Chapter 6
Data Detective Methods for Revealing Questionable Research Practices

Gregory Francis and Evelina Thunell

Abstract There are many types of Questionable Research Practices (QRPs) that all tend to generate statistical information that misrepresents reality. This chapter discusses some methods for detecting the presence of QRPs, mostly by looking for conflicts in different sources of information. These methods typically cannot identify precisely which QRPs were used, and sometimes the conflicts are due to typos or simple mistakes, but either way readers should be skeptical about the validity of studies with inconsistent statistical information. An appropriate mindset for identifying inconsistencies is that of a "data detective" who looks for patterns that do not make sense. We start by describing mathematical inconsistencies between sample sizes and the degrees of freedom in hypothesis tests, which are easy to detect and indicate either a QRP, unreported outlier removal, or sloppiness in reporting. A similarly easy check is the use of the STATCHECK program to identify inconsistencies between reported test statistics and *p*-values, which may indicate sloppiness in reporting or improper rounding to conclude statistical significance. Similar problems can also be discovered with the GRIM test, which identifies situations where reported means or proportions are impossible for the given measurement and sample size(s). Two additional tests explore inconsistencies across experiments. First, the Test for Excess Success compares the frequency of reported successful outcomes to the expected frequency if the tests were run properly, fully reported, and analyzed without QRPs. Too much success indicates a problem with the reported results (possibly because of QRPs). Second, the *p*-curve analysis examines the distribution of reported *p*-values for properties that indicate invalid data sets (that are perhaps the result of QRPs).

G. Francis (✉)
Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA
e-mail: gfrancis@purdue.edu

E. Thunell
Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA

Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden

W. O'Donohue et al. (eds.), *Avoiding Questionable Research Practices in Applied Psychology*, https://doi.org/10.1007/978-3-031-04968-2_6

Keywords Questionable research practice · Clinical psychology · Excess success · Data detective methods for revealing questionable research practices · STATCHECK program · GRIM test

Introduction

As discussed in other chapters, questionable research practices (QRPs) and *p*-hacking can turn non-conclusive data sets into seemingly interesting findings. While such practices might be tempting for a researcher who is desperate to publish their work in fancy journals, they come at the expense of the credibility and reproducibility of the findings. Examples of QRPs are publication bias (reporting significant findings but not reporting relevant non-significant findings), inappropriate sampling (e.g., adding data points until achieving statistical significance), inappropriate analyses (e.g., trying various analyses and reporting only the ones that give the wanted result), and hypothesizing after the results are known (HARKing; inventing a new theory and hypothesis that matches your results). Hypothesis testing is the dominant statistical analysis method in clinical psychology, and it comes with strict requirements and rules that are violated in different ways by QRPs. The impact of using QRPs is a kind of bias that misrepresents reality.

QRPs can make studies appear to provide strong support for effects that do not exist in reality. That is, the results seem to support the alternative hypothesis, but the null hypothesis is actually true. How then can we distinguish scientific results that are valid from results that are based on QRPs? Luckily, QRPs tend to leave a pattern of statistical evidence that can be used to identify their presence. In this chapter, we show how to detect and interpret such patterns.

In many respects, revealing the patterns generated by QRPs is similar to a detective trying to crack a case. The information may not be right in front of you, but different clues can be combined to demonstrate problems with experimental results and conclusions that are based on QRPs. In this chapter, we describe a number of methods that help you act like a data detective and identify problems in reported statistics.

Mathematical Inconsistencies and Data Gleaning

A valuable skill for a data detective is recognizing how to extract relevant information from what the authors themselves report. Here, we review some approaches that have proven useful for identifying problems with reported results.

A simple approach for detecting errors in reported results is to look for numerical inconsistencies. For example, many statistical tests (e.g., *t* and *F* tests) are based on distributions with a "degrees of freedom" (df) value. For example, a one-sample *t*-test has df = $n - 1$, where n is the sample size, while a two-sample *t*-test has

$df = n_1 + n_2 - 2$, where n_1 and n_2 are the sizes of the two samples. Likewise, an independent one-way ANOVA F-test has two degrees of freedom terms called $df_{numerator} = K - 1$ and $df_{denominator} = N - K$. Here, K is the number of conditions and N is the sum of sample sizes across all conditions. Scientific papers usually report the sample sizes and the number of conditions, so it is relatively easy to calculate the degrees of freedom. Thus, you can easily check the following text: "As predicted, with $n_1 = 35$ and $n_2 = 27$, we found a significant difference between the control and experimental means $t(58) = 2.1$, $p = 0.04$." The authors report 58 degrees of freedom, but using the formula above for the two-sample t-test you know that the degrees of freedom should actually be $n_1 + n_2 - 2 = 60$. An inconsistency of this type might indicate that the authors removed some participants from their data set without reporting this, but still properly reported the degrees of freedom for the remaining data. Outlier removal is not necessarily a QRP, but sometimes participants are removed because their absence allows the remaining data to show a significant ($p < 0.05$) result. At any rate, data removal should be fully reported and justified. Errors of this type are rather common. At best they indicate sloppiness, and regardless of their source should prompt you to feel less confident in the reported results and their associated conclusions. The next section describes a conceptually similar check for inconsistencies that often have more severe consequences.

STATCHECK

Most statistical analyses in psychology use hypothesis testing to determine whether there is an "effect." Typically, this is done by defining an "alternative hypothesis" that there is a true effect and a null hypothesis that indicates "no effect." For example, when testing whether a drug is effective at reducing the duration of a cold, the null hypothesis H_0 might look like:

$$H_0 : \mu_1 = \mu_2$$

where μ_1 and μ_2 denote the duration of the cold with and without the drug, respectively. Thus, the null hypothesis states that there is no difference in the population mean durations with or without the drug whereas the alternative hypothesis states that the drug does change the duration. The goal of the hypothesis test is to decide whether to reject the null hypothesis. This decision is based on "statistical significance," which is determined by a test statistic that is derived from the experimental data. A two-sample t test for independent equal means has a test statistic of:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means and $s_{\bar{X}_1-\bar{X}_2}$ is the standard deviation of the sampling distribution of the difference of means, which is a function of the standard deviation of each sample, s, and the sample sizes n_1 and n_2. If the null hypothesis is true, the t-value is usually close to 0. The hypothesis test will then not reject the null hypothesis. If the alternative hypothesis is true and the sample sizes are large enough, the t-value will typically deviate substantially from 0. In this case, the researcher rejects the null hypothesis and can argue for their alternative hypothesis. However, just due to random sampling, the t-value will sometimes deviate from 0 even if the null hypothesis is true, and the researcher will then erroneously reject the null hypothesis. How often this so-called Type I error happens is controlled by the researcher through a significance criterion, α.

Oftentimes, the criterion is set to $\alpha = 0.05$, meaning that the probability of concluding that an effect exists when it truly does not is 5%. The decision about whether to reject the null hypothesis and thus conclude that an effect exists (concluding statistical significance) is based on the p-value (the area under the tail, beyond the observed t-value, of the t sampling distribution if the null hypothesis is true). If $p < \alpha$, then the observed t-value deviates more from 0 than what should be common if the null hypothesis is true. Therefore, the researchers conclude that there seems to be an effect: they reject the null hypothesis and claim that the observed difference of means is "statistically significant."

When reporting the results of a hypothesis test it is common to report the computed t-value, the corresponding degrees of freedom (which depends on the sample size(s)), and the p-value. It often looks like: $t(48) = 2.55$, $p = 0.014$. It is actually redundant to report both the t- and p-values, as there is a one-to-one relationship between them for a given degrees of freedom. This redundancy can be used to check the reported statistics.

For example, suppose you read an article that reports: "As predicted we found a significant difference between the control and experimental conditions, $t(22) = 2.00$, $p < 0.05$." For the given degrees of freedom (df = 22) and t-value, one can recompute the corresponding p-value[1] to discover that $p = 0.058$. Thus, the reported t-value is incompatible with the statement $p < 0.05$. Instead, the result is actually not statistically significant (because $p > \alpha = 0.05$). The mathematics in the original text, therefore, indicates that something is wrong with the numbers. p-value inconsistencies can come about from simple typos (e.g., typing 0.014 instead of 0.14), or honest mistakes (e.g., copying the wrong line from the output of statistical software). In some cases (as in the above example), p-value inconsistencies might be because authors "round down" a reported p-value in order to make readers believe an experiment produced statistical significance. This kind of inappropriate rounding is a QRP. Regardless of how they appear, p-value inconsistencies should raise concerns about the reported results and their associated conclusions.

[1] For example, with the online calculator at https://introstatsonline.com/chapters/calculators/t_dist.shtml

STATCHECK is an online program (http://statcheck.io) that automates this kind of consistency check. To use it, simply upload a copy of an article and let STATCHECK scan it for statistical information. Just as we did above, STATCHECK identifies test statistics and their accompanying degrees of freedom, recomputes the *p*-value based on these numbers, and then compares it to the reported *p*-values. STATCHECK includes some additional computations (such as checking on whether the recomputed *p*-value is close enough to the reported *p*-value for appropriate rounding to be an explanation, and identifying whether an inconsistency in *p*-values changes the decision on statistical significance). STATCHECK works for a variety of statistical tests.

Some limitations of STATCHECK include that it cannot process certain file formats, it typically does not distinguish between one- and two-sided tests, and it cannot parse non-standard formats for reporting statistical outcomes. These limitations cause STATCHECK to sometimes omit or misinterpret statistical test results, and it is therefore always advisable to manually check the statistics flagged by STATCHECK.

Errors of this type are shockingly common. Systematic investigations of scientific articles have found that around half of them have at least one inconsistent *p*-value and that around 12–14% of the articles contain an inconsistency that alters the interpretation of statistical significance.

GRIM Tests

Another way of identifying inconsistencies in statistical reporting is to notice a relationship between sample sizes and measured values. Let's consider a simple case. Suppose you receive a marketing report for a survey to evaluate how many people might be interested in a new product (a macaroni-and-cheese pizza) at your restaurant. One of your employees runs a survey on $n = 37$ people and reports that 56% of the people expressed interest in the new product. Your first reaction might be that the survey seems pretty promising for your new product. A bit of data detective work, however, suggests that you should assign the survey task to a different employee. The percentage calculation is computed from the following formula

$$\%\text{Interest} = \frac{f}{n} \times 100$$

where f is the number of survey respondents who are interested in your product and $n = 37$ is the number of people who participated in the survey. Let's deduce the value for f by plugging in the values reported by your employee

$$56 = \frac{f}{37} \times 100$$

With a bit of algebra, we find that $f = 20.72$. We know this value for f cannot be quite right because there cannot be fractions of respondents. Could the reported percentage have been rounded from the true value? We can check by looking at nearby values of f. For example, if $f = 21$ then we would get

$$\%\text{Interest} = \frac{21}{37} \times 100 = 56.76$$

Unfortunately, this value does not explain why your employee reported 56% because rounding of 56.76% would produce 57%. What if $f = 20$? Then we get

$$\%\text{Interest} = \frac{20}{37} \times 100 = 54.05,$$

which is too small to be rounded up to 56%. In fact, with $n = 37$ people in the survey it is impossible for the percentage to equal 56%, even after rounding. So, either your employee misreported the number of people in the survey or simply made up the numbers. At any rate, you should hold off on making changes to your menu until you resolve the inconsistency.

Similar logic applies to reported values of means. For example, suppose a survey asks people to rate, on an integer scale from 1 to 7, how much interest they have in a macaroni-and-cheese pizza. A rating of 1 indicates no interest at all and a rating of 7 indicates that they want it *now!* The employee responsible for the survey reports that 55 responders gave a mean value of 4.74, which indicates interest above the middle point of the scale. The computation of the mean, $\bar{X}$, is based on the following formula:

$$\bar{X} = \frac{\sum X_i}{n},$$

where X_i refers to the score for responder number i, and the capital sigma indicates to sum the scores of all the responders. Thus, with the reported mean and sample size, we can solve for the sum of scores:

$$\sum X_i = (n)\bar{X} = (55)4.74 = 260.7$$

Importantly, the scores can only take integer values (1, 2, 3, 4, 5, 6, or 7) because that is the nature of the rating scale. This means that the sum of scores must also be an integer value, which it is not in the above calculation. Did we get a decimal value for the sum because the reported mean was rounded from its true value? We can check this possibility by considering nearby values for the sum of scores and seeing if the corresponding mean value matches what was reported. For example, using $\sum X_i = 261$ (e.g., rounding up to the nearest integer) gives

$$\bar{X} = \frac{\sum X_i}{n} = \frac{261}{55} = 4.7455$$

which would round up to 4.75 and so cannot correspond to the reported mean of 4.74. Likewise, using a smaller value for the sum of scores such as 260 would give

$$\bar{X} = \frac{\sum X_i}{n} = \frac{260}{55} = 4.727$$

which would round up to 4.73, and thus is too small to match the reported mean of 4.74. Once again, a mean value of 4.74 is mathematically impossible for a sample of size $n = 55$ when measuring ratings with this kind of 1–7 scale.

Note that this kind of inconsistency is sometimes explained by rounding of reported statistics. If the sample size was $n = 46$, a mean of $\bar{X} = 4.74$ would be fine because

$$\sum X_i = (n)\bar{X} = (46)4.74 = 218.04$$

which rounds down to 218. A re-computation of the sample mean gives:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{218}{46} = 4.739$$

which rounds up to match the reported value of 4.74. Thus, here the reported mean is consistent with the sample size, the nature of the scale, and a bit of rounding for reported values.

These types of calculations are referred to as exploring the Granularity-Related Inconsistency of Means (GRIM). Many of the calculations described above can be automated in a spreadsheet. We have provided such a spreadsheet, GrimTest.xls, at the Open Science Framework (https://osf.io/k8yjc/). Enter a reported mean (or proportion) and a sample size, and the spreadsheet indicates whether the numbers make sense.

With a bit of ingenuity and algebra, one can apply the GRIM analysis also to other situations. For example, sometimes an article reports the combined sample size across two samples and proportions or means for each sample but not the specific size of each sample. A variation of the GRIM test might consider all possible sample size combinations that add up to the reported combined sample size and see if any combination is consistent with the reported means or proportions. In some cases, it is possible to use both means and standard deviations to identify inconsistencies.

GRIM inconsistencies can occur because of typos or other forms of sloppiness. They can also happen through QRPs such as removing data from the sum of scores but not taking their removal into account when reporting the sample size. In some cases, a GRIM inconsistency may indicate that the reported data is simply “made

up." Whether based on fraud, tinkering, or a typo, readers of data with a GRIM inconsistency should be skeptical about the reported results and their implications.

Data Extraction Techniques

Many GRIM inconsistencies could be easily resolved if scientists shared their data and analysis code. Regrettably, this is not the norm. Even though many journals formally require authors to share their data, it is uncommon for authors to do so, and the journals often do not ensure that authors follow the rules.

Data sharing also has other advantages, such as allowing a scientific field to take full advantage of a scientist's empirical work by allowing other researchers to explore additional aspects of the data or use it to guide new experiments. Until data sharing becomes common, data detectives can use a variety of techniques to glean some statistical information from reported statistics. Here, we show how some of these techniques can be combined.

Figure 6.1a schematizes the stimuli in a spatial cuing experiment. On each trial, a participant looks at a computer screen that briefly flashes a central arrow pointing to the left or to the right and then shows a target letter either to the left or to the right. The observer's task is to identify the target letter as quickly as possible by making a button-press, and the computer measures their response time. On 80% of the trials, the arrow points to where the target letter is about to appear, so observers learn to attend to the indicated location. The experiment investigates how much such

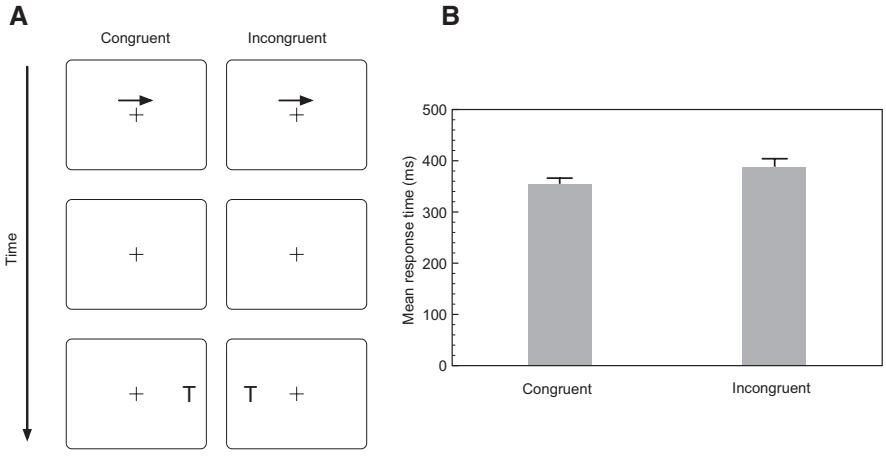

Fig. 6.1 Stimuli and results for a spatial cueing experiment. (**a**) shows stimuli for congruent and incongruent trials. In congruent trials the arrow points to the location of a subsequent target letter. In incongruent trials the arrow points to the opposite location of the target letter. (**b**) shows the results. Mean response time is shorter for congruent trials than for incongruent trials. The error bars indicate one standard error of the mean

attention affects the speed of letter identification. Each observer produces a mean response time for congruent trials (when the arrow points to the location of the target letter) and for incongruent trials (when the arrow points to the opposite side of the location of the target letter). Figure 6.1b shows typical data from $n = 31$ observers (the data are available in a spreadsheet, SpatialCueingData.xlsx, at the Open Science Framework). It indicates that the mean response time is shorter for the congruent than for the incongruent trials. The error bars indicate the standard error across observers for each condition.[2]

It is common to present findings with a data plot (like Fig. 6.1b) along with a summary of a statistical test. Here, the test is a dependent *t* test that compares mean response times for congruent and incongruent conditions: $t(30) = 2.13$, $p = 0.04$. For a data detective, there is more quantitative information than what is directly reported. For example, you might want to know the standard deviations for the conditions and the correlation across observers. The standard error for a condition, $S_{\bar{X}}$, is related to the standard deviation of the data, *S*, by the formula:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

So if we know the standard error, we can easily solve for the standard deviation. The error bars in Fig. 6.1b indicate the standard error, so we just need to extract the information from the plot. We do this using a program called *Plot Digitizer*, which prompts the user to identify the ends of each axis and then click on points of interest in the plot. The program computes the position of each marked point in the plot. Figure 6.2 shows the two windows from *Plot Digitizer* that report the height of each bar and its associated error bar.

The values under the "Condition" column in the small window to the left indicate the x-value of each point, in the order they were clicked. In this bar plot, the x-values simply indicate the two conditions. We are more interested in the values in the "Response Time" column. The first two values refer to the mean and top of the error bar for the congruent condition, and the last two values refer to the mean and top of the error bar for the incongruent condition. The height of the error bar above the mean is the standard error, and so we can compute the standard error $S_{\bar{X}}$ with the following formula:

$$S_{\bar{X}} = \text{Error bar height} - \text{Mean}$$

We can then easily compute the standard deviation *S* for each condition as:

$$S = S_{\bar{X}}\sqrt{n}$$

[2] Sometimes authors compute an error bar using the standard deviation across observers or the range of a 95% confidence interval; it is typical for the figure caption to indicate the basis of each error bar.

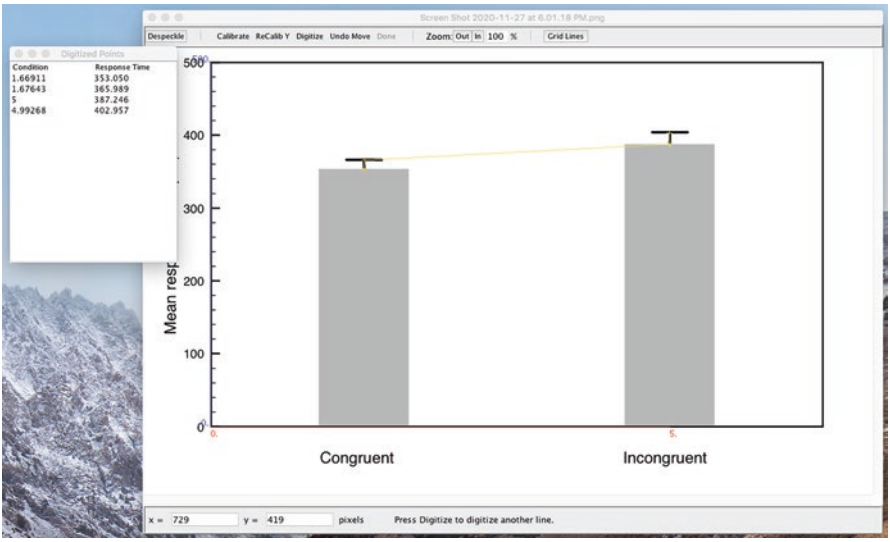

Fig. 6.2 Data gleaning of spatial cueing data using the *Plot Digitizer* program. The yellow lines on the plot connect selected points

Finally, we can compute the correlation between the congruent and incongruent conditions by using the variance sum law, which describes how the variance of difference scores S^2_{x-y} is related to the variance of each score and their correlation r:

$$S^2_{x-y} = S^2_x + S^2_y - 2rS_xS_y$$

Here, we use variables x and y to refer to the two correlated measures (e.g., congruent and incongruent response times). Some algebra shows that the correlation must be:

$$r = \frac{S^2_{x-y} - S^2_x - S^2_y}{-2S_xS_y}$$

We can compute the variance of difference scores from the means and t-value because the t-value is given by:

$$t = \frac{\bar{X} - \bar{Y}}{S_{\bar{X}-\bar{Y}}} = \frac{\bar{X} - \bar{Y}}{S_{x-y} / \sqrt{n}}$$

A bit of algebra results in a formula for the standard deviation of the difference scores S_{x-y}:

$$S_{x-y} = \frac{\bar{X} - \bar{Y}}{t}\sqrt{n}$$

Table 6.1 compares the values gleaned from Fig. 6.1b and the above computations against the values computed directly from the raw data. One can see that the gleaned values are quite close to the actual values. Small discrepancies exist because it is difficult to place the clicks directly on the top of the bars in the plot and because the reported *t*-value is rounded to two decimal places. Using the gleaned values to estimate the correlation between congruent and incongruent response times gives $r = 0.385$. The true correlation (computed from the raw data) is $r = 0.391$.

These extraction techniques can also be used to identify non-obvious inconsistencies in a data set. For example, suppose the text describing a dependent samples *t*-test reported the following, "As predicted, there was a significant difference, $t(30) = 2.8$, $p = 0.009$, between the control ($\bar{X} = 45$, $s = 7.3$) and experimental ($\bar{X} = 55$, $s = 7.6$) conditions." While this result might seem like convincing support for there being a difference in means, it actually makes no sense at all. The reported degrees of freedom for the dependent *t*-test indicates that $n = 31$. Combining this sample size with the reported mean and *t*-values, the standard deviation of the difference scores can be computed using the formula above, $S_{x-y} = 19.88$. Now, we can check whether this value is possible with the standard deviations given for each condition. Solving for the correlation between scores in the control and experimental conditions using the formula above gives $r = -2.56$, which violates the constraint that correlations must always be between plus and minus one. Thus, we can conclude that the reported numbers cannot be correct.

This section has mostly dealt with mathematical inconsistencies in statistical reports. Standard reporting formats include redundant information that sometimes allow data detectives to identify inconsistencies. With these methods, the data detective checks for inconsistencies in the reported results of a single experiment. In the next section we identify two methods for characterizing inconsistencies *across* experiments.

Table 6.1 True and gleaned values for the means, standard errors, and standard deviations of the spatial cueing data

Statistic	Congruent		Incongruent	
	True	Gleaned	True	Gleaned
Mean	353.64	353.05	387.85	387.25
SE	12.45	12.94	16.10	15.71
SD	69.30	72.04	89.63	87.48

Experimental Inconsistencies

Hypothesis testing is often presented as a way of drawing conclusions within a single experiment. However, sometimes conclusions are based on statistical outcomes *across* experiments, and the properties of hypothesis testing impose important constraints in such situations. We will describe two analysis methods that look for violations of these constraints. Conceptually, identifying inconsistencies across experiments is similar to identifying mathematical inconsistencies within an experiment. However, there are two important differences. First, mathematical inconsistencies could potentially be due to typos or calculation errors rather than QRPs. The same interpretation is usually not plausible for inconsistencies across experiments. Second, mathematical inconsistencies are definitive in the sense that there is no way for the numbers to make logical sense. Inconsistencies across experiments, on the other hand, are defined as improbable (rather than impossible) inferential outcomes. These inconsistencies suggest the involvement of QRPs because observed outcomes would be very rare if QRPs were not involved.

Test for Excess Success

In most experiments in clinical psychology, conclusions are based on hypothesis testing. Due to how samples are randomly selected for such tests, it sometimes happens that a test draws the wrong conclusion. For example, it is possible that a population with a true null hypothesis produces a significant outcome simply due to the scientist happening to get an unusual sample of data. The hypothesis testing procedure for drawing a conclusion controls the rate of making such a Type I error; and scientists typically set that rate to be 5%. Likewise, it is possible that a population with a true effect produces a non-significant outcome due to the scientist happening to get an unusual data sample. The probability of making such a Type II error is not directly controlled in hypothesis testing, unless the scientist has a good idea of the size of the true effect and gathers a large enough sample of data.

An important implication of drawing conclusions based on hypothesis tests is that mistaken conclusions are *inevitable*. Even when doing everything correctly (in terms of random sampling, analyzing the data, and reporting the results), scientists *must* sometimes make the wrong decision. Consider the *power* of an experiment. Power is the complement of Type II error, meaning that it refers to the probability that a hypothesis test based on a random sample of data will reject the null hypothesis when this is the correct conclusion (there really is an effect). Power depends on the size of the effect and on the size of the sample, in that larger effects and larger samples give higher power. Oftentimes, scientists do not try to control power, because the effect size is unknown. When power is considered, scientists often aim for sample sizes that give at least 80% power. However, this is an arbitrary target, and it is sometimes inappropriate. Consider a scientist who plans two independent

experiments, and will draw a conclusion in favor of some theoretical conclusion only if both experiments show significant effects. If each experiment has 80% power, then the probability of both experiments producing significant results is $0.8^2 = 0.64$. Thus, even though the power of each experiment is acceptable when considered alone, the odds of the scientist finding support for their theoretical conclusion are only slightly better than a coin flip.

As additional successful experiments are added to the list of requirements for drawing a theoretical conclusion, the probability of consistent success decreases. Out of 20 experiments, one should expect on average $0.8 \times 20 = 16$ significant outcomes. The probability of all 20 experiments producing significant outcomes is only $0.8^{20} \approx 0.01$. Thus, if a scientist reports that 20 out of 20 experiments each with a power of 0.8 produced significant outcomes, this should not be interpreted as strong evidence for the theoretical conclusions but instead as an indication that something has gone wrong; in particular it suggests that the scientist engaged in some types of QRPs. The absence of non-significant findings in experiments with limited power is a marker for flaws in the scientific process because the reported findings seem "too good to be true."

These observations can be quantitatively formalized with the Test for Excess Success (TES). By estimating effect magnitudes from the reported experiments, this method estimates the success rate of future experiments that use the same sample sizes. The success rate is an estimate of the probability of future replication experiments to produce the same degree of success as the original experiments. If this rate is low (0.1 is a common, if arbitrary, threshold), then the reported results of the original studies are deemed problematic (too good to be true).

To demonstrate how to perform a TES analysis, consider a prominent paper that reported six experiments investigating the impact of poverty on cognitive performance. The main claim was that poverty-related concerns use mental resources that would otherwise be available for other tasks. This claim implies that poor people make bad choices partly because they are poor, rather than being poor because they make bad decisions. If true, this finding has many important policy implications. When deciding on how to best help poor people, one needs to consider their lower cognitive capabilities, which may vary with their financial situation. The paper describing these six experiments was published in the journal *Science*, which is widely regarded as the most prestigious scientific academic journal, and the findings were considered important enough to merit mention in the *New York Times* and numerous other media outlets. Below, we use a TES analysis to show that these results actually do not adequately support the theoretical claims. Arguably, some of the findings were produced with QRPs.

For each of the six studies, we can extract the statistics for the relevant hypothesis tests. For most of the studies, multiple hypothesis tests were performed. However, to keep the current analysis simple, we estimate an upper limit of the success rate for each experiment by considering only the statistically weakest relevant test. This approach is conservative, since an experiment is always less likely to produce multiple specific outcomes than only one of the outcomes.

A key result from Experiment 1 was an interaction between income (rich or poor, defined by a median split) and condition (scenarios describing hard or easy to manage financial difficulties). The measurements included performance on a Raven's matrices task (a measure of fluid intelligence) and a cognitive control task. To estimate an upper limit of the power of Experiment 1, we used the weaker of the results from these two measures. The calculation of power is done in an R program (TESAnalysis.R) that is available for download at the Open Science Framework. Without going into the specific formulas, the program converts the sample sizes and test statistics (F- or t-value) into a standardized effect size (Hedges' g). This standardized effect size is then used to estimate the probability that a new experiment with the same sample size as the original experiment would produce a significant outcome. As the first row of Table 6.2 indicates, the power is around 0.6. So, if the effect is real and similar to what was originally reported, future replication studies with the same sample size have around a 60% chance of producing a significant outcome.

Experiment 2 was similar to Experiment 1, but with nonfinancial scenarios. The prediction of the authors was that this design would *not* produce a significant difference between rich and poor participants; and that was precisely what they reported. The success probability for this experiment is computed as one minus power, which gives the probability of a random sample *not* producing statistical significance. As shown in Table 6.2, the success probability is rather high because it is easy to not produce a significant outcome with a small sample.

Experiment 3 added monetary incentives for correct responses and found similar effects as for Experiment 1. Namely, there was a significant interaction between income and scenario for measures of cognitive control. If the effect is similar to what is reported in Experiment 3, then the power of a replication experiment with the same sample sizes is just above 0.5.

Table 6.2 Estimated success probabilities for six experiments investigating poverty and cognition. The probability of all six experiments producing successful outcomes is so low (0.065) that the results seem too good to be true

Experiment	Test	Reported statistics	Success probability
1	Interaction for Raven's matrices	$F(1,97) = 5.12$, $p = 0.03$	0.602
2	Non-significant difference for rich and poor on cognitive control	$F(1,35) = 1.69$, $p = 0.20$	0.764
3	Interaction for cognitive control	$F(1,98) = 4.31$, $p = 0.04$	0.532
4	Interaction for Raven's matrices	$F(1,92) = 4.04$, $p = 0.04$	0.505
Field 1	Pre- and post-harvest differences	$p < 0.001$	~1
Field 2	Pre- and post-harvest heart rate (stress)	$t(187) = 1.715$, $p = 0.088$	0.523
P_{TES}			0.065

Experiment 4 was very similar to Experiment 1, but with a different order of some tasks. A key result is an interaction between income and condition for the Raven's matrices task. Power for a replication experiment is barely above 0.5. We should note that the reported statistics for Experiment 4 show a *p*-value inconsistency. A recalculation shows that $F(1,92) = 4.04$ corresponds to $p = 0.047$ rather than the reported $p = 0.04$. For the TES analysis, we assume that the reported *F*-value is correct.

To explore the generality of the findings beyond the controlled settings of Experiments 1–4, two field studies were run to investigate cognitive performance for farmers in India. The first field study found strong differences in cognitive performance for farmers pre-harvest (when they are relatively poor) compared to post-harvest (when they are relatively wealthy). The original text does not report sufficient statistical information to compute power of a replication study, but the reported *p*-values are small, so the estimated power will be close to 1.

The second field study also found cognitive effects pre- and post-harvest, and this study concluded that the effect is not because of nutritional differences (food consumption was similar pre- and post-harvest) but seemed to be due to stress (farmers had a higher heart rate pre- compared to post-harvest). The authors of the study used a non-typical significance criterion of 0.1 rather than the usual 0.05. In our analysis, we suppose that a deviation from the norms of hypothesis testing was appropriate, and we calculated power with this atypical significance criterion. Regardless of these details, the probability of a replication study showing a significant result is only a bit above 0.5.

The probability that six independent experiments like these should *all* be successful (a non-significant test outcome for Experiment 2 and significant test outcomes for the other studies) is the product of the probabilities in Table 6.2, which is 0.065. Thus, if the effects are real and similar to what is reported, studies like these are unlikely to produce six successful outcomes. Given the rarity of the observed results, scientists should be skeptical that the reported experiments are representative of reality. The studies described in the original paper do not make a strong argument for poverty having the hypothesized impact on cognition, and it remains an open question whether this effect actually exists.

A reasonable interpretation of our TES analysis result is that the authors of the original study engaged in some kind of QRPs in order to produce their reported results. The TES analysis cannot differentiate between different types of QRPs, and it is possible that the authors themselves do not know what kinds of choices they made to produce success across their experiments. Regardless of the origins of the problems, the bottom line is that the reported results are unlikely to represent reality. We advise readers to ignore the reported findings and wait for (or plan) better experiments.

P-Curve Analysis

If the null hypothesis is true, then *p*-values across experiments are approximately uniformly distributed. That is, the *p*-value is equally likely to take any value between 0 and 1. At first glance, this might seem like a very strange claim, but it is actually intuitive once you understand how *p*-values are related to Type I error control.

Remember that in hypothesis testing the scientist defines a significance criterion, α, to set the probability of picking a random sample that rejects a true null hypothesis. The scientist then computes the *p*-value for their data and compares it to α. If $p < \alpha$, then the null hypothesis is rejected. Importantly, this procedure works for any value of α. Thus, if $\alpha = 0.05$ and the null hypothesis is true, there is a 0.05 probability of picking a random data set that produces a *p*-value smaller than 0.05. If $\alpha = 0.10$, then there is a 0.1 probability of picking a random data set that produces a *p*-value lower than 0.1. Just to continue the example, if $\alpha = 0.34294$, then there is a probability of 0.34294 of picking a random data set that produces a *p*-value below 0.34294. This property indicates that the probability of observing a *p*-value smaller than any value *x* is precisely *x*. This is the definition of a uniform probability distribution.[3]

Figure 6.3a shows the distribution of *p*-values for simulated one-sample *t*-tests when the null hypothesis is true (effect size equals zero). Here, simulated data were drawn from a standard normal distribution and then analyzed with a one-sample *t*-test for H_0: $\mu = 0$. This simulated experiment was repeated 10,000 times, and the histogram shows that the resulting *p*-values are approximately uniformly distributed across the interval 0 to 1. The gray vertical line indicates the 0.05 criterion for statistical significance. As intended for a true null hypothesis, about 5% of the *p*-values fall below this criterion. R code, pValues1.R, to reproduce the plots in Fig. 6.3a is available at the Open Science Framework.

The situation is quite different when the null hypothesis is false. When there really is an effect, the distribution of *p*-values is positively skewed, with more small *p*-values than large *p*-values. Figure 6.3b shows the distribution of *p*-values when the standardized effect size is 0.2 and the sample size is $N = 50$. The skew is intuitive if we consider the fact that increasing the effect size leads to increased power (the probability of picking a sample that rejects the null hypothesis). In this case, power is 0.284, and so 28.4% of the *p*-values must fall below the 0.05 criterion. As power increases, the distribution of *p*-values becomes more skewed with more very low *p*-values. Figures 6.3c and d show this property for larger effect sizes (and thus higher power). In fact, the shape of the *p*-value distribution for a given test is entirely determined by the power of the test. Figures 6.3e and f show *p*-value distributions

[3] There are some situations where *p*-values do not follow a uniform distribution even when the null hypothesis is true. For example, a test of proportions with a small sample size is constrained by combinatorics to produce some *p*-values and not others; therefore the *p*-values will not follow a uniform distribution. Likewise, a test of means may show a small preference for some *p*-values due to rounding characteristics of mean measurements. These issues aside, the distribution of *p*-values is close to uniform for many hypothesis tests when the null is true.

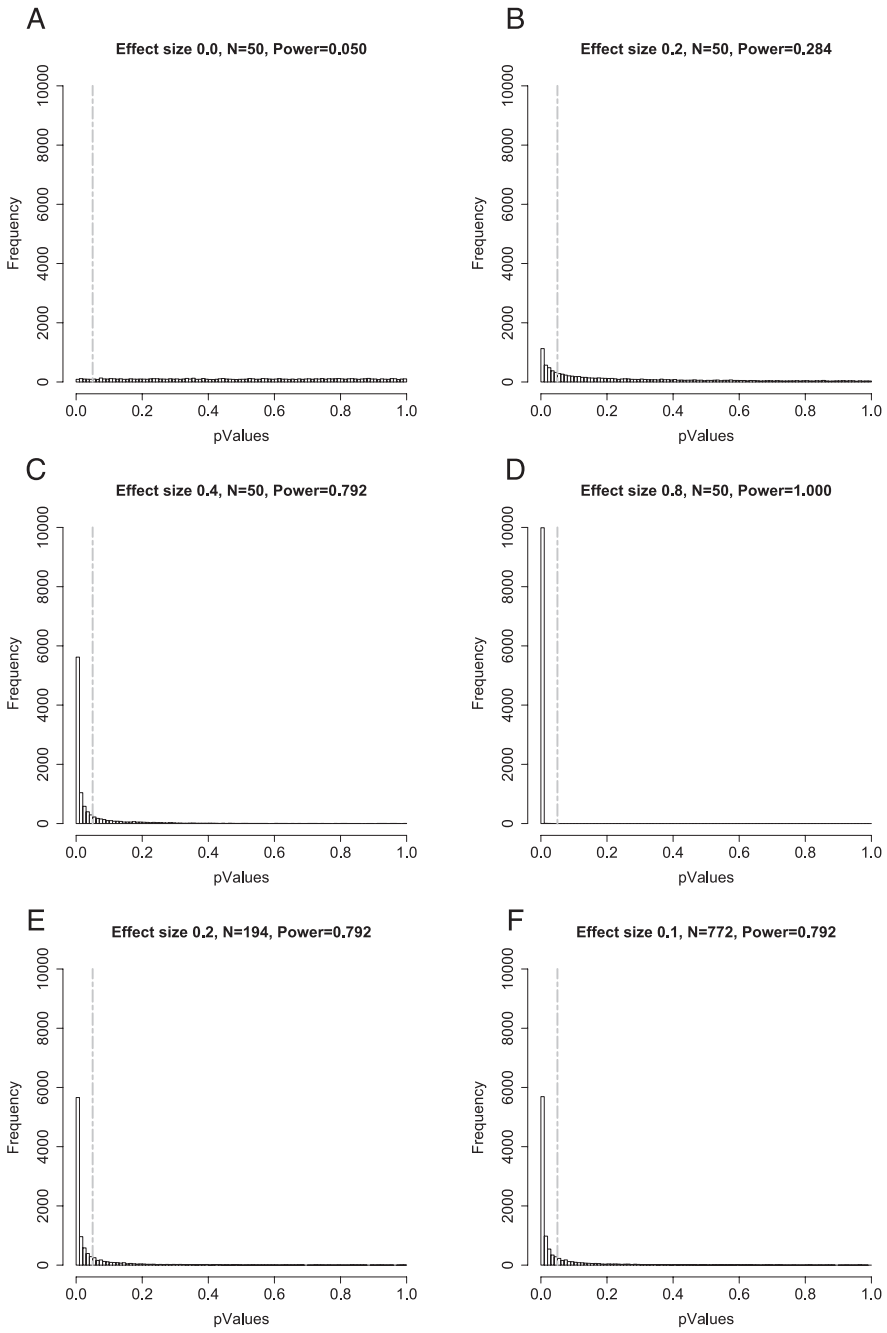

Fig. 6.3 Histograms characterizing *p*-value distributions for tests with different power values. The vertical gray line indicates the significance criterion (0.05). The histogram interval width is 0.01

for combinations of effect sizes and sample sizes that give the same power value as in Fig. 6.3c. The *p*-value distributions are essentially the same (small deviations are due to random sampling in the simulations).

Importantly, these properties hold even when considering only significant (e.g., $p < 0.05$) findings. Figure 6.4 plots *p*-value distributions for significant *p*-values (between 0 and 0.05). When the null hypothesis is true, the distribution is uniform (Fig. 6.4a). For non-zero (real) effects, the *p*-value distribution is skewed, with a preponderance of very small *p*-values (Fig. 6.4b–d). The code to reproduce these simulations, pValues2.R, is available at the Open Science Framework.

The *p*-values for each histogram in Figs. 6.3 and 6.4 were generated from experiments that have the same effects and sample sizes (and thus the same power). Should experiments differ in sample sizes or effect sizes (and thus in power), the curves are different, but the general shape (e.g., positive skew) continues to hold. Thus, a set of experiments with some (different) real effects should produce a distribution of

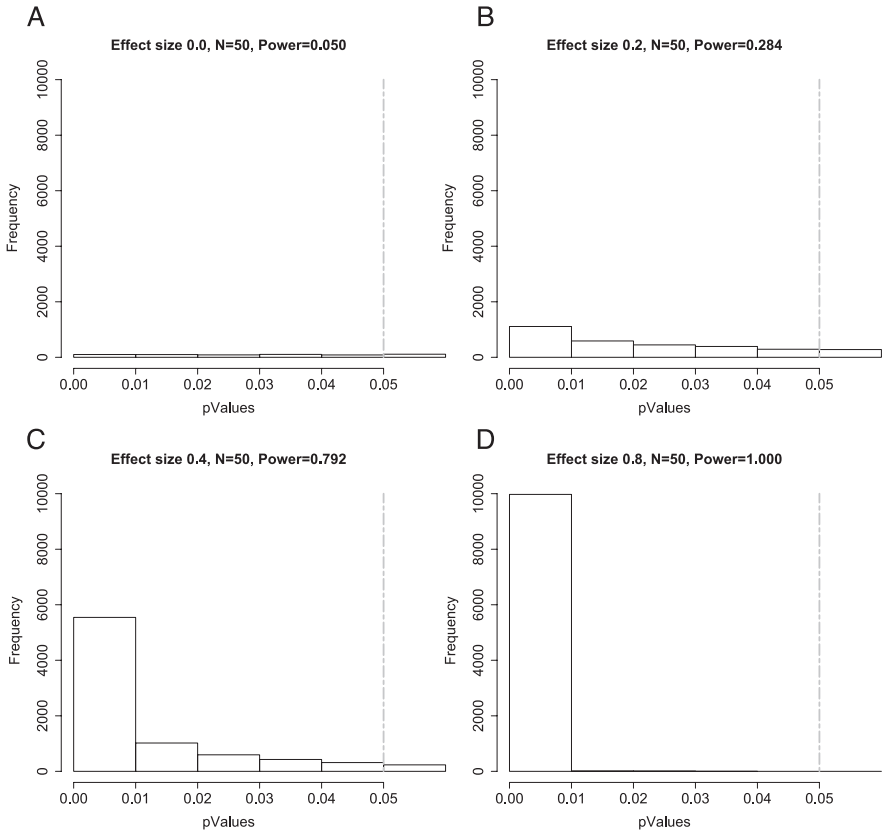

Fig. 6.4 Histograms characterizing *p*-value distributions between 0 and 0.05 for tests with different power values. The properties are the same as for the histograms in Fig. 6.3. The vertical gray line indicates the significance criterion (0.05)

p-values that has positive skew. Likewise, a set of experiments that entirely investigates (maybe different) null effects should produce a distribution of *p*-values that is flat. A set of experiments that contains some true null effects and some real effects will produce a *p*-value distribution with positive skew.

As data detectives, we can make use of the *p*-value distribution. First, we note that its shape is essentially unaffected by publication bias (a bias to only report significant outcomes): Even if only significant outcomes are published, the distribution of *p*-values below the significance criterion differs between null and real effects and true null effects will produce something close to a uniform distribution. Moreover, there are other problems that an analysis of the *p*-value distribution can efficiently identify. For example, left-skewed distributions are a sign of QRPs because such distribution shapes should be very unlikely if data collection and statistical analyses are done properly. The online app at http://p-curve.com automates analyses of the *p*-value distribution. Figure 6.5 plots the *p*-curve generated by a set of experiments that explored how the placement of calorie labels (before or after a menu item) influenced selection of foods with high calories. Across six studies (three in the main text and three in supplemental material), researchers consistently found significant effects that indicated that placing the calorie labels before a menu item led people to order lower calorie foods. The test statistics for these studies are shown in the small window in Fig. 6.5 (note that the test statistic for one study was taken from a corrigendum provided by the authors to fix a small error in their data set). The

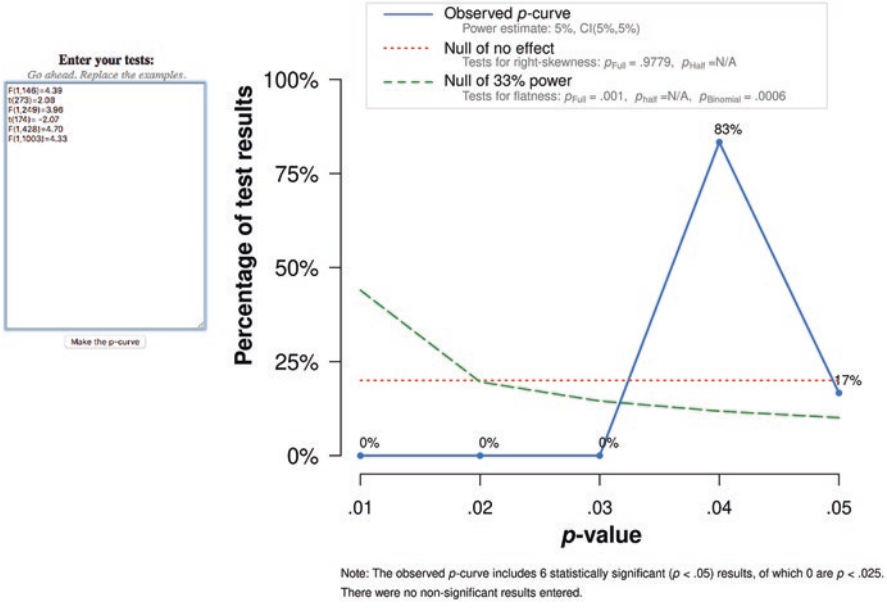

Fig. 6.5 Results of the *p*-curve app for six studies investigating the impact of calorie information on menu choices. The solid blue curve reflects the frequency of reported *p*-values. It is left skewed, which is not how *p*-values should be distributed

researchers argued that putting the calorie information in the leading position makes it more prominent in memory and therefore more influential than when it is placed after the menu item (importantly, placing the calorie information after the menu item is standard in the United States). However, the distribution of *p*-values for these six studies suggests that something is wrong with this set of results. The blue curve in Fig. 6.5 reflects the reported *p*-values, and there are none smaller than 0.04. Such a left-skewed distribution should be very unlikely if the studies are run correctly. The online app includes statistical tests for evaluating the distribution of *p*-values relative to a null (uniform) distribution and to what they refer to as an "inadequate" distribution (the *p*-value distribution for studies with power of 0.33), which is described by the green line in Fig. 6.5. The app also reports a test for whether the studies contributing to a *p*-value distribution contain "evidential value," meaning that the distribution is right skewed. For these tests, the *p*-curve analysis indicates that the evidential value is inadequate (the empirical curve is flatter than a curve with power of 0.33) and it does not indicate evidential value (the empirical curve is not right skewed).

A reasonable interpretation of the *p*-value distribution in Fig. 6.5 is that some of the experimental results were generated with QRPs. It is not possible to identify precisely what QRPs were used, but we should not trust that the reported results or corresponding conclusions reflect reality. The conclusions may yet be correct, but the reported experiments do not provide appropriate support for those conclusions. Scientists who want to investigate this topic further need to start over with better experiments.

It is fairly easy to apply the *p*-curve analysis, but it is important to understand its requirements and interpretation. One requirement is that the *p*-values that contribute to the distribution must be independent. It is sometimes the case that a set of data is analyzed with multiple hypothesis tests (e.g., an ANOVA reports an interaction and specific contrasts with the same data set). The *p*-values from these tests are (typically) not independent, and so the tests to explore the distribution shape can be misleading. To address this concern, researchers using the *p*-curve analysis use just one *p*-value from each data set or experiment. Unfortunately, it is not always clear how to select a *p*-value from the set, and the choice can make a big difference. For example, choosing the smallest *p*-value from each experiment will often result in a distribution with right skew even when the null hypothesis is true. Likewise, choosing the biggest significant *p*-value from each set will often produce a left skewed distribution, even when there is a real effect. To avoid this problem, some researchers apply an arbitrary rule, such as using the *p*-value from the first reported relevant test; but this does not really address the fundamental problem: the analysis should be based on the *p*-values that are relevant to the question of interest. It often requires subject matter expertise to identify such *p*-values, and sometimes there is not a unique *p*-value that relates to the question of interest.

For the *p*-curve graph in Fig. 6.5, the question of interest is, "does the location of caloric information influence menu choices?" and we picked the *p*-values that specifically investigated that question. The resulting left skewed *p*-curve distribution indicates that the six studies reported here were not produced by proper hypothesis

tests. Importantly, this conclusion does not mean that *each* of the six tests is flawed. The identified problem is with the *set* of hypothesis tests (their distribution of *p*-values). Surely some of the individual studies are problematic as well (else the set could not be), but it is possible that some studies are flawed and some studies are fine.

This aspect of interpretation can matter quite a bit for other types of questions of interest. For example, suppose you applied a *p*-curve analysis to a specific researcher because you wonder if he engages in QRPs. You select one *p*-value from each of seven articles published by this researcher. Figure 6.6 shows the (entirely made up) *p*-curve for the selected *p*-values. It is right skewed, so the *p*-curve analysis suggests that there is "evidential value" in this set of *p*-values. Unfortunately, this conclusion does not really answer the question of whether the researcher engages in questionable research practices. It could be that the researcher does not engage in QRPs, but it could also be the case that for some investigations the researcher does use QRPS and for some investigations he does not. Publishing some studies with evidential value means that a combination of studies with evidential value and studies without evidential value (e.g., a flat distribution) might produce a right skewed distribution of *p*-values. The point is that a property of the set does not necessarily apply to each

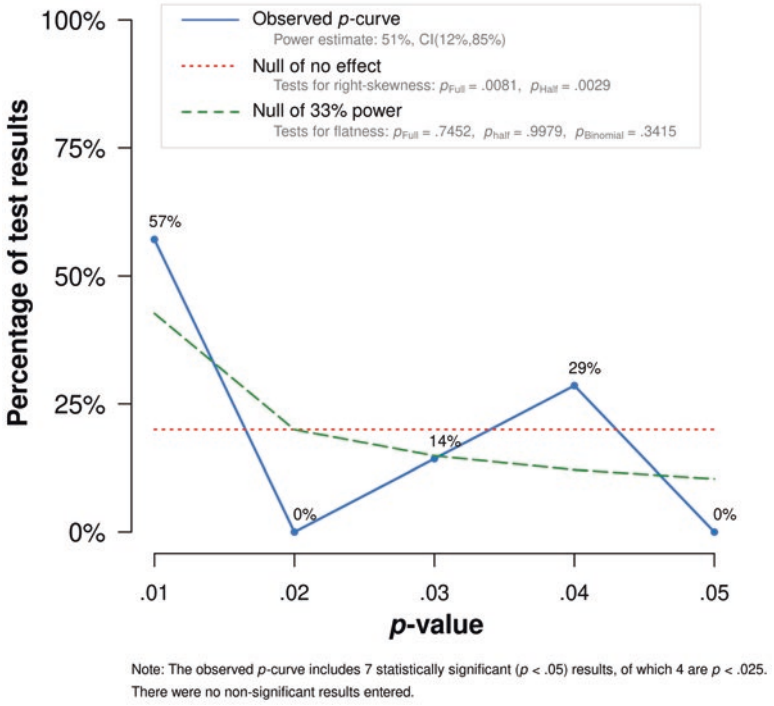

Fig. 6.6 Results of the *p*-curve app for seven (hypothetical) studies investigating a researcher who investigates two different topics. Although the distribution is right-skewed, thereby indicating some "evidential value," this finding is difficult to interpret

member of the set. A right skewed *p*-curve does not mean that every study a researcher reports is fine, and a left skewed *p*-curve does not mean that every study a researcher reports is problematic. For this reason, it usually does not make sense to apply *p*-curve analyses to an author, a specific scientific journal, or a field of study. Instead, *p*-curve analyses should be used to evaluate specific claims or conclusions, when those claims or conclusions are based on a reported set of *p*-values. For the studies producing the *p*-values in Fig. 6.6, it might make sense to look into the set of studies related to specific conclusions made by the researcher, and use the *p*-curve analysis to evaluate the evidential value of the studies relative to those claims.

Conclusions

Questionable Research Practices (QRPs) often leave a trail of evidence that indicates they were involved in producing the reported outcomes. Proper experiments (without QRPs) have fundamental properties that can be identified across experiments. One such property is how success should relate to experimental power. Excess success for a set of experiments indicates that the results were generated in a way that violates good data collection, analysis, or reporting. This discrepancy can be identified with the Test for Excess Success. A second such property is the distribution of *p*-values, which should be right-skewed for proper experiments that investigate a real effect. The distribution of *p*-values should almost never be left-skewed for experiments that were generated without QRPs. A left-skewed distribution indicates that the results were generated in a way that violates good data collection, analysis, or reporting. These problems can be identified by the *p*-curve analysis.

Within a single experiment, it is often useful to look for various discrepancies between reported statistics. Such discrepancies do not necessarily indicate the involvement of QRPs, but they do suggest that something has gone wrong in the reporting of the experiment. Thus, readers should be somewhat skeptical about the validity of the reported results and the associated conclusions.

As is the case for many types of detective work, a data detective may be able to conclude that there is something "odd" about reported results but not pinpoint exactly what has gone wrong. Inconsistencies between statistics might arise from fraud or they might be the result of simple typos. In a similar way, neither the Test for Excess Success nor the *p*-curve analysis can identify precisely *how* researchers produced results that are too-good-to-be-true or that generate a left-skewed distribution of *p*-values. Still, the burden of proof is on the scientists; they should always provide evidence to support their claims. If the reported results seem unbelievable, other scientists should dismiss the claims until sufficient evidence is produced.

While some scientists may deliberately set out to deceive others, we suspect that most scientists introduce QRPs without realizing it. Indeed, one very beneficial use of the various methods for detecting the impact of QRPs is for scientists to apply them to their own work before publishing. Hopefully, this could motivate scientists to examine their research methods in detail and root out QRPs. Such applications will greatly improve scientific work.

Further Reading

GRIM Test

Brown, N. J. L., & Heathers, J. A. J. (2016). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science, 8*(4), 363–369. https://doi.org/10.1177/1948550616673876

Heathers, J. (2017). Introducing SPRITE (and the case of the carthorse child). *Hackernoon*, https://medium.com/hackernoon/introducing-sprite-and-the-case-of-the-carthorse-child-58683c2bfeb

Heathers, J. A., Anaya, J., van der Zee, T., & Brown, N. J. (2018). Recovering data from summary statistics: Sample parameter reconstruction via Iterative techniques (SPRITE). *PeerJ Preprints, 6*, e26968v1. https://doi.org/10.7287/peerj.preprints.26968v1

Test for Excess Success

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review, 19*, 151–156. https://doi.org/10.3758/s13423-012-0227-9

Francis, G., & Thunell, E. (2019). Excess Success in "Ray of hope: Hopelessness increases preferences for brighter lighting". *Collabra: Psychology, 5*(1), 22. https://doi.org/10.1525/collabra.213

P-Curve

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General, 143*, 534–547.

Calorie Labels

Dallas, S. K., Liu, P. J., & Ubel, P. A. (2019). Don't count calorie labeling out: Calorie counts on the left side of menu items lead to lower calorie food choices. *Journal of Consumer Psychology, 29*(1), 60–69. https://doi.org/10.1002/jcpy.1053

Francis, G., & Thunell, E. (2020). Excess success in "Don't count calorie labeling out: Calorie counts on the left side of menu items lead to lower calorie food choices". *Meta-Psychology, 4*. https://doi.org/10.15626/MP.2019.2266

Poverty and Cognition

Mani, A., Mullainathan, S., Shafir, E., & Zhao, J. (2013). Poverty impedes cognitive function. *Science, 341*, 976–980.