mixed	iii-TOST	0.40	0.21	0.09
$\delta = 0$	δ is not mixed	iii-KS	0.30	1.00	0.51
δ is mixed	$\delta = 0$	iv-AIC	0.16	0.98	0.68
δ is mixed	$\delta = 0$	iv-F	0.25	0.96	0.54
δ is mixed	$\delta = 0.01$	iv-AIC	0.14	0.98	0.71
δ is mixed	$\delta = 0.01$	iv-F	0.24	0.96	0.55
δ is mixed	$\delta = 0.2$	iv-AIC	0.10	0.98	0.78
δ is mixed	$\delta = 0.2$	iv-F	0.16	0.96	0.67
δ is mixed	$\delta = 0.5$	iv-AIC	-0.01	0.98	1.00
δ is mixed	$\delta = 0.5$	iv-F	-0.02	0.96	1.00
δ is mixed	$\delta = 0.8$	iv-AIC	-0.01	0.98	1.00
δ is mixed	$\delta = 0.8$	iv-F	-0.02	0.96	1.00
δ is mixed	$\delta = 1.2$	iv-AIC	-0.01	0.98	1.00
δ is mixed	$\delta = 1.2$	iv-F	-0.02	0.96	1.00
δ is mixed	$\delta > 0$	iv-AIC	0.04	0.98	0.90
δ is mixed	$\delta > 0$	iv-F	0.06	0.96	0.84
δ is mixed	δ is non-zero	iv-AIC	0.03	0.98	0.91
δ is mixed	δ is non-zero	iv-F	0.05	0.96	0.86
δ is mixed	δ is not mixed	iv-AIC	0.06	0.98	0.86
δ is mixed	δ is not mixed	iv-F	0.08	0.96	0.79

3 References

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PLoS Biol* 13 (3): e1002106. https://doi.org/10.1371/journal.pbio.1002106.

Hicks, Daniel J. 2021. "Open Science, the Replication Crisis, and Environmental Public Health." *Accountability in Research* 0 (July): null. https://doi.org/10.1080/08989621.

2021.1962713.

Kass, Robert E., and Adrian E. Raftery. 1995. “Bayes Factors.” *Journal of the American Statistical Association* 90 (430): 773–95. https://doi.org/10.2307/2291091.

McShane, Blakeley B., Ulf Böckenholt, and Karsten T. Hansen. 2016. “Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes.” *Perspectives on Psychological Science* 11 (5): 730–49. https://doi.org/10.1177/1745691616662243.

Romeijn, Jan-Willem. 2017. “Philosophy of Statistics.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/statistics/.

Schweder, T., and E. Spjøtvoll. 1982. “Plots of P-values to Evaluate Many Tests Simultaneously.” *Biometrika* 69 (3): 493–502. https://doi.org/10.1093/biomet/69.3.493.

Simonsohn, Uri, Leif D. Nelson, and Joseph P. Simmons. 2014. “P-Curve: A Key to the File-Drawer.” *Journal of Experimental Psychology: General* 143 (2): 534–47. https://doi.org/10.1037/a0033242.

Young, S. Stanley, Mithun Kumar Acharjee, and Kumer Das. 2019. “The Reliability of an Environmental Epidemiology Meta-Analysis, a Case Study.” *Regulatory Toxicology and Pharmacology* 102 (March): 47–52. https://doi.org/10.1016/j.yrtph.2018.12.013.