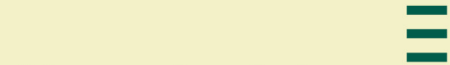

Cambridge Elements

Perception

Hypothesis Testing Reconsidered

Gregory Francis

ISSN 2515-0499 (online)
ISSN 2515-0499 (print)

Cambridge Elements

Elements in Perception
edited by
James T. Enns
The University of British Columbia

HYPOTHESIS TESTING RECONSIDERED

Gregory Francis
Purdue University

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108730716
DOI: 10.1017/9781108582995

First published 2019

A catalogue record for this publication is available from the British Library.

ISBN 978-1-108-73071-6 Paperback
ISSN 2515-0502 (online)
ISSN 2515-0499 (print)

Hypothesis Testing Reconsidered

Elements in Perception

DOI: 10.1017/9781108582995
First published online: May 2019

Gregory Francis
Purdue University

Author for correspondence: gfrancis@purdue.edu

Abstract: Hypothesis testing is a common statistical analysis for empirical data generated by studies of perception, but its properties and limitations are widely misunderstood. This Element describes several properties of hypothesis testing, with special emphasis on analyses common to studies of perception. The author also describes several challenges and difficulties with using hypothesis testing to interpret empirical data. Many common applications of hypothesis testing inflate the intended Type I error rate. Other aspects of hypothesis tests have important implications for experimental design. Solutions are available for some of these difficulties, but many issues are difficult to deal with.

Keywords: hypothesis testing, statistics, signal detection theory, Type I error, excess success

ISBNs: 9781108730716 (PB), 9781108582995 (OC)
ISSNs: 2515-0502 (online), 2515-0499 (print)

Contents

1 Introduction

In many scientific fields, there is no better start to a results section than, "As predicted, we found a significant difference between . . . " Finding a significant difference (e.g., $p < 0.05$) allows authors to affirm their beliefs about, for example, color perception, attention, the workings of visual circuits, or how people search for targets in a cluttered display. Many scientists learned the basics of hypothesis testing as undergraduate students, and they learned to deal with more complicated tests (e.g., multi-way ANOVA, ANCOVA, mediation, moderation) as graduate students. Statistical analyses are central to modern investigations of psychology, including perception, and hypothesis testing is a common approach to statistical analysis.

Despite playing a central role, many properties of hypothesis tests are misunderstood. These misunderstandings can lead to scientific articles that make no sense and to experiments that are so poorly conceived that it was never appropriate to run them. Over the past seven years, psychology has experienced a "replication crisis," whereby some important findings do not hold up when independent scientists repeat the experiment; much of the crisis seems to be related to inappropriate uses of hypothesis testing.

With this issue in the background, it might be useful to characterize some confusions about hypothesis testing and to describe its assumptions and limitations. Throughout this Element, we provide examples of how the issues impact the design and interpretation of perception studies. This discussion is not meant to be a critique of hypothesis testing itself; although after considering all the challenges, you may decide that hypothesis testing is not worth the effort. Alternative approaches include a focus on estimation (Cumming, 2014), Bayesian methods (Kruschke, 2010; McElreath, 2016), and information criterion methods (Burnham & Anderson, 2002), but they are not discussed here.

The target audience for this Element is someone who has already taken one (or more) statistics courses and uses hypothesis testing. The discussion requires little explicit mathematics (and there are no theorems!), but a general understanding of sampling distributions, *p*-values, and power is probably going to be necessary for the reader to follow all the arguments. The selected topics represent issues that have been raised over the past few years in discussions with colleagues and students. Readers may be disappointed to discover that the text sometimes identifies problems without proposing solutions, but it may be useful to discover that there remain unsolved problems in the use of hypothesis testing. Indeed, an overall theme of the Element is that the proper use of hypothesis testing is rather more complicated than generally believed. While

the basic idea is simple and appealing, the actual use is often quite complicated, and some common practices undermine the tenets of hypothesis testing.

2 The Basics of Hypothesis Testing

Hypothesis testing offers an appealing approach to data analysis. Follow the rules and you will make a Type I error (conclude there is an effect when there really is no effect) only 5% of the time (or whatever criterion you set). Such Type I error control sounds really good because it aligns with the natural skepticism of a scientist who doubts an effect exists unless there is sufficient reason to believe otherwise.

Hypothesis testing is also pretty easy to apply. We create a quantitative null hypothesis that indicates "no effect" (e.g., population means equal each other across two conditions) and then predict properties of our data set if that null hypothesis is true. A fundamental concept here is the sampling distribution, which describes how common it should be to find various values of a sample statistic if the null hypothesis is true. The test essentially checks whether the statistic computed from the observed data is among the "rare" statistics in the sampling distribution by computing the probability that the observed data or something even more extreme would occur. This probability is the *p*-value. See Figure 1.

The details get more complicated for other analyses, but the basic reasoning is the same as that given earlier. Assume the null is true and estimate the probability of the observed (or more extreme) statistic under that assumption. If the probability is low (e.g., less than 0.05), reject the null hypothesis: conclude statistical significance. By definition, if everything is done properly, you should only make a Type I error (reject the null hypothesis when it is actually true) at your criterion rate (e.g., 0.05).

A key part of that last sentence is "if everything is done properly." Lots of things can go wrong when doing hypothesis testing, even when scientists are operating with the best of intentions. As we will see in the following sections, even seemingly small deviations from the proper procedures for hypothesis testing can cause the Type I error rate to be much larger than intended.

2.1 An Example from Perception

The stimuli in Figure 2a show the Muller–Lyer illusion: the horizontal line with outward wings appears to be longer than the horizontal line with inward wings. To measure the size of the illusion, n=310 observers adjusted the length of a line with wings so that it appeared to be the same length as a comparison line of 100 pixels long with no wings. See the Appendix for details on how to get the data set. Each observer made eight matches for the inward wing and outward wing

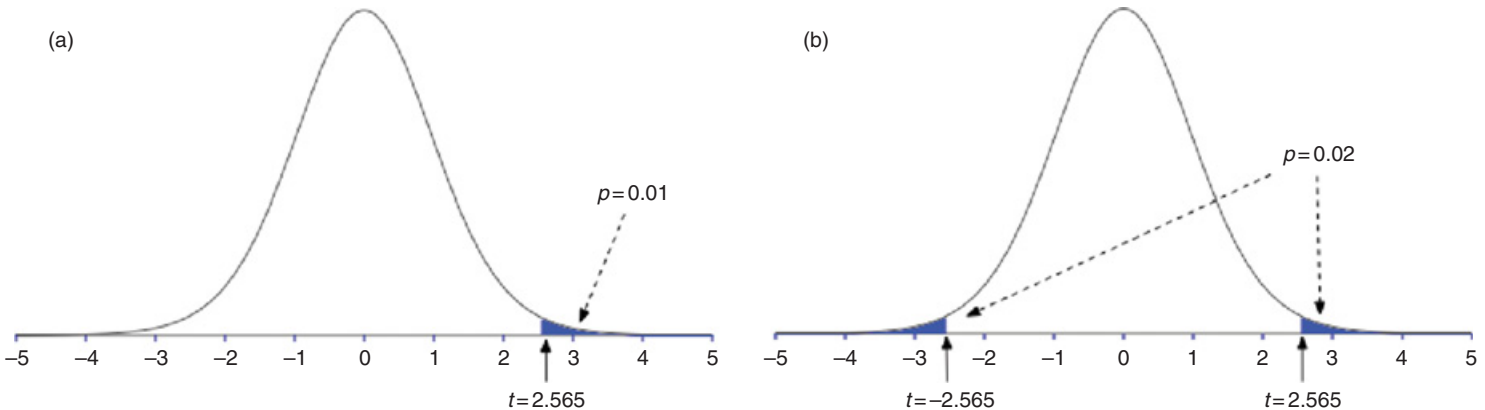

Figure 1 The *p*-value is the area under the curve of the sampling distribution beyond the observed test statistic. Here, the sampling distribution is for the *t*-value statistic that compares two sample means. (a) For a positive one-tailed test, the area is more extreme in the observed direction. (b) For a two-tailed test, the area is more extreme than the observed value in both tails.

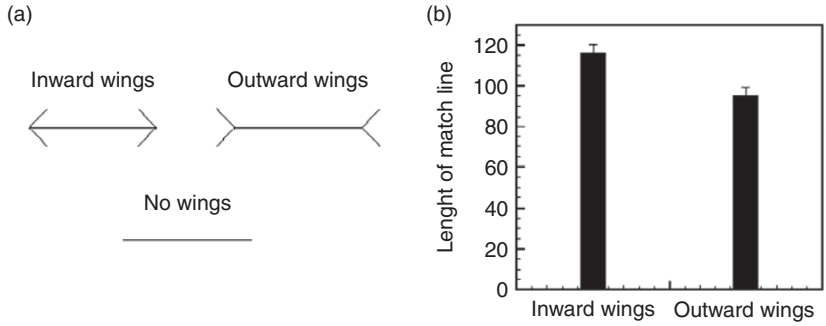

Figure 2 Stimuli and summary data for an experiment on the Muller–Lyer illusion. (a) A line with outward wings looks longer than a line with inward wings. (b) Mean line lengths for lines with the inward or outward wings so that they appeared to be the length of a 100-pixel line with no wings. The error bars indicate the standard deviation.

conditions, and the observer's score was the mean match length across the comparisons for each condition. Figure 2b plots the average match length across the 310 observers for each wing type. As expected, the match length is smaller than 100 pixels when the line has outward wings (a 92-pixel-long line with outward wings looks to be 100 pixels long). Likewise, the match length is longer than 100 pixels when the line has inward wings (a 112-pixel-long line with inward wings looks to be 100 pixels long).

A dependent two-sample hypothesis test comparing the means for the two wing conditions requires the sample size (in this case n=310) and computation of the sample means, standard deviations, and correlation of subject scores across the conditions ($\overline{X}_{\text{Inward}} = 112.3$, $s_{\text{Inward}} = 8.1$, $\overline{X}_{\text{Outward}} = 91.5$, $s_{\text{Outward}} = 8.0$, $r = 0.522$). With this information, the standard deviation of the difference of paired scores is computed to be $s_{\text{Difference}} = 7.87$ and the test statistic is t=46.6 with df=309, which corresponds to p<0.001. If there were truly no difference in the mean line lengths for the population of observers, then a random sample of 310 observers that produced a t-value test statistic at least as large as what we observed would be extremely rare. In practice, we say that the observed difference is "significant."

Take away message: When done properly, hypothesis testing controls the Type I error rate and the calculations are fairly easy to perform.

3 Robustness of the Two-sample *t*-test

A canonical hypothesis test is the two-sample t-test that compares two independent means. Our undergraduate classes told us that the t-test requires two

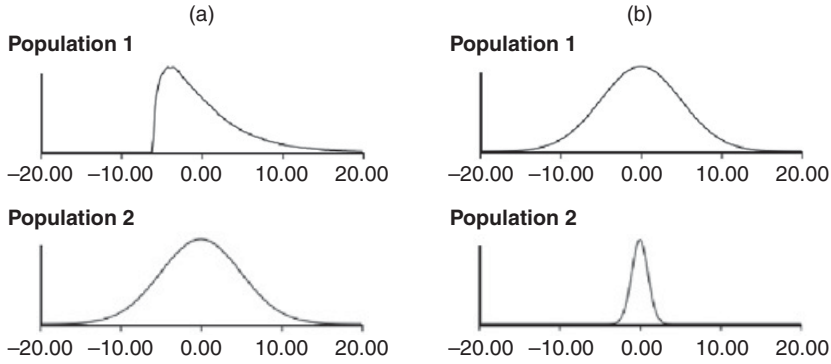

Figure 3 Exploring robustness of the *t*-test for two independent sample means. Here, every population distribution has a mean of zero. (a) Although the *t*-test assumes normal population distributions, even very skewed population distributions do not cause severe problems. For these populations the Type I error rate is 0.051. (b) Normal population distributions with unequal standard deviations. Here, the Type I error rate can be very different from the intended 0.05.

assumptions: the population distributions are normally distributed and the population standard deviations are the same. The mathematical theorems about Type I error rates no longer hold if the population distributions are non-normal, but in practice it matters only a little bit. For example, the distribution for population 1 in Figure 3a is strongly skewed, while the distribution for population 2 is a normal distribution; but both distributions have the same mean value (0), so a test of population means is for a true null hypothesis. Out of 10,000 simulated *t*-tests based on samples drawn from these distributions, the Type I error rate for the standard *t*-test is 0.051, which is just a bit above the intended 0.05. (See the Appendix for access to the simulation code.) In general, as long as the population distributions are unimodal and close to a normal distribution, the Type I error rate will be close to the intended value.

As long as the samples drawn from each population are of equal sizes, the *t*-test is also quite robust when the population standard deviations are different. In Figure 3b, population 1 has a standard deviation of 5, while population 2 has a standard deviation of 1. From 10,000 simulated *t*-tests with equal sample sizes (n_1=n_2=25), the Type I error rate is 0.059, which is only somewhat bigger than the intended 0.05.

In contrast to these situations, unequal standard deviations coupled with unequal sample sizes can be a disaster. If a large sample size (n_1=25 scores) is combined with the large standard deviation for population 1 and a smaller

sample size (n_2=5 scores) is combined with the small standard deviation for population 2, then the Type I error rate is around zero (none of the 10,000 simulated *t*-tests rejects the null hypothesis). On the other hand, if the larger sample size is paired with the smaller standard deviation, then the Type I error rate is around 0.38, even when the 0.05 criterion is used to decide statistical significance.

The good news is that there is an easy solution to this problem. Welch's test is an alternative to the *t*-test that maintains the desired Type I error rate even when unequal standard deviations are paired with unequal sample sizes. In the cases presented earlier, Welch's test produces Type I error rates of 0.046 and 0.05, respectively. Welch's test is not perfect; for example, if the population standard deviations are equal but the sample sizes are different, a Type I error rate of around 0.06 is produced. Nevertheless, it avoids the really egregious cases that can occur for the standard *t*-test.

Take away message: The *t*-test is quite robust to deviations from some of its assumptions; but if you have unequal sample sizes, you should use Welch's test rather than a standard *t*-test.

4 Adding Data Increases the Type I Error Rate: Optional Stopping

A not uncommon situation is that after gathering some initial data, your analysis produces a promising but not significant result (e.g., p=0.08). Some people describe such a result as a "marginal effect" and move on, but that feels unsatisfying since the whole point of your experiment was to test for the effect (and it is not clear what "marginal" means anyhow). What some scientists do is add more subjects to the data set and rerun the analysis. That approach is problematic because when you make a final decision, you have given yourself two chances to reject the null hypothesis. Since the first decision (assuming everything else is appropriate) had a 5% chance of making a Type I error, the second decision inflates the error rate. The amount of increase in Type I error depends on a variety of factors (notably the sizes of the first and added samples). Moreover, suppose after adding some subjects to the original data set, your analysis produces p=0.07. You face the same issue and may decide to add still more subjects to the data set. If you are willing to keep adding subjects, the probability of making a Type I error approaches 1.0!

The problem is actually worse than it seems because Type I error control in hypothesis testing is not a property of any individual test. Rather, it is a property of the *procedure* you use to make a final decision about whether an effect exists (e.g., your result is statistically significant). If your procedure has many decision points (e.g., you will add subjects before making your final decision if p=0.08,

but not if p=0.3), then you have to consider all those decision points, whether or not you actually follow them in a given situation. Thus, if your first data set produces p=0.02 and you report a significant result as your final decision, then your Type I error rate may be much higher than your intended 0.05. The Type I error rate has to consider what you *would have done* with results different from what you observed. Thus, if you would have added subjects had the p-value been larger, then that fact has to be included when considering the Type I error rate of your procedure.

The more principled way of describing the problem is to flip it around and describe it as *optional stopping*. It is not the adding of subjects that is truly problematic; rather the problem comes from stopping data collection when you are satisfied with the outcome. What is the absolute upper limit of resources (e.g., sample size) you would commit to a study? In many situations, scientists pick a sample size to "start," but they know that they will run more subjects if necessary. Having possible stopping points along the way up to that absolute upper limit sample size must inflate the Type I error rate. Oftentimes, scientists do not know their absolute upper limit sample size, nor (until faced with the choice) do they know what they would do if they found p=0.07 on their third analysis check. Such scientists cannot know the Type I error rate for their hypothesis-testing procedure.

What to do? There are sequential sampling methods that let you specify stopping points in advance and still maintain a desired Type I error rate. A simple approach is called the composite open adaptive sequential test (COAST; Frick, 1998). Here you gather an initial data set and run a t-test. If the p-value is below 0.01, you stop and conclude that you found a significant result. If the p-value is above 0.36, you stop and conclude that you did not find a significant result. Otherwise, you add another score and repeat. This procedure has a Type I error rate of 0.05, and it tends to use fewer subjects than a traditional t-test where sample size has been identified by a power analysis. There are costs, of course; you cannot decide whether or not to use COAST *after* looking at your data set. For example, if your first sample produces 0.02, you cannot claim significance; instead, the COAST procedure requires you to keep adding subjects. Moreover, for a given sample size, sequential sampling approaches have (somewhat) lower power than the traditional t-test. Finally, COAST does not have an upper limit on the sample size. As a result, if data collection stops with a p-value between 0.01 and 0.36, the scientist would not conclude evidence for an effect, and so COAST has a Type I error rate a bit below the intended 0.05. Other sequential sampling approaches allow for upper limits on the sample size, but you must have the resources to generate such sample sizes, even though you are unlikely to use them. (You cannot say

that you will run up to 250 subjects if you only have enough money to pay for 75 subjects.)

You might say that the solution to optional stopping is obvious: pick a sample size in advance and stick to it. That can work in some situations, but then what do you do when you get p=0.08? If you run another experiment with entirely new data, then you inflate the Type I error rate by having multiple chances to reject the null. Meta-analysis (pooling data across experiments) does not help either because it is just a variation of optional stopping; you would not have run the follow-up studies if the original study were sufficiently convincing (Ueno, Fastrich, & Murayama, 2016). Even worse, although you might have a fixed sample size in mind for your study, someone else might have a different maximum sample size in mind and use your study as a starting point for further investigation. These different analyses would have different procedures and therefore different Type I error rates, even if they reported the same results for the same samples.

Taken to an extreme, the fixed sample size requirement for hypothesis testing seems to suggest that each experiment can only be run once, that you have to specify the sample size in advance, and then you (and everyone else) have to accept the decision of that experiment. That extreme view seems rather ludicrous, but if you relax the fixed sample size requirement of hypothesis testing, then you lose control of the Type I error rate, which is the whole point of hypothesis testing. In some sense, this view emphasizes that science cannot be too closely tied to statistical analyses. Statistical analysis is a means of double-checking scientific reasoning, but it cannot do the reasoning itself.

4.1 An Example from Perception

Optional stopping causes problems in addition to an inflation of the Type I error rate. Consider the Muller–Lyer experiment that produced the results in Figure 2, but suppose that your research interest was the correlation across subjects of matching lengths for inward and outward wings. Using the entire data set (n=310), this correlation is r=0.52, which is significant (t_{308}=10.7, p<0.001). If instead of gathering all the data and then analyzing, you analyzed data from just the first 30 participants and then added data one participant at a time until finding a significant result, then you would stop after getting data from n=53 subjects, when (for this data set) the correlation is r=0.3, which just satisfies the significance criterion (t_{51}=2.25, p=0.03). Analyses with earlier data sets do not produce significant results. For example, with the first n=52 subjects, the correlation is r=0.24, which corresponds to t_{50}=1.78, p=0.08. Generally speaking, optional stopping tends to produce results that just satisfy the significance

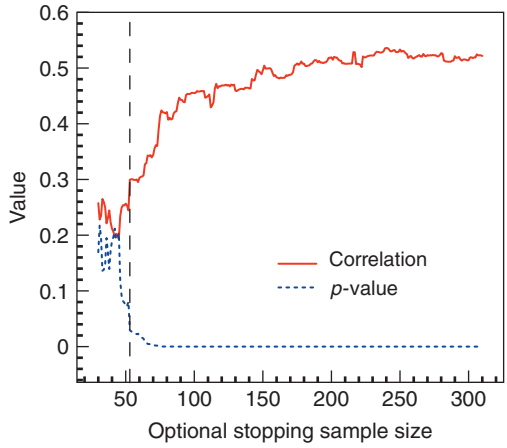

Figure 4 Using optional stopping for the Muller–Lyer data set from Figure 2 dramatically underestimates the correlation. The vertical line indicates the first time the updated sample produced a *p*-value less than the 0.05 criterion. Here, the sample correlation is small compared to what it would be with the full data set.

criterion. This means that the estimated effect size can overestimate or underestimate the true effect size, depending on its magnitude in the initial sample. For example, if the estimated effect happens to be small for the initial sample, then you rarely find strong effects because data collection stops before a strong result appears. Figure 4 demonstrates this property by plotting the sample *r* and *p* values generated by an optional stopping approach for the Muller–Lyer data set in Figure 2. The early samples happen to underestimate the correlation, and significance is found before the correlation is pulled toward the value calculated from the entire data set.

Take away message: Unless you are in a situation where you can fix the sample size, hypothesis testing does not necessarily do a good job controlling the Type I error rate. Unfortunately, it is difficult to avoid optional stopping.

5 ANOVA Can Be Extremely Conservative

Undergraduate statistics classes often introduce analysis of variance (ANOVA) as a way to resolve the multiple testing problem. If you have multiple tests (for example, to compare means against one another), then each test has a risk of making a Type I error and that risk accumulates, so that the probability of making at least one Type I error from the multiple comparisons is much larger than the intended 0.05 (or whatever rate you choose). ANOVA cleverly solves

this problem by testing an omnibus null hypothesis (all means equal one another).

The cost of using an omnibus null hypothesis is that it does not indicate which means differ from other means. Thus, a significant ANOVA is usually followed up with additional tests to compare means (or groups of means) against one another. These additional tests have the feel of being the "dark arts" of hypothesis testing because they all seem a bit ad hoc. In many cases, these methods err on the side of being extra conservative.

For example, consider a situation where a scientist is testing search times for four visual maps. The scientist wants to compare her preferred map design to the other three designs. To convince other scientists that her design is better, she needs to show the following outcomes:

- A significant one-way ANOVA, which indicates that there is some difference among the map designs.
- A significant contrast of design 1 compared to design 2.
- A significant contrast of design 1 compared to design 3.
- A significant contrast of design 1 compared to design 4.

The three contrasts are necessary to conclude that her preferred design is better than each of the other designs. What is the Type I error rate for concluding that her preferred design is better than the other designs? Each hypothesis test has a Type I error rate of 0.05. But if all of the nulls are true and there really is no difference between any of the map designs, the Type I error rate of all four tests is around 0.003. It should intuitively make sense that requiring three significant contrasts *in addition to* a significant ANOVA has to reduce the Type I error rate. The reader can verify these calculations and create variations using the online ANOVA power calculator in Francis (2018) by setting up four levels, entering zero for each mean, and creating three appropriate contrasts.[1] Since all the population means are equal, the computed power for all tests will correspond to the Type I error rate.

Thus, if a scientist has specific comparisons in mind for drawing her conclusions, following standard analysis approaches may be setting up enormous statistical hurdles. Simulation studies using the power calculator in Francis (2018) show that a Type I error rate of just under 0.05 is generated across the full set of four tests if you set the significance criterion to be α=0.3 for each test. With such a criterion, each test has a fairly high risk of making a Type I error, but it is rather unlikely that *all* the tests will make a Type I error.

[1] https://introstatsonline.com/chapters/calculators/OneWayANOVAPower.shtml

However, you would not want to use this inflated criterion because it would generate a high Type I error for other cases. Suppose maps 1 and 2 have identical population mean reaction times, so a significant difference will be a Type I error. Further suppose that maps 3 and 4 have very different means than those for maps 1 and 2. Thus, the tests that compare map 1 to map 3 and to map 4 cannot make Type I errors. Since your conclusion that map 1 is best requires all tests to be significant, the probability of making a Type I error (because map 1 is not actually better than map 2) is the probability of getting significant outcomes for all four tests (ANOVA and three contrasts). Suppose that any comparisons involving map 1 and map 3 or map 4 have high power, then the overall conclusion hinges on the test comparing map 1 and map 2. The Type I error for the overall conclusion is thus the α-value used for that test. Scientists typically want to consider the worst-case scenario for Type I error control, so the safe approach is to have each test use the desired α-value. There is a cost for this safety, though; as we will see in Section 7, an experiment that requires multiple significant outcomes may have low power.

5.1 An Example from Perception

The Muller–Lyer illusion experiment described in Figure 2 also included a No wings condition. When comparing the conditions, a scientist might want to show that compared to the No wings condition, the Outward wings condition produces smaller matching line lengths (thus it is perceived to be larger than it really is) and that the Inward wings condition produces larger matching line lengths (thus it is perceived to be smaller than it really is). To convincingly demonstrate these effects, the scientist would need the following:

- A significant one-way ANOVA, which indicates that there is some difference among the conditions.
- A significant contrast where the Outward wings match length is smaller than the No wings match length.
- A significant contrast where the Inward wings match length is larger than the No wings match length.
- A significant contrast where the Inward wings match length is larger than the Outward wings match length.

The three contrasts are necessary to conclude that the illusion is found; should any of these tests not produce a significant result, the experiment would not be interpreted as entirely consistent with the presence of the Muller–Lyer illusion. Suppose there is no illusion effect at all. What is the Type I error rate for concluding that the illusion exists? Figure 5 shows how to set up an online

Enter the Type I error rate, α = 0.05

Enter the population standard deviation, σ = 8

Enter the population correlation between levels, ρ = 0.52

How many levels (groups) do you have in your ANOVA? K = 3

Number of iterations
(bigger values produce better estimates, but take longer) 50000

Level Name	Population Mean
OutwardWing	100
InwardWings	100
NoWings	100

Add a contrast test

Specify hypotheses for Contrast1

H_0: 1 $\mu_{OutwardWings}$ + 0 $\mu_{InwardWings}$ + −1 $\mu_{NoWings}$ = 0

H_a : Negative one-tail

α 0.05

Specify hypotheses for Contrast2

H_0: 0 $\mu_{OutwardWings}$ + 1 $\mu_{InwardWings}$ + −1 $\mu_{NoWings}$ = 0

H_a: Positive one-tail

α 0.05

Specify hypotheses for Contrast3

H_0: −1 $\mu_{OutwardWings}$ + 1 $\mu_{InwardWings}$ + 0 $\mu_{NoWings}$ = 0

H_a: Positive one-tail

α 0.05

Power for all tests = 0.00008 Calculate minimum sample size

Sample size n = 310 Calculate power

Test	Estimated Power
ANOVA	0.0499
Contrast1	0.04982
Contrast2	0.0487
Contrast3	0.05086

Figure 5 Using an online power calculator to estimate the Type I error rate when an ANOVA and three contrasts must all be significant. The Type I error for all four tests being significant is 0.00008

calculator using values similar to those in the data set.[2] You can easily verify that the values for the standard deviation, correlation, (equal valued) means, and sample size do not matter. Instead, the Type I error rate is determined by the number and type of hypothesis tests. Clicking on the *Calculate power* button runs a simulation of 50,000 experiments that generates data sets sampled from populations having the specified properties. Each data set is then subjected to the dependent ANOVA and three (one-tailed) contrast tests. Since the null hypothesis is true in these simulations, significance for any test is a Type I error. The table at the bottom of Figure 5 indicates that, as expected, each hypothesis test has a Type I error rate of approximately 0.05. However, the probability of all four tests being significant is indicated in the *Power for all tests* text field; and the Type I error rate of all four tests is 0.00008.

Take away message: If your conclusion requires multiple test outcomes, then the Type I error rate for your conclusion might be much smaller than the criterion you set for any individual test.

6 ANOVA Handles Only One Type of Multiple Testing Problem

Note the critical difference between the previous section and the traditional multiple testing problem. In the previous section, multiple tests must be significant to provide support for the scientist's conclusion. These multiple tests tend to reduce the Type I error rate because multiple outcomes must be simultaneously satisfied. In the traditional multiple testing problem, the concern is that you might find *at least one* Type I error from multiple tests. These multiple tests tend to increase the Type I error rate because there are multiple ways to make a Type I error. The latter is a concern about controlling the Type I error rate for exploratory studies, where you do not have specific outcomes in mind.

Scientists want to control Type I error for the worst-case scenario. In an exploratory study, the worst-case scenario is that all the null hypotheses are true but one or more of the tests may produce a significant result. Sometimes this scenario can be handled by analyzing data with ANOVA (to test whether one condition is different from other conditions). Unfortunately, ANOVA only does a good job controlling the Type I error rate for exploratory investigations using a one-way ANOVA. Multi-way ANOVA introduces a new kind of multiple testing problem (Cramer et al., 2016).

Suppose you are doing exploratory work and you use a 2×2 ANOVA to search for significant results. You set α=0.05 to control the Type I error rate, but your ANOVA has three tests: a main effect for factor 1, a main effect for factor

[2] https://introstatsonline.com/chapters/calculators/OneWayANOVADependentPower.shtml

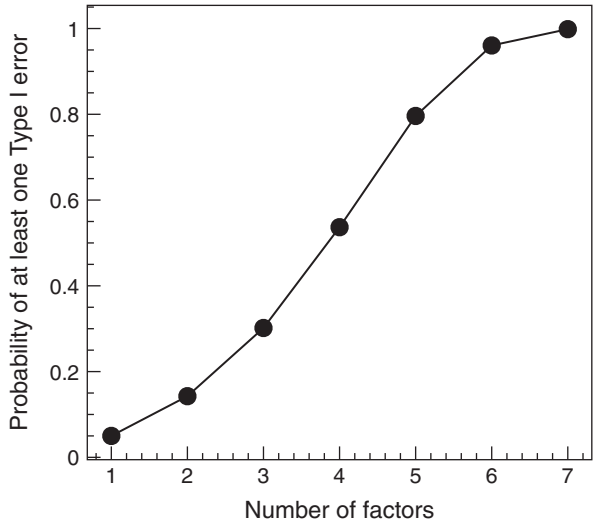

Figure 6 The probability of a multi-way ANOVA producing at least one Type I error as a function of the number of factors.

2, and an interaction between factor 1 and factor 2. If there is actually no difference between populations, then each test has 0.05 probability of making a Type I error, but the probability of *at least one* of the tests making a Type I error is 0.14. See the Appendix for reference to code that estimates this probability from simulated experiments.

The multiple testing problem gets worse with additional factors in the ANOVA. A 2×2×2 ANOVA has seven tests (three main effects, three two-way interactions, one three-way interaction) and a Type I error rate of 0.3. An ANOVA with six factors (they exist in our journals!) has 63 tests and a Type I error rate of 0.96. Any ANOVA with more than six factors is almost guaranteed to produce at least one Type I error among the main effects and various interactions (see Figure 6). It does not matter how many levels are within each factor.

You can apply various methods such as Bonferroni to reduce the α criterion value and thereby reign in the Type I error rate, but then it becomes less likely that you will find any significant results, even if there are population differences. Perhaps the fundamental point is that ANOVA should not be used for exploratory investigations; certainly exploratory results should not be treated the same as confirmatory results.

6.1 An Example from Perception

The Muller–Lyer illusion experiment described in Figure 2 also had conditions where the match line was oriented vertically (the comparison line

was always horizontal). Without wings, vertical lines are perceived to be longer than horizontal lines having the same physical length; this phenomenon is known as the horizontal-vertical illusion. Thus, when observers adjust the vertical matching line to the horizontal comparison line, they will (on average) set the vertical line to be physically smaller than the horizontal line. If the data were analyzed with a 2 (wing characteristics) × 2 (orientation) ANOVA with α=0.05, the Type I error rate of finding at least one significant effect is 0.14.

Take away message: Multi-way ANOVA designs generally do a poor job controlling the Type I error rate for exploratory investigations that look for some significant results out of many tests.

7 Power Analyses Should Consider All Relevant Tests

Many journals now insist that researchers report a power analysis that demonstrates their study has a sufficiently high probability of producing a significant result, if there actually is an effect. For simple cases, such as a two-sample *t*-test, power calculations are straightforward if one has a specific alternative hypothesis. Unfortunately, for more complicated designs, such as a one-way ANOVA with several contrasts, scientists rarely consider the full set of hypotheses; this oversight can lead to dramatically underpowered studies.

Consider a situation where a researcher plans to test a preferred visual map design against three other designs by using an independent ANOVA and three contrasts that compare map 1 to each of the other maps. Suppose the researcher has a good guess as to the population means and standard deviation (search times in seconds) for each map: μ_1=2.4, μ_2=2.6, μ_3=2.8, μ_4=3.0, and σ=0.5. A program such as G*Power (Faul, Erdfelder, Lang, & Bucher, 2007) will convert the means and standard deviation to a standardized effect size called Cohen's *f*=0.447 (this is the standard deviation of the population means, using the "population formula" divided by the population standard deviation). If the researcher plans for power of 0.9 with α=0.05, G*Power will indicate that the experiment requires a sample size of 76 subjects (19 in each sample for the different maps).

A researcher using this sample size will likely be sorely disappointed. If there really are the planned-for differences in population means, her study does have a 0.9 probability of producing a significant ANOVA; but the study has only a 0.21 probability of producing the full set of desired outcomes. In particular, the contrast to test between μ_1 and μ_2 has only around 0.23 power when the sample sizes are 19 in each group. To have 0.9 power for the entire set of tests requires a sample size of at least 131 in each group (a total of 524 subjects).

Obviously, there are enormous differences between studies requiring 76 subjects and 524 subjects. A power analysis that does not consider all the relevant tests can lead to extremely poor experimental designs. These calculations and variations thereof can be reproduced in the Independent ANOVA power calculator in Francis (2018). You can, for example, consider designs with unequal sample sizes and discover that an experiment with n_1=n_2=150 and n_3=n_4=100 also produces a 0.9 probability of all tests being significant.

Researchers sometimes do not perform a power analysis because they think that they do not know how to estimate the effects of interest. Oftentimes, such researchers are not giving themselves enough credit. If forced to do the analysis, they would discover that their experience or a perusal of existing literature does give some guidance. A power analysis does not need to be especially precise; even sloppy estimates can help identify appropriate sample sizes.

7.1 An Example from Perception

Suppose you want to repeat the experiment of Figure 2 to explore the Muller–Lyer and horizontal-vertical illusions with a new sample (maybe a patient population that possibly does not experience the illusions) and a more complicated set of analyses. Table 1 lists the various tests that might be used in this type of study. It might seem excessive to run 13 hypothesis tests to demonstrate the Muller–Lyer and horizontal-vertical illusions, but actually such combinations of tests are quite common in academic papers, and the reasoning is actually pretty sound. With an ANOVA across the full data set, you want to show a main effect of orientation (the horizontal-vertical illusion) and a main effect of wings (the Muller–Lyer illusion). Without these significant main effects, your experiment would not be consistent with the illusions. However, just those main effects are not enough. You would also want to show a main effect of wings for each orientation. Moreover, for each orientation you would run three contrasts to compare each combination of wing types. All of these contrasts need to produce significant results to demonstrate the Muller–Lyer illusion for each orientation. Finally, you need to show an effect of orientation for each of the three wing types. To compute the power of your study to show significant results for all of these tests, you need to specify the expected population values if the illusion exists and the details of your design. A reasonable source for expected values might be the sample statistics from the original study. Table 2 lists the statistics that are needed to do the analysis.

To compute power, the values in Table 2 were used to define a population. An R computer program (see the Appendix for how to download the code) sampled values from the population for a given sample size and then ran the various tests

Table 1 Various tests to analyze a study of the combined Muller–Lyer and horizontal-vertical illusions.

Test	**H_a**
2×3 ANOVA on full data set	
Main effect of orientation	$\mu_V \neq \mu_H$
Main effect of wings	$\mu_O \neq \mu_I$, $\mu_O \neq \mu_N$, or $\mu_I \neq \mu_N$
One-way ANOVA on horizontal orientations	
Main effect of wings	$\mu_{HO} \neq \mu_{HI}$, $\mu_{HO} \neq \mu_{HN}$, or $\mu_{HI} \neq \mu_{HN}$
Contrast for outward vs. none wings	$\mu_{HO} \neq \mu_{HN}$
Contrast for inward vs. none wings	$\mu_{HI} \neq \mu_{HN}$
Contrast for inward vs. outward wings	$\mu_{HI} \neq \mu_{HO}$
One-way ANOVA on vertical orientations	
Main effect of wings	$\mu_{VO} \neq \mu_{VI}$, $\mu_{VO} \neq \mu_{VN}$, or $\mu_{VI} \neq \mu_{VN}$
Contrast for outward vs. none wings	$\mu_{VO} \neq \mu_{VN}$
Contrast for inward vs. none wings	$\mu_{VI} \neq \mu_{VN}$
Contrast for inward vs. outward wings	$\mu_{VI} \neq \mu_{VO}$
One-way ANOVA on outward wings	
Main effect of orientation	$\mu_{HO} \neq \mu_{VO}$
One-way ANOVA on inward wings	
Main effect of orientation	$\mu_{HI} \neq \mu_{VI}$
One-way ANOVA on none wings	
Main effect of orientation	$\mu_{HN} \neq \mu_{VN}$

Table 2 Sample statistics from the Muller–Lyer experiment that are needed to perform a power analysis.

Statistic	**Value**
Mean for horizontal, wings in	112.3
Mean for horizontal, wings out	91.5
Mean for horizontal, no wings	101.8
Mean for vertical, wings in	104.3
Mean for vertical, wings out	83.4
Mean for vertical, no wings	92.5
Standard deviation (all conditions)	8.0
Correlation (all conditions)	0.5

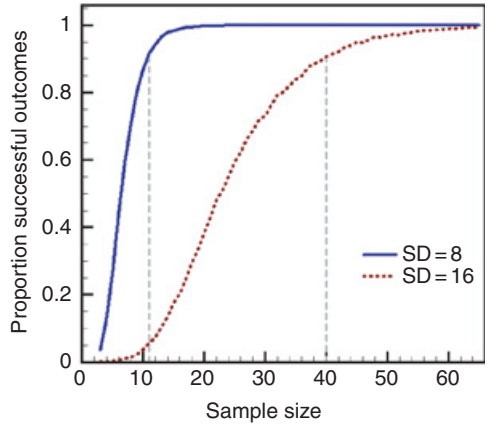

Figure 7 The proportion of successful outcomes from simulated experiments for a replication of the line length illusion experiment in Figure 2. The solid curve supposes that the sample statistics from the data in Figure 2 are representative of the population. The dashed curve is similar but supposes that the standard deviation across participants is twice the original value. The vertical gray lines indicate the sample size for each simulation that is needed to produce a 0.9 success rate.

to see if they all produced significant outcomes. This procedure was repeated 5000 times, and the solid curve in Figure 7 plots the proportion of successes (all 13 hypothesis tests produced significant results) against the sample size. The curve indicates that a sample of size n=11 is sufficient to produce a 0.9 success rate. A small sample is sufficient here because the differences between means are quite large relative to the standard deviation (see Table 2).

Naturally, the power analysis is only as good as the estimates provided in Table 2. In particular, if your planned study investigates a different population than the original study, you may want to consider that the means, standard deviations, and correlations might be different. For example, the data set in Figure 2 is based on a population of undergraduates at a large Midwestern university. A more diverse population with a broader range of ages might have a larger standard deviation. Likewise, a more specialized population, such as patients taking a certain kind of medication, may have quite different statistics. These differences can have important implications for the design of your experiment. Suppose that your new population has a standard deviation in line-length judgments of σ=16, which is twice the value listed in Table 2. Rerunning the simulated experiments with this new standard deviation produces the dashed line in Figure 7. This curve suggests that a sample of size n=40 is required to have a 0.9 success rate for all the hypothesis tests.

Take away message: Power analyses and sample size determinations need to consider the full set of tests that are relevant to the conclusions. Not including the full set can lead to dramatically underpowered experiments.

8 The Only *p*-value You Can Plan for Is Zero

The proper interpretation of the *p*-value in hypothesis testing is rather confusing, so many people have a kind of intuition about how it is produced and what it means. Unfortunately, this intuition can lead to inappropriate ideas about how to design experiments. Intuition correctly points out that (all else equal) larger samples tend to produce smaller *p*-values. However, some people interpret this observation as indicating that a researcher should pick a sample size large enough to produce a significant *p*-value but not waste resources by using an excessively large sample size, which will generate an unnecessarily tiny *p*-value. Indeed, some scientists seem to have a knack for picking just the right sample size, and they report multiple studies with *p*-values just below the α=0.05 criterion. Rather than indicating experimental skills, such results should be a warning that something has gone wrong.

Figure 8 plots how *p*-values are distributed between 0 and 1 for a two-sample, two-tailed, *t*-test based on sample sizes of n_1=n_2=50 with five standardized effect sizes (Cohen's δ). To better show the curves, the *y*-axis is on a log scale. (The particular numbers on the *y*-axis reflect that probability was calculated for

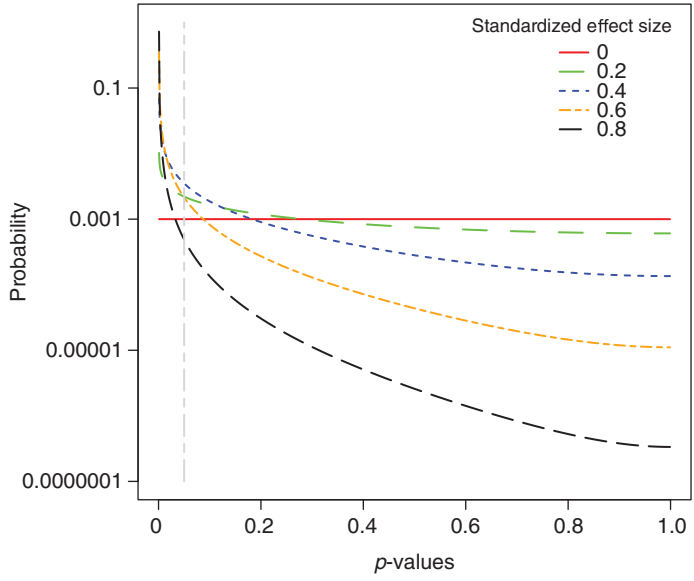

Figure 8 The distribution of *p*-values for different standardized effect sizes.

1000 bins of the *x*-axis; different bins would produce different numbers, so the relative order of the curves is what matters.) When the standardized effect size is zero (no effect), the distribution of *p*-values is uniform. This may seem surprising, but it directly follows from the way *p*-values relate to Type I error. If the null hypothesis is true, then 5% of the *p*-values will be more extreme than the α=0.05 significance criterion. Likewise, 6% of the *p*-values will be more extreme than the α=0.06 criterion. More generally, the proportion of *p*-values more extreme than criterion α is α itself, which happens only if the distribution of *p*-values follows a uniform distribution.

As the effect size increases, large *p*-values become less common and *p*-values closer to zero become more common. The vertical gray line indicates the traditional criterion for significance, α=0.05. The area under each curve to the left of this line is power for the respective experiment (or Type I error rate, when the effect size is zero). The areas are 0.05, 0.17, 0.51, 0.84, and 0.98, respectively, as the effect size increases.

A key feature of these curves is that except when the effect size is zero, more *p*-values are close to zero than any other value. When effects are non-zero, tiny *p*-values are more common than big *p*-values, and this holds true even in the significance range (0, 0.05). Which *p*-value you actually get in any given experiment is due to random sampling, but if you have high power, then you will almost never get *p*-values close to the significance criterion. These properties of *p*-values form the basis of meta-analytical approaches such as *p*-curve (Simonsohn, Nelson, & Simmons, 2014) and *p*-uniform (van Assen, van Aert, & Wicherts, 2015).

When you plan an experiment, you do not know if you are in a high-power or a low-power situation because you do not know the population standardized effect size. However, if your experimental conclusions consistently rely on *p*-values a bit below the criterion (say, 0.02–0.05), then you are probably doing something wrong: you are ignoring relevant *p*-values that are above the criterion (publication bias), you are stopping data collection when the *p*-value drops below the criterion (optional stopping), you are reporting weak statistical tests instead of more appropriate strong tests, or you are drawing conclusions from exploratory work (filtering effects by the *p*-value). The latter situation seems to be very common, but because of the multiple testing problem, it can produce findings based on high Type I error rates.

Take away message: Experiments measuring real effects tend to produce very small *p*-values. If you consistently find *p*-values just below the significance criterion, then you are probably doing something wrong. When planning an experiment, you can only pick sample sizes that likely produce very small

p-values (e.g., high power); you cannot plan for a *p*-value to be just below the significance criterion.

9 Subjects and Trials Do Not Trade off Evenly

The Muller–Lyer and horizontal-vertical illusion data set in Figure 2 used n=310 subjects who each ran m=8 trials for each of the six conditions (three wing types and two orientations). When analyzing the data, we computed a single mean score for each subject for each condition, so each observer provides six (correlated) scores for the analysis. We could have analyzed the data differently by considering trial as another factor to include in our analysis. However, studies of perception are usually not interested in the trial-to-trial variability and instead only care about the average performance of each subject, which tends to cancel out the trial-to-trial variability. Implicitly, we can think of each subject *j* at trial *i* having a score for each condition determined by

$$X_{ij}=\mu + \alpha_j + \epsilon_{ij}$$

where μ is the overall grand population mean across subjects, α_j is the deviation from the grand mean for subject *j*, and ϵ_{ij} is noise for a particular trial for a particular subject.

The grand mean, μ, is an unknown fixed value. Subject-based deviation from the grand mean, α_j, varies across subjects who might be randomly selected for the experiment. Assume the α_j values in the subject population are distributed normally as $N(0, \sigma_\alpha)$. The ϵ_{ij} term describes variability within a subject across trials. We assume that it also follows a normal distribution, $N(0, \sigma_\epsilon)$, which has a different standard deviation than for variability across subjects.

If we follow common practice within studies of perception, then the mean of a subject's trials for a given condition becomes an individual score, Y_j. If we are just interested in a single condition, the set of Y_j values are then fed into, say, a one-sample *t*-test. The variability of a given Y_j value is derived from the standard error of the mean. Moreover, since variances add, the variability across subjects will be

$$\sigma^2 = \sigma_\alpha^2 + \frac{\sigma_\epsilon^2}{m}$$

Larger variability within a participant across trials, σ_ϵ, leads to more variability in mean scores across participants, but the contribution is divided by the number of trials, m, so it may not have a big effect. When computing the standard error of the mean across participants for the *t*-test, one uses the number of subject scores, so

$$\sigma_{\overline{X}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{\sigma_\alpha^2}{n} + \frac{\sigma_\epsilon^2}{nm}}$$

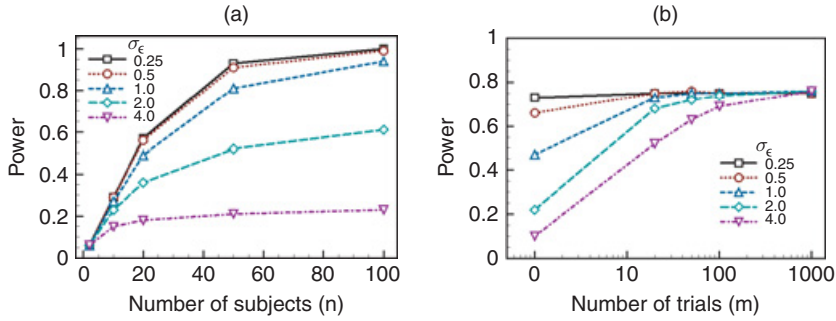

Figure 9 Simulation results showing how the number of subjects or the number of trials affects power. In (a), the product of the number of subjects and number of trials is fixed at 100. So, an increase in the number of subjects also causes a decrease in the number of trials for each subject. Nevertheless, power increases with the number of subjects. This is true for every value of within-subject variability, σ_ϵ. In (b), for a fixed sample size, n=30, power increases with more trials. However, the effects are small unless σ_ϵ, the variability within subjects, is large relative to $\sigma_\alpha = 1$, the variability between subjects.

Notice that the σ_ϵ term is divided by nm so it contributes less variability than a corresponding σ_α term. In the standard t-test analysis, any observed variability is assumed to come from σ_α, which is not necessarily correct but can often be close to accurate.

Suppose an experiment is being run that can gather $m \times n$=100 trials. How should the experimenter distribute trials across subjects to best test the mean? There may be practical issues that limit how many trials each subject can produce and how many subjects are available; however, from a statistical perspective, it is almost always best to maximize the number of subjects (e.g., n=100 and m=1). The following discussion provides an intuitive feel for the issues by using simulated experiments (see the Appendix for how to access the simulation code); for a more detailed discussion see Rouder and Haaf (2018).

Figure 9a shows the result of simulated experiments that set (n, m) to be (100, 1), (50, 2), (20, 5), (10, 10), or (2, 50); thus $m \times n$=100 for every condition. For each subject, $\mu = 0.5$ and the simulation generated a value α_j from a standard normal distribution ($\sigma_\alpha = 1$). For each subject, the simulation generated m trials by drawing random values of ϵ_{ij} from a normal distribution with a mean of zero and a standard deviation of σ_ϵ equal to 0.25, 0.5, 1, 2, or 4 (these different conditions correspond to the lines in Figure 9a). By using the previously presented equation for X_{ij}, this process defines individual trial scores.

The mean subject scores for each simulated sample were analyzed using a one-sample t-test with H_0: $\mu = 0$. Each point in Figure 9a shows the proportion of 10,000 simulated experiments that rejected the null (estimated power). There are two notable effects. First, as σ_ϵ increases, power decreases. This makes sense, because increased variability across trials generally produces a larger standard error of the mean and thus a smaller t statistic. Second, for a fixed value of σ_ϵ, the maximum power is always for n=100, m=1, which will minimize the standard error of the mean. This makes sense because when computing the standard error of the mean, the σ_ϵ term is divided by nm, which always equals 100 trials for these simulated experiments. Thus, to minimize standard error, a researcher should increase n at the expense of decreasing m. An additional advantage of increasing n is that the t-test uses degrees of freedom $n-1$, so a larger n reduces the t critical value needed to determine statistical significance.

So it is not appropriate to just trade off subjects for trials. Everything else equal, it is always best to maximize the number of subjects. However, subjects are often more difficult to acquire than trials, and for a fixed number of subjects there is an advantage to gathering more trials. Figure 9b shows the proportion of 10,000 new simulated experiments that reject the null. Each experiment used n=30 subjects and varied the number of trials, m, and the value of σ_ϵ (as before, $\sigma_\alpha = 1$ and $\mu = 0.5$). In general, larger m values give rise to higher estimated power. The advantage is largest when σ_ϵ is big, which makes sense given the contribution of σ_ϵ to the standard error of the mean. Increasing the number of trials for a fixed number of subjects means that σ_ϵ is divided by a larger denominator and contributes a smaller amount of variability to the standard error.

Note that even with $\sigma_\epsilon = 2$, which is twice the size of σ_α, an increase of trials from m=50 to m=1000 only leads to a power increase of 0.02. So increasing the number of trials by a factor of 20 hardly benefits the researcher but greatly bothers the subjects. There are situations where increasing the number of trials has large effects. For $\sigma_\epsilon = 1$, increasing m from 1 to 20 produces a 0.26 increase in power. Generally, unless σ_ϵ is much larger than σ_α, there is little gain in gathering more than 20–50 trials per subject. Rouder and Haaf (2018) argue that for many situations in cognition and perception, σ_ϵ is much larger than σ_α, and thus it is often worthwhile to run a large number of trials.

This discussion supposes that scientists are interested in the mean behavior across a population of people. However, a good argument can be made, especially for studies of perception where nearly every subject shows the same pattern of behaviors, that psychologists should primarily focus on individual behaviors (Smith & Little, 2018). When the focus is on measuring the properties of an individual, more trials are always beneficial because n=1.

9.1 An Example from Perception

Suppose you wanted to design a new experiment that explores the Muller–Lyer and horizontal-vertical illusions (Figure 2) for lines 150 pixels long (compared to the original experiment, where lines were 100 pixels long). The subject pool at your university provides a single course credit for a student who participates in a 30-minute experiment. Taking into account 5 minutes needed for instructions, handling consent forms, and debriefing, and taking into account another 5 minutes for breaks, you estimate that participants in your experiment have about 20 minutes of experiment time, which is sufficient to go through roughly 100 trials. For each line length, there are six conditions (wings: outward, inward, none and orientation: vertical, horizontal). Thus, you could run approximately 16 trials for each condition to have a total of 96 trials during a 30-minute experimental session.

You want to use 16 credits for this experiment, and you can thus run either n=16 participants in a 30-minute experiment (each participant earns 1 credit), or you can run n=8 participants in a 60-minute experiment with more trials (each participant earns 2 credits). Since the start-up and debriefing time is constant for both a short and a long experiment, participants in the long experiment have roughly 45 minutes of experiment time (the longer experiment does require 10 minutes for breaks), which corresponds to 225 trials. To ensure an equal number of trials for each condition, we round down to 37 trials per condition for a total of 222 trials in the longer experiment.

Which experiment gives more power? From the discussion in this section, we know that an increase in the number of subjects has a bigger impact on power than an increase in the number of trials. However, we do not have an equal trade-off of trials and subjects in these two designs. The 60-minute experiment has a total of 1776 trials across all 8 participants, while the 30-minute experiment has a total of 1536 trials across all 16 participants. To explore which is the better design, we create simulated experimental data (source code is available as described in the Appendix).

From the data in the Muller–Lyer and horizontal-vertical illusion experiment described in Figure 2, we compute the standard deviation of each participant for the trials of each condition and estimate that $\sigma_\epsilon \approx 9$. We previously noted in Table 2 that the standard deviation of means across subjects is approximately 8 for each condition. Using the variance formula given earlier, we can compute (using m=8 for the experiment described in Figure 2) that

$$\sigma_\alpha^2 = \sigma^2 - \frac{\sigma_\epsilon^2}{m} = 8^2 - \frac{9^2}{8} \approx 54$$

so $\sigma_\alpha \approx 7.3$. The simulation generates 10,000 simulated experimental data sets for each experimental design, and for each data set it runs the tests described in Table 1. A simulated experiment is deemed successful only if it produces significant outcomes for all the tests. The short experiment with 16 participants, each running 16 trials for every condition has a predicted success rate of 0.99. In contrast, the long experiment with 8 participants each running 37 trials for every condition has a predicted success rate of 0.78. Thus, the short experiment, with more subjects, is a better design, even though it has fewer trials overall and uses the same number of university credits.

Take away message: It is not true that trials and subjects can be traded off. An increase in the number of subjects has a bigger impact than an increase in the number of trials. In general, you should consider the full characteristics of your data collection and analysis design when planning sample sizes.

10 Replication Is a Poor Way to Control Type I Error

Some fields in psychology are suffering from a replication crisis (Gelman & Loken, 2014; Open Science Collaboration, 2015), whereby new studies do not support previously published experimental findings. In response to this crisis, some people suggest that researchers should run replication studies before reporting findings; in other cases, people express skepticism about a result until it has been successfully replicated. There are good reasons to double-check your result and to have a skeptical attitude; however, those reasons are not related to the statistical properties of hypothesis testing.

When done properly, hypothesis testing is a decision-making process that generates a Type I error (when the null is actually true) with a probability of α (e.g., 0.05). Suppose you are in a situation where hypothesis testing can be done properly, so you have (and want) this kind of Type I error control. You are concerned about the replication crisis, so after finding a result that rejects the null hypothesis ($p<0.05$), you decide to replicate the study to be sure that you have not made a Type I error. If the replication study also rejects the null, you will conclude there is an effect; if the replication fails, you will not conclude an effect exists.

What is the Type I error rate for this decision-making process? If the null hypothesis is true, then each test has a 0.05 probability of rejecting the null. Since the two experiments involve independent samples, the probability that *both* tests reject the null is the product of the probabilities $0.05 \times 0.05 = 0.0025$. If controlling the Type I error is important for making your decision about whether an effect exists, it seems it would be easier to just run one test with $\alpha = 0.0025$.

Moreover, for the same resources (e.g., total sample size) and equivalent Type I error control, running one test is always better than running two tests because the single test is more powerful than requiring two significant outcomes. Consider a specific example of a one-sample *t*-test with null and alternative populations having $\mu_0=0$, $\mu_a=1$, $\sigma=2$. If you run one study with $n=50$ and $\alpha=0.0025$, the power of the test is 0.636. If you take two smaller samples ($n=25$), increase the criterion to $\alpha=0.05$, and require both tests to be significant before concluding there is an effect, then you have used the same resources and had the same Type I error rate as the single experiment. However, the power for each individual test is 0.670, and the probability of both tests being significant is the product of each test's power, which is only $0.670\times0.670=0.449$. At least with regard to power, the replication requirement is much less effective than applying standard hypothesis testing with a smaller criterion (see Schimmack, 2012).

This observation does not mean that replication has no important role to play in scientific investigations, only that the role should not be to control the Type I error rate. There are broadly two situations where replication is useful. First, the original study may have been exploratory rather than confirmatory. As noted in an earlier section, hypothesis testing does not control the Type I error rate for exploratory analyses. Thus, it might be prudent to run a replication study with proper Type I error control. Second, you may want to test the generality of the result with new methods (perhaps very small changes, such as using a different computer or gathering subjects from a different location). Examining generalizability requires that both studies have high (estimated) power; otherwise, it will be difficult to distinguish sampling variability from methodological variability. A mix of the two cases is where you suspect improper reporting, improper analysis, or improper sampling in an original study and you want the replication study to check the accuracy of what was concluded.

Take away message: There may be good reasons to run replication studies, but they do not include avoiding Type I errors.

11 Identifying Improper Methods through Excess Success

The previous section noted that replication is one way to check whether the conclusion of a previous study might be the result of some kind of inappropriate reporting, analysis, or sampling methods. Unfortunately, replication takes a lot of work and sufficient expertise to be able to reproduce the conditions of the original experiment. However, it turns out that many inappropriate methods leave markers that identify their presence. Analysis of these markers can be

done in a few hours and involves much less effort than replicating a study. The main insight here is that multiple studies should succeed at a rate that reflects their power. If studies with low or modest estimated power nevertheless are reported to show a high rate of success, readers should suspect that something has gone wrong in data collection, reporting, analyzing, or interpreting the empirical findings.

A comforting example of this situation is the analyses in Francis (2012) and Schimmack (2012) showing that a set of 10 studies (Bem, 2011) purporting to find evidence that people could see into the future (precognition) seemed "too good to be true." Studies of precognition are rather odd, so it might be good to give a description of Bem's study 1. A computer screen showed two curtains with one curtain occluding an unseen image. The participant's task was to choose the curtain that contained the picture. The act of choosing caused the curtain to be removed, thereby displaying the image if the participant chose correctly. Importantly, the computer program was structured so that the location of the image was not determined (by a random number generator) until after the participant made their response. The key statistical finding was that participants identified the correct curtain with a rate of 0.531 when the image to be displayed was erotic; this rate is significantly above the 0.5 rate that indicates chance guessing. There was also a significant difference in proportion correct for erotic and non-erotic images. Bem (2011) interpreted these findings as evidence that participants had knowledge about the future content of the computer display for some types of images. Other studies reported by Bem explored similar effects of precognition on choosing pictures and on memorability of seen words.

Using the data from the reported studies, the estimated power of the studies ranged from 0.37 to 0.89. Nine of the 10 studies rejected the null hypothesis; however, given the estimated power values of the experiments, such an outcome (or better) should happen with a probability of only 0.06 for samples of similar size as those in the original studies. The estimated success rate is so low that it seems strange that Bem (2011) reported results that are so successful, and the reported excess success engenders skepticism rather than belief in the effects. Better-powered experiments would not face this issue. If each of the 10 experiments had an (estimated) power of 0.9, then the probability of getting 9 (or more) significant results out of 10 experiments is 0.74. There are many good reasons to not believe the findings in Bem (2011), as physics makes a convincing argument that precognition is not possible; so it is good to see that the reported statistics (by rejecting the null too often) are actually consistent with that perspective.

However, the same reasoning applies to investigations that seem much more plausible (Francis, 2014; Francis, Tanzman, & Matthews, 2014). One of my

favorite papers from 2014 was by Firestone and Scholl (2014), who argued that some studies of top-down effects on visual perception were subject to the "El Greco fallacy." The fallacy refers to the observation that the painter El Greco drew figures unusually long and thin; some people speculated that his painting style was due to severe astigmatism. With astigmatism, perceived shapes can be stretched vertically, and it might follow that El Greco simply drew what he saw. This theory is a fallacy because the astigmatism should also apply to the figures in El Greco's paintings; making them seem – to him – even *more* stretched out than what he saw when looking at his models. Indeed, if El Greco painted images to look like how he saw his models, then he would paint accurate figures (both the model and the figure would look stretched out to El Greco, but they would be the same in terms of actual shape).

Firestone and Scholl (2014) point out that several investigations of top-down effects (e.g., action capabilities on perception and morality on lightness perception, which are more fully described below) are subject to the El Greco fallacy: subjects show a top-down influence on perception when they should not. Their implication is that the presence of the reported effect must be due to some other factor because the top-down influence should be involved in both the perception and the matching judgment. I find the argument in Firestone and Scholl (2014) to be very convincing. However, they then proceeded to report five experimental results to demonstrate the presence of top-down effects when they should not occur (the El Greco fallacy). Contrary to what was intended, the empirical studies do not support their claim that many reported top-down effects seem to be "demand characteristics," whereby subjects produce responses that they believe are implicitly requested by the experimenter. Instead, the reported results seem too good to be true, so they do not support any claims at all.

Experiment 1 was a replication of Stefanucci and Geuss (2009) that reported a top-down effect on size judgments. Subjects who held their arms out (by holding a long rod) estimated a doorway to be narrower compared to subjects who held their arms next to their body (not holding a rod). Table 3 reports the relevant statistics and estimated success probabilities (power) for replication studies with the same sample sizes. If the means and standard deviation estimated from Experiment 1 are accurate, then future studies with the same sample size should reject the null hypothesis around 65% of the time.

Experiment 2 was similar in design (changing only how the estimated size of the doorway was produced) and result. Based on the reported means and standard deviations, replication experiments with the same sample sizes are expected to produce a significant outcome around 55% of the time.

Table 3 Statistical properties, hypotheses, and estimated probability of success for the tests in the five experiments from Firestone and Scholl (2014)

	Statistics	Supporting hypotheses	Probability of success
Experiments 1 and 3	n_1=10, n_2=10, $t(18)$=2.57, p=0.019	$\mu_1 \neq \mu_2$	.644
		$\mu_1 = \mu_3$	.932
	n_3=10, $t(18)$=0.43, p=0.67	Joint	.619
Experiment 2	n_1=10, n_2=10, $t(18)$=2.33, p=0.032	$\mu_1 \neq \mu_2$	.560
Experiment 4	n_1=41, n_2=41, $t(80)$=1.73, p=0.088	$\mu_1 \neq \mu_2$	.522
Experiment 5	n_1=44, n_2=45, $t(87)$=2.13, p=0.036	$\mu_1 \neq \mu_2$	.551
All tests			.100

Experiment 3 was designed to remove the suspected demand characteristics that contributed to the significant result produced by Experiment 2. Success for this experiment was a nonsignificant finding that compared the results of Experiment 3 (where subjects held a rod) and the results of the no-rod subjects in Experiment 1. Statisticians warn that a nonsignificant result should not be interpreted as supporting the null, but we follow the lead of Firestone and Scholl (2014) here. Thus, the probability of success indicated in Table 3 for Experiment 3 (second row) is the probability, 0.932, of a replication study *not* rejecting the null hypothesis.

The tests for Experiments 1 and 3 involve a common data set (the no-rod condition of Experiment 1). When estimating success for both experiments, we consider this dependency by running 100,000 simulated experiments (using the means, standard deviations, and sample sizes reported by the original experiments) and measure how often the simulated data produce a significant outcome for Experiment 1 and a nonsignificant outcome for Experiment 3 (source code is available as described in the Appendix). We call the proportion of successes for both experiments the "Joint" probability of success. As Table 3 indicates, this probability is around 0.62.

Experiment 4 is a bit odd in that Firestone and Scholl (2014) used a marginal result (p=0.088) as support that their experiment replicated a finding in Banerjee, Chatterjee, and Sinha (2012) that thinking unethical thoughts made the world seem darker. In estimating power for this kind of experiment, I followed their lead and used a significance criterion of α=0.1. The estimated power is just a bit over 0.5. Experiment 5 was similar to

Experiment 4 but changed the reporting method; the estimated power is around 0.55.

If we accept the analyses in Firestone and Scholl (2014) to be appropriate (e.g., treat a marginal result as good enough and use a nonsignificant result to accept the null), then these data seem very strange. The results all perfectly match the ideas expressed by Firestone and Scholl (2014) even though random sampling would be expected to make such success very rare. Most individual experiments like these have only a bit more than a 50% chance of producing a significant result, but Firestone and Scholl (2014) report success four out of four times (and one anticipated null result). The probability of a set of experiments like these being so successful is found by multiplying the estimated success probabilities in Table 3 (using the Joint probability for Experiments 1 and 3). Experiments like these are predicted to have full success only 10% of the time.

Given the low odds of success for experiments like these, I think it is reasonable for readers to wonder how Firestone and Scholl (2014) got their reported results. Luck is a possibility, but then we should hardly trust the empirical results as being representative of their underlying populations. Indeed, even if the reported results are representative of the populations, replication studies using the same sample sizes are unlikely to be so successful. More generally, the empirical results in Firestone and Scholl (2014) seem little better than the precognition findings reported by Bem (2011).

It might seem that any set of five experiments would have these kinds of problems, but there are two ways to avoid the criticism of excess success. First, each experiment could have high (estimated) power. For example, if each experiment has power of 0.95, then the probability of all five experiments being successful is 0.95^5=0.77. When power is lower, the criticism can be avoided by reporting experimental failures. If estimated power is 0.6, then one expects an average of two failures from a set of five experiments. The binomial distribution indicates that the probability of getting three or more successes from five experiments is 0.68. Four or more successes from a set of five such experiments has a probability 0.34. Five or more successes from a set of five such experiments has a probability of only 0.078.

In hindsight, there are several oddities about the design and results reported by Firestone and Scholl (2014). Experiments 1 and 2 do show the same kind of results as previous work, but with half the number of subjects. Based on the effect size reported in Stefanucci and Geuss (2009), the new experiments had a slightly less than 50% chance of producing a significant result (see Table 4). The sample sizes for Experiments 4 and 5 seem better, with estimated power being around 0.8 (which some people treat as a target for experimental design).

Table 4 Standardized effect sizes from previous studies and the sample sizes used in the five experiments from Firestone and Scholl (2014); the success probabilities for these experiments indicate that the full set had very low odds of success

	Standardized effect sizes (Hedge's *g*)	**Sample sizes from *Firestone and Scholl (2014)***	**Probability of success**
Experiments 1 and 3	*g=0.914*	n_1=10, n_2=10	0.49
	g=0	n_3=10	0.95
Experiment 2	*g=0.914*	n_1=10, n_2=10	0.49
Experiment 4	*g*=0.63	n_1=41, n_2=41	0.80
Experiment 5	*g*=0.63	n_1=45, n_2=44	0.84
Full set			0.15

Even so, the probability of all experiments being successful (the product of the success probabilities for all experiments) is around 0.15, based on the effect sizes reported by previous studies. (Hedge's *g* is a estimate of the standardized effect size; it is similar to Cohen's *d*.) It should have been clear before gathering any data that the experiments planned by Firestone and Scholl (2014) had very low odds of success. What is baffling is that this long-shot investigation worked!

We can speculate about ways that the reported set of studies ends up showing too much success. Optional stopping seems unlikely for Experiments 1–3, because Firestone and Scholl (2014) seem to have used the same number of subjects per condition as Stefanucci and Geuss (although Stefanucci and Geuss collapsed data across two conditions and so ultimately used twice as many subjects as Firestone and Scholl). Firestone and Scholl (2014) also used the same methods and measures as Stefanucci and Geuss (2009), so it seems unlikely that they combed through the data to find significant results. Given that their experiments were rather unlikely to be successful, it seems that Firestone and Scholl (2014) were either (un)lucky, or they ran variations of the experiments several times and only reported the significant results (publication bias).

For Experiments 4 and 5, Firestone and Scholl (2014) used more than double the number of subjects as Banerjee et al. (2012). Firestone and Scholl (2014) do not provide any justification for the larger sample size, and it is curious that their results were just below the significance criterion (a marginal criterion in the case

Table 5 Standardized effect sizes from previous studies and sample sizes that would produce a 0.8 success probability for the five experiments in Firestone and Scholl (2014)

	Standardized effect sizes (Hedge's *g*)	**Minimum sample sizes**	**Probability of success**
Experiments 1 and 3	*g=0.914*	n_1=34, n_2=34	0.96
	g=0	n_3=34	0.95
Experiment 2	*g=0.914*	n_1=34, n_2=34	0.96
Experiment 4	*g*=0.63	n_1=71, n_2=71	0.96
Experiment 5	*g*=0.63	n_1=71, n_2=71	0.96
Full set			0.806

of Experiment 4). Results just below the criterion are consistent with optional stopping, where subjects are added until the data produce the desired result. The marginal reported result in Experiment 4 might imply that additional subjects were run, produced a still larger *p*-value, and then were dropped from the report.

We do not know what really happened in the experiments reported by Firestone and Scholl (2014). The previous speculations are just one set of investigative approaches that could produce excess success. Even the authors may not know what happened in their set of studies, but this lack of knowledge does not engender much confidence from the perspective of a reader.

Suppose we were starting over and wanted to run a set of experiments with an 80% chance of success for the entire set. We impose the traditional α=0.05 criterion for each study. Working backward, we can compute that each significant result needs to have a power of 0.96 so that their product (and the 0.95 from the anticipated null result of Experiment 3) equals 0.806. Table 5 shows that using effect sizes from Stefanucci and Geuss (2009) and Banerjee et al. (2012), the full set of experiments requires a minimum of 454 subjects, which is more than double the 214 subjects used by Firestone and Scholl (2014). It is worth keeping in mind that the original studies may have overestimated the effect sizes, so these sample sizes should probably be considered optimistic.

I want to emphasize that I think the basic ideas in Firestone and Scholl (2014) are excellent. Their observations about the El Greco effect seem valid regardless of the quality of their empirical studies. It is unfortunate that their reported empirical results seem rather unbelievable and thereby detract from the paper overall.

Take away message: Experimental results based on faulty uses of hypothesis testing often leave a pattern of results that reveal the presence of the misuse. Estimating the replication probability of a set of experiments can reveal that the reported results seem too good to be true.

12 Preregistration May Be Useful but Is Not Necessary for Good Science

We saw that Type I error control is strongly influenced by specifying the hypotheses to be tested. Indeed, with such specification, scientists may end up using conservative hypothesis tests. In contrast, without such specification, scientists often engage in exploratory work, where Type I error control is hardly possible. This observation suggests that whenever feasible, scientists should specify the hypotheses that properly test their claims. I think this suggestion is not controversial, and it seems to fit in with recent calls for researchers to preregister their experiment, analysis methods, and data collection plans (Chambers, 2013; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Wolfe, 2013).

The idea of preregistration is that before actually running an experiment, a scientist describes the total experiment plan in a place where the scientist cannot subsequently alter it (e.g., the Open Science Framework or AsPredicted .org). This experiment plan describes the stimuli, tasks, experimental methods, measures, number of subjects and how they are sampled, the questions to be investigated, and the data analysis plan. After writing down these details, the experiment is run and any deviation from the preregistered plan is noted (perhaps with justification). Proponents of preregistration note that it prevents researchers from generating theoretical ideas or methods of data analysis after looking at the data, which is sometimes called "hypothesizing after the results are known," HARKing (Kerr, 1998), or the "garden of forking paths" (Gelman & Loken, 2014). With preregistration, it would also be obvious that a researcher stopped data collection early or added observations (perhaps due to optional stopping) or that various measures were combined in a way that is different from what was originally planned. If preregistered documents are in a public place, preregistration might also reduce the occurrence of publication bias because there is a public notification about the researcher's intention to run the experiment. Along similar lines, journals might agree to publish preregistered experiments prior to data collection, which would prevent reviewers and editors from publishing only significant results.

These attributes all seem like good pragmatic reasons for scientists to practice preregistration. However, I want to consider what should be inferred when a researcher sticks to the preregistered plan. Does success for a preregistered strategy lend some extra confidence in the results or in the theoretical

conclusion? Does it increase belief in the process that produced the preregistered experimental design? A consideration of two extremes suggests that it does not.

Extreme case 1. Suppose for some topic a researcher proposes a standardized effect size that was picked randomly from values between 0 and 1. The number ends up being d=0.37, so the researcher plans an experiment and analysis (one-sample t-test, n=79 subjects, which gives 90% power if the true effect corresponds to d=0.37) and preregisters everything. The experiment is subsequently run and finds the predicted effect. Whether or not the population truly has an effect, surely such an experimental outcome does not actually validate the process by which the hypothesis was generated (a random number generator). For the experiment to validate the *prediction* of the hypothesis (not just the hypothesis itself), there needs to be some justification for the theory/process that generated the prediction. Preregistration by itself does not, and cannot, provide such justification; so preregistration seems rather silly for unjustified experimental designs.

Extreme case 2. Suppose a researcher generates a hypothesis by deriving an effect size (d=0.37) from a quantitative theory that has previously been published in the literature. The researcher preregisters this hypothesis and the corresponding experimental design (one-sample t-test, n=79 subjects, which gives 90% power if the true effect corresponds to d=0.37). The subsequent experiment finds the predicted difference. Such an experimental finding may be interpreted as validation of the hypothesis and of the quantitative theory, but it does not seem that preregistration has anything to do with such validation. Since the theory has previously been published, other scientists could follow the steps of the researcher and derive the very same predicted effect size and thereby conclude that the experimental design was appropriate. In a situation such as this, it seems unnecessary to preregister the experimental design because describing its justification achieves the same ends. (Or, other scientists could find an error in the prediction or design, which would undermine the conclusions. Preregistration does not protect against this possibility.)

Most research situations are neither of these extremes, but researchers in psychology often design experiments based both on vague ideas, intuition, or curiosity and on well-defined past experimental results or quantitative theories. It is impossible to gauge the quality of the experimental design for the vague parts, and preregistration does not change that situation. For those parts of the predicted hypotheses (and methods and measures) that are quantitatively derived from existing theory or knowledge, it is possible to gauge the quality of the experiment from readily available information; preregistration does not add anything to the quality of the design.

Preregistration does force researchers to commit to making a real prediction and then creating experiments and specifying analyses that properly test that prediction. This type of prediction and experimental design is a laudable goal. But such a goal does not make sense if researchers do not have any hope of achieving it. When researchers design their experiments based on vague ideas, they are doing exploratory work, and it is inappropriate to ask (or even to invite) such researchers to make predictions. If forced to do so, researchers may generate some predictions, but those predictions will not be meaningful with regard to the process by which they were generated. At best, such studies would provide information about a researcher's intuition, but scientists are generally not interested in whether researchers can generate good guesses. They run studies to test aspects of theoretical claims.

At a practical level, many researchers who are motivated to preregister their hypotheses may quickly realize that they cannot do it, because their theories are not sufficiently precise. A field that insists on preregistration may find that it trades a "replication crisis" for a "theory crisis." That might be a good discovery for those researchers and for the broader field, and it may lead to better science in the long term. Likewise, preregistration does deal with some potential problems, such as optional stopping, dropping unsuccessful conditions, and HARKing. But these are exactly the issues that are handled by good justification for experimental design. Without good justification for experimental design, researchers are engaged in exploratory work, and then practices such as HARKing make sense, along with an appropriate cautionary interpretation.

In summary, writing down the justifications for an experimental design may be a good activity for scientists to self-check the quality of their planned experiment. Moreover, when attempting to be so precise, it may often be the case that scientists learn that part of their work is exploratory. Recognizing the exploratory parts of research can help guide how scientists interpret and present their empirical findings. However, justification for an experimental design should be part of a regular scientific report about the experiment; so there seems to be no additional advantage to publishing the justification in advance as a preregistration.

Take away message: The benefits of preregistration should be part of normal scientific practice of justifying your experimental design. If you cannot provide such justification, then preregistration does not help.

13 Hypothesis Testing Is a Variation of Signal Detection Theory

Many judgments about perception involve decision making under noise. An often useful analysis of such a situation involves the application of Signal

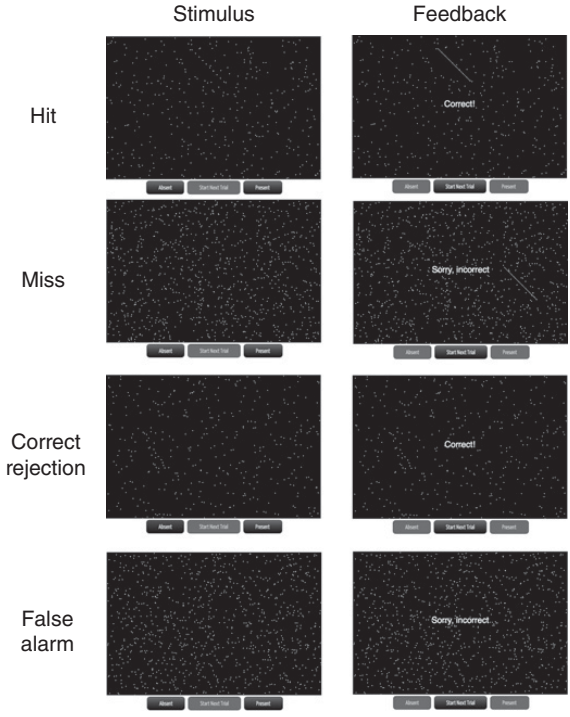

Figure 10 Stimuli and feedback in a signal detection experiment. The different rows indicate the four possible combinations of stimulus and response feedback.

Detection Theory (Macmillan & Creelman, 1991). Readers already familiar with Signal Detection Theory can skip ahead to Section 13.1, where the approach is used to clarify properties of hypothesis testing.

The left side of Figure 10 shows stimuli in an experiment (Francis & Neath, 2018) that can be evaluated with a signal detection analysis. On each experimental trial, the participant is shown a field with many randomly placed white dots. On some trials, there is a target (signal) that consists of an additional set of 10 equally spaced dots forming a straight line oriented 45 degrees left of vertical. The line is randomly placed among the other dots. The participant's task on each trial is to determine if the display contains the target (signal-and-noise) or the target is absent from the display (noise-alone). The right side of Figure 10 shows the experimental feedback provided to the participant after making their response. In addition to text feedback, if the target is present, then a line connects the 10 target dots. The response and feedback combinations are labeled for each row. A "Hit" corresponds to a situation where the target is present and the participant reports it is present. A "Miss" occurs when the target

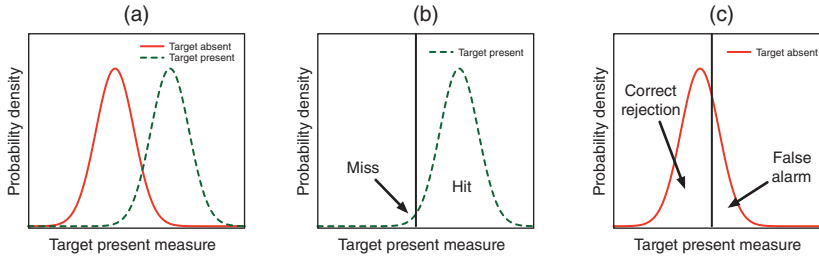

Figure 11 Hypothetical distributions for the target present measure under conditions where the target is actually present (dashed curve) or absent (solid curve). In (b) and (c), the vertical black line indicates a decision criterion. A sampled value above the criterion is interpreted as indicating the target is present, while a sampled value below the criterion is interpreted as indicating the target is absent. Figure 11b shows the areas under the curve of the target present condition that correspond to the miss and hit rates. Figure 11c shows the areas under the curve of the target absent condition that correspond to the correct rejection and false alarm rates.

is present but the participant mistakenly says it is absent. A "Correct rejection" is when the target is absent and the participant reports it is absent. Finally, a "False alarm" happens when the target is absent but the participant mistakenly reports it is present.

It might not seem like it at first, but the information that determines how well a participant can discriminate a target present from a target absent stimulus is the same kind of information used in hypothesis testing. In Signal Detection Theory, we suppose that a participant computes a value for each stimulus that, in some way, measures the "presence of a target." We do not know the details of this computation for the stimuli in Figure 10, nor do we know the units of measurement. Regardless of the details, different stimuli will produce different values because different arrangements of random dots can mask, highlight, or mimic the presence of the target (c.f., the stimuli in Figure 10). Signal Detection Theory supposes that the computed value is normally distributed across different stimuli. When the target is present, the mean of this distribution will be larger than when the target is absent. Figure 11a schematizes how this measure might be distributed for target absent and target present situations. Each curve has a distribution of values for the presence of a target because different stimuli mask or emulate a target to different degrees.

Signal Detection Theory supposes that viewing a stimulus is equivalent to taking a random sample from one of the distributions. In the theory, the sampled value is compared to a fixed classification criterion. If the sampled value is

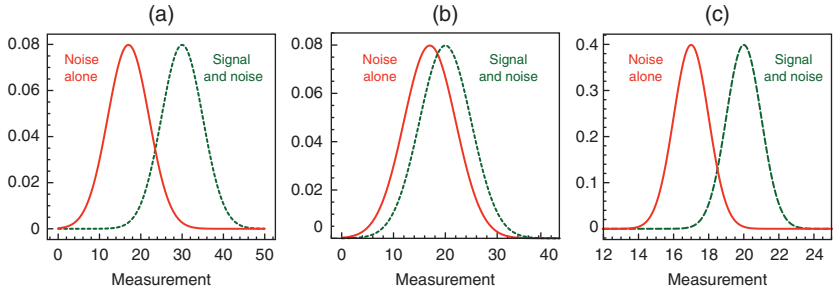

Figure 12 Three examples of hypothetical signal-and-noise and noise-alone distributions.

larger than the criterion, the stimulus is classified as "target present." If the sampled value is smaller than the criterion, the stimulus is classified as "target absent." Figures 11b and 11c show how the area under each curve corresponds to hits, misses, correct rejections, and false alarms for the target present and target absent situations.

To understand why Signal Detection Theory is useful for studies of perception, we need to better understand how decisions are related to the overlap of the distributions. Figure 12a schematizes distributions for the signal-and-noise situation and the noise-alone situation. Here, the noise-alone distribution has a mean value μ_{NA}=17, while the signal-and-noise distribution has a mean value μ_{SN}=30. Both distributions have a standard deviation of σ=5, which indicates how random noise makes a measurement differ from the mean of each distribution. An observed measurement involves drawing a random sample from one or the other distribution.

Suppose you get a measurement of unknown origin and you want to determine whether it came from the signal-and-noise or the noise-alone distribution. For the distributions in Figure 12a, it should be clear that a measurement value above 40 almost surely came from the signal-and-noise distribution, while a measurement value below 10 almost surely came from the noise-alone distribution. Measurement values between these extremes could have come from either distribution, but values between 23.5 and 40 are more common for the signal-and-noise distribution than for the noise-alone distribution. Similarly, values between 10 and 23.5 are more common for the noise-alone distribution than for the signal-and-noise distribution. If the two conditions are equally probable, then a good strategy might be to classify a given measurement according to which distribution has the higher probability density for that measurement. This would suggest that any measurement above the intersection point of the distributions (23.5) would be classified as from the signal-and-noise

distribution, while any measurement below the criterion of 23.5 would be classified as from the noise-alone distribution.

This strategy works fairly well for the distributions in Figure 12a, but mistakes are inevitable. Sometimes the noise-alone distribution will produce measurement values bigger than 23.5 and sometimes the signal-and-noise distribution will produce measurement values smaller than 23.5. When making decisions in noise, there is no way to avoid making some errors. In general, the ability to make good decisions depends on the separation of the two distributions. For example, Figure 12b shows distributions with more overlap because the signal-and-noise distribution now has a mean of μ_{SN}=20. Even with a criterion placed at the intersection of the distributions (18.5), many classification mistakes will be made because the two distributions generate such similar values.

Figure 12c shows distributions with the same mean values of Figure 12b but with a smaller standard deviation (σ=1). With less noise, it should be possible to fairly well classify a measurement as coming from the signal-and-noise distribution or from the noise-alone distribution. When making decisions in noise, the standardized separation of the distributions can be described by the signal-to-noise ratio:

$$d' = \frac{\mu_{SN} - \mu_{NA}}{\sigma}$$

For the pairs of distributions in Figures 12a–c, we have $d' = 2.6$, $d' = 0.6$, and $d' = 3.0$, respectively.

So, if we know the properties of the distributions, then d' is a convenient summary of how well a signal-and-noise (e.g., target present) condition can be distinguished from a noise-alone (e.g., target absent) condition. Unfortunately, in studies of human perception, the distributions cannot be directly observed, and even the unit of the target/signal present measure is unknowable. Despite these challenges, Signal Detection Theory allows scientists to estimate d' for a pair of distributions. Across multiple trials of an experiment with equal proportions of target present and absent conditions, an observer will produce a proportion of hits (H), misses (M), false alarms (FA), and correct rejections (CR). The experimenter creates the stimuli and thus knows what kind of stimulus is presented on each trial and also knows how to classify the observer's response on each trial. Traditionally, the calculations for d' are described using hits and false alarms, but it is more intuitive to use misses and correct rejections. We will show how to transform the formulas to the more traditional version later.

To compute d' from response proportions, we first identify the relative position of the criterion for the signal-and-noise distribution. We do this by

finding the percentile score for the rate of misses on a standard (mean of zero and standard deviation of 1) normal distribution. The percentile of a distribution is the score that divides the distribution into a bottom portion with an area under the curve equal to the given proportion and a top portion being one minus the given proportion. For example (using the distributions in Figure 12b), if the miss rate is 0.21, the percentile is −0.806. Next, we identify the relative position of the criterion for the noise-alone distribution. We do this by finding the percentile score for the rate of correct rejections on a standard normal distribution. For example, if the correct rejection rate is 0.42, then the percentile is −0.202; d' is simply the difference of these percentiles:

$$d' = \text{Percentile}(CR) - \text{Percentile}(M) = 0.604$$

which is very close to the 0.6 that we computed earlier for Figure 12b using the formula for means and standard deviation. The small difference is due to rounding in the proportions. The percentile formula works because each percentile is based on an (unknown) fixed decision criterion value. If the signal-alone and signal-and-noise distributions were the same, then the miss rate and the correct rejection rate would be the same. Within the theory, these rates are different only when the signal-and-noise distribution is shifted to the right of the noise-alone distribution. The magnitude of the shift is reflected by the difference in the rates, and the percentile calculation tells us exactly how large the shift must be to produce the rate difference. Thus, critical information about the distributions can be computed from the miss and correct rejection rates, even when the criterion, means, and standard deviation are unknown.

Traditionally, d' is computed using the hit and false alarm rates. Although less intuitive than using miss and correct rejection rates, this approach is mathematically equivalent because of a simple relationship between percentiles. Namely, since hits and misses make up all possibilities for the signal-and-noise distribution, it follows that

$$\text{Percentile}(M) = -\text{Percentile}(H)$$

Likewise, correct rejections and false alarms make up all possibilities for the noise-alone distributions, so

$$\text{Percentile}(CR) = -\text{Percentile}(FA)$$

Plugging these alternative percentile terms in the d' equation presented earlier and rearranging terms give the typical formula used in Signal Detection Theory:

$$d' = \text{Percentile}(H) - \text{Percentile}(FA)$$

Typically, studies of perception compute d' to estimate how distinctly the brain represents different types of information (noise-alone versus signal-and-noise).

Importantly, the calculation of d' is unrelated to the value used for the decision criterion. Even though a more liberal criterion might lead to a higher hit rate, it will also generate a higher false alarm rate; together they will correspond to the same value of d'. If we arbitrarily suppose that the mean of the noise-alone distribution is zero, then we can compute the criterion value in standard units as

$$\text{Criterion} = -\text{Percentile}(FA)$$

Because this value depends on an arbitrary assumption about the mean, researchers instead usually calculate a term that represents bias. When the signal is present as frequently as it is absent, the best criterion (in the sense that it maximizes percentage correct) is at the intersection of the two distributions:

$$\text{Criterion} = \frac{d'}{2}$$

Bias is measured relative to this optimal criterion:

$$\text{Bias} = -\text{Percentile}(FA) - \frac{d'}{2}$$

Somewhat confusingly, this bias term is often called C to indicate the criterion, but it is not actually the criterion that is used to make decisions. A liberal criterion (below the best criterion) produces more hits and false alarms and a negative bias value, while a conservative criterion (above the best criterion) produces fewer hits and false alarms and a positive bias value.

13.1 Signal Detection Theory and Hypothesis Testing

In hypothesis testing, d' is used in two ways. The first use of d' is for the population standardized effect size. The standardized effect size Cohen's δ is simply a d' that describes the standardized separation of population distributions. For example, it is well known that the average American male is about 14 centimeters (6 inches) taller than the average American female. The standard deviation of heights is around 15.4 centimeters, so $d' = 0.9$. Figure 13a shows the population distributions. Now, suppose you are given a height and asked to classify it as corresponding to a male or a female (which distribution is signal-and-noise and which is noise-alone is rather arbitrary in this situation). Your intuition about people's height should tell you that for a broad range of values, knowing only the measured height of a person is not especially informative about whether they are male or female. Indeed, it is fairly easy to deduce that (if

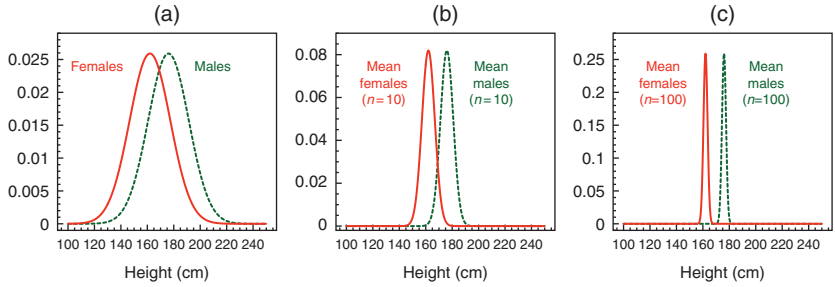

Figure 13 Distributions of heights for US males and US females. (a) The population distributions have a separation of $d' = 0.9$. (b) For sampling distributions of means from samples of n=10, $d' = 2.8$. (c) For sampling distributions of means from samples of n=100, $d' = 9.0$.

males and females are equally common) a criterion at the intersection of the distributions produces a misclassification (male for female or female for male) with a probability of 0.32 (any other criterion does even worse). The bad news is that the d' value for height differences between males and females is quite large compared to many phenomena in psychology. As rules of thumb, Cohen (1988) characterized δ values of 0.2, 0.5, and 0.8 as indicating small, medium, and large effect sizes, respectively. Using the best possible criterion, misclassification errors for these values are 0.46, 0.40, and 0.34, respectively.

Thus, for many of the phenomena commonly investigated in psychology, it is very challenging to identify which distribution produced a given measurement. However, all is not lost because instead of focusing on individual measurements, we can investigate the properties of the *mean* of measurements. The advantage of working with the mean is that the variability of the distribution of mean values (known as the sampling distribution) is affected by the sample size, n. The standard deviation of the sampling distribution of the mean is called the *standard error of the mean*, and it can be estimated from a sample standard deviation, s, as

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

Figures 13b and 13c show the sampling distributions of the means for samples of size n=10 and n=100, respectively. Notice that the sampling distributions for the male and female means have much less overlap ($d' = 2.8$ and $d' = 9.0$) than the population distributions. Thus, misclassification probabilities for means are rather small. For samples of size n=10, the misclassification probability is 0.08. For samples of size n=100, the misclassification probability is nearly zero.

Hypothesis testing leverages the small standard deviation of the sampling distribution so that d' is in a range where classification of effects (or not) can be discriminated. This is the second use of d' in hypothesis testing. The t statistic that is used for one-sample and two-sample tests of means simply estimates the d' of the sampling distributions. For a one-sample t-test the formula is

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}}$$

where $\overline{X}$ is an estimate of the signal-and-noise distribution mean from the sample data and μ is the mean of the noise-alone distribution, as specified in the null hypothesis. The situation is similar for two-sample t-tests (Herzog, Francis, & Clarke, in press).

To actually make a decision, we have to specify a decision criterion. If we know the properties of the distributions, then the intersection of distributions is a good choice. However, in hypothesis testing we can only *estimate* the signal-and-noise distribution, so we cannot use that estimate to determine the criterion. Since bigger d' values correspond to easier classifications, if t is big enough, then we conclude that we are in a situation where it should be easy to find a signal if it exists. We have a "significant" result. Roughly speaking, psychology uses a criterion value of 2; while physics opts for a criterion value of 5. More precisely, what counts as "big enough" for the t value is determined by an acceptable false alarm rate (Type I error). We set the criterion so that we (mistakenly) conclude signal-and-noise when the measurement actually came from the noise-alone distribution only at the specified rate (e.g., 0.05).

Take away message: Hypothesis testing is an application of Signal Detection Theory as applied to sampling distributions. The criterion for hypothesis testing is established to control the Type I error rate (false alarm rate).

14 Using Signal Detection Theory to Analyze Reported Results of Hypothesis Testing

We can use Signal Detection Theory to evaluate bias of the hypothesis testing approach. We know the false alarm rate in an hypothesis test because we set it as the Type I error rate (e.g., 0.05). The hit rate is the power of the experiment. Figure 14 plots bias as a function of power for three values of Type I error. The horizontal gray line indicates no bias. We see that for all of these Type I error rates, hypothesis testing is biased toward the noise-alone (null hypothesis) distribution, except for very high power values. Technically, we do not accept the null hypothesis (noise-alone distribution), but that is often the default position of skeptical scientists. It should not be surprising that hypothesis

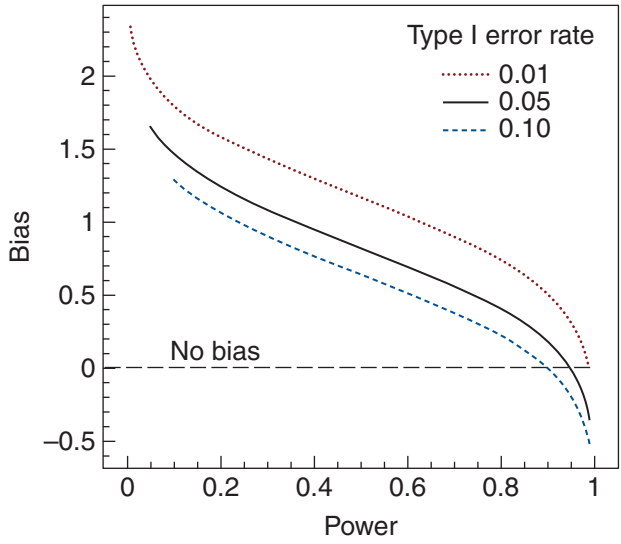

Figure 14 The bias of hypothesis testing as a function of power for three Type I error rates. In most cases a hypothesis test is biased against the alternative hypothesis.

testing is biased against the alternative because the hypothesis-testing procedure sets a criterion to minimize false alarms (Type I errors) regardless of the unknown d' value. It is only when d' is so large that power is greater than $1-\alpha$ that the test becomes biased toward the alternative hypothesis.

You could do the same kind of analysis to compute d' from the Type I error rate and power. However, the calculation would simply report the t value (for a one-tailed test) that produces the given power value. Remember, the t value we use in an hypothesis test is just the d' of the sampling distributions.

This discussion is mostly for pedagogical purposes. We cannot, for example, adjust the criterion to reduce bias in hypothesis testing because we do not know the true effect size for the alternative hypothesis. We should note, however, that Signal Detection Theory emphasizes that the choice of the criterion always involves a trade-off between hits and false alarms. For example, recent calls (Benjamin et al., 2018) to reduce the desired Type I error from the typical 0.05 to 0.005 should have the benefit of decreasing Type I errors (false alarms), but at the cost of decreasing power (hits).

Take away message: In terms of Signal Detection Theory, hypothesis testing tends to be biased against the alternative hypothesis. This bias seems appropriate given that the goal of hypothesis testing is to fix the rate of Type I errors.

15 Conclusions

Hypothesis testing is a common method of analyzing experimental data. When everything works well, it is easy to understand the appeal. Being able to control the Type I error rate is a good thing, and the methods are (generally) easy to apply.

However, we have seen that deviating just a bit from the textbook examples can cause serious problems with hypothesis-testing approaches. Sampling, analysis strategies, reporting, and measurement issues can cause hypothesis testing to produce much higher Type I error rates than might be expected. At the same time, some standard hypothesis-testing strategies produce Type I error rates much lower than what is intended, thereby making it difficult for scientists to convince their peers about a new discovery. To reiterate some of the main points of the text, Table 6 recapitulates the "take away" messages from the earlier sections.

Table 6 Summary of the take away messages from this Element

Topic	Take away message
Basics	When done properly, hypothesis testing controls the Type I error rate and the calculations are fairly easy to perform.
Robustness	The *t*-test is quite robust to deviations from some of its assumptions; but if you have unequal sample sizes, you should use Welch's test rather than a standard *t*-test.
Optional stopping	Unless you are in a situation where you can fix the sample size, hypothesis testing does not necessarily do a good job controlling the Type I error rate. Unfortunately, it is difficult to avoid optional stopping.
Conservative ANOVA	If your conclusion requires multiple test outcomes, then the Type I error rate for your conclusion might be much smaller than the criterion you set for any individual test.
Multiple testing	Multi-way ANOVA designs generally do a poor job controlling the Type I error rate for exploratory investigations that look for some significant results out of many tests.
Power for all tests	Power analyses and sample size determinations need to consider the full set of tests that are relevant to the conclusions. Not including the full set can lead to dramatically underpowered experiments.
Planning *p*-value	Experiments measuring real effects tend to produce very small *p*-values. If you consistently find *p*-values just

Table 6 (cont.)

Topic	Take away message
	below the significance criterion, then you are probably doing something wrong. When planning an experiment, you can only pick sample sizes that likely produce very small *p*-values (e.g., high power); you cannot plan for a *p*-value to be just below the significance criterion.
Subjects and trials	It is not true that trials and subjects can be traded off. An increase in the number of subjects has a bigger impact than an increase in the number of trials. In general, you should consider the full characteristics of your data collection and analysis design when planning sample sizes.
Replication and Type I error	There may be good reasons to run replication studies, but they do not include avoiding Type I errors.
Excess success	Experimental results based on faulty uses of hypothesis testing often leave a pattern of results that reveals the presence of the misuse. Estimating replication probability of a set of experiments can reveal that the reported results seem too good to be true.
Preregistration	The benefits of preregistration should be part of normal scientific practice of justifying your experimental design. If you cannot provide such justification, then preregistration does not help.
Signal Detection Theory	Hypothesis testing is an application of Signal Detection Theory as applied to sampling distributions. The criterion for hypothesis testing is established to control the Type I error rate (false alarm rate).
Signal Detection Theory analysis of hypothesis testing	In terms of Signal Detection Theory, hypothesis testing tends to be biased against the alternative hypothesis. This bias seems appropriate given that the goal of hypothesis testing is to fix the rate of Type I errors.

The reader might now want this section to explain how to solve the challenges that arise from these messages. Unfortunately, the author does not have straightforward solutions. Some problems (such as optional stopping and multiple testing) are more easily handled with Bayesian methods, but these methods introduce new challenges (such as picking a prior distribution or ways of comparing models).

Perhaps the best advice is for scientists to remember that statistical results are not the goal of empirical studies. Rather, the goal of scientists is to *explain* how things work by measuring and characterizing mechanisms. At best, statistics can only direct us toward models and insights that help clarify the mechanisms that underlie the phenomena we study. Without an understanding of mechanisms, scientists will be dabbling with phenomena that they do not really understand.

Appendix

Most of the data sets and software scripts (R core team, 2017) for the reported analyses are available at the Open Science Framework (https://osf.io/pzc4n/). This Appendix briefly describes the available files and explains how they are used.

A.1 Muller–Lyer and Horizontal-Vertical Illusion Data Set

Each trial from the 310 participants is in the file MLHVTBT.csv. Each participant is identified by a unique "SUBJECT_ID." In addition to the conditions described in the Element text, there were also judgments where the comparison line length was 150 pixels. In the data file, this variable is called "TargetLength."

A.2 Analysis of Muller–Lyer Data

The R script file IllusionAnalysis.R reads in the data file and creates a data frame consisting of only the experimental trials with 100-pixel-long comparison stimuli (practice and demographic trials are ignored). For each participant, the code computes the mean across trials for each stimulus condition. It then computes means, standard deviations, and correlations across participants for the various wing conditions for horizontal and vertical trials. These values are printed on the console window.

To produce the analysis in Section 2.1, the code runs a paired *t*-test between the outward wings and inward wings for the horizontal orientation. The result of the test is printed on the console window.

For the correlation analysis in Section 4.1, the code first runs a significance test for the correlation between the outward wings and inward wings for the horizontal orientation using the entire data set. The result is printed to the console. Next, the code simulates an optional stopping approach that takes a subset of the data and tests for a significant correlation. The loop starts with the first 30 participants and gradually increases the sample size until it includes the entire data set. As discussed in Section 4.1, the smallest such subset that produces a significant result is for n=53. The loop prints the correlation and corresponding p-value for each subset of data to the console.

A.3 Robustness of *t*-test Simulations

The robustness simulations in Section 3 were generated by code in the online textbook of Francis (2018). The simulation is online at https://introstatsonline

.com/chapters/chapter12/homogeneity_variance_sim.shtml, but access is restricted to those who have purchased a registration code and set up an account.

A.4 Multiple Testing in ANOVA

The file ANOVATypeIErrorRate.R generates null data to run 10,000 simulated 2×2 independent ANOVAs. Each test looks for either main effect or an interaction. The result printed to the console reports the Type I error rate for each test (around 0.05) and the Type I error rate for at least one significant result across the three tests (around 0.14). The Element reports these results in Section 6.

The file ANOVATypeIErrorRate222.R runs a similar analysis for a 2×2×2 ANOVA. The results are reported in Section 6.

The file ANOVAFactors.R produces the data plotted in Figure 6.

A.5 Power for All Tests

The power analysis in Section 7.1 for the Muller–Lyer and vertical-horizontal illusion experiment can be reproduced with file IllusionsPower.R. The console reports the power of each test and the power for the conjunction of all specified tests. Change the value of the standard deviation (variable SD in the code) to 16 to produce the corresponding curve in Figure 7.

A.6 Distribution of *p*-values

The file pValues.R computes the distribution of *p*-values, as discussed in Section 8 and displayed in Figure 8.

A.7 Trials and Subjects

The file SubjectsVsTrials.R produces simulation results like those in Figure 9a. The file SubjectsVsTrials2.R produces simulation results like those in Figure 9b. The file PowerSubjectsVsTrials.R generates the simulation results reported in Section 9.1 that demonstrate an advantage of more participants over more trials-per-participant for a replication study of the Muller–Lyer and horizontal-vertical illusion experiment.

A.8 Test for Excess Success

Most of the power calculations (Tables 3, 4, and 5) were done with the "pwr" library in R (Champely et al., 2018). The file TESAnalysis.xls provides a summary of intermediate calculations. For Table 3, experiments 1 and 3 had

multiple tests and comparisons, so power was estimated with simulated experiments. The file TestExcessSuccessExp1-3.R repeats this analysis.

A.9 Signal Detection Theory Analysis of Bias in Hypothesis Testing

The data to produce Figure 14 is generated by the file SDTofHypothesisTesting.R.

References

Banerjee, P., Chatterjee, P., & Sinha, J. (2012). Is it light or dark? Recalling moral behavior changes perception of brightness. *Psychological Science*, **23**(4), 407–409.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, **100**, 407–425.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R. . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behavior*, **2**, 6–10.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, 49, 609–610.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Ford, C., & Volcic, R. (2018). Package 'pwr': Basic functions for power analysis.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cramer, D. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. . . . Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANVOA: Prevalence and remedies. *Psychonomic Bulletin & Review*, **23**(2), 640–647.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, **25**(1), 7–29.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods*, **39**, 175–191.

Firestone, C., & Scholl, B. J. (2014). "Top-down" effects where none should be found: The El Greco fallacy in perception research. *Psychological Science*, **25**(1), 38–46.

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, **19**, 151–156.

Francis, G. (2014). The frequency of excess success for articles in *Psychological Science*. *Psychonomic Bulletin & Review*, **21**, 1180–1187.

Francis, G. (2018). *IntroStats Online 2*. Sage Publications. https://introstatsonline.com/

Francis, G., & Neath, I. (2018). *CogLab 5*. Cengage Publishing. https://coglab.cengage.com/

Francis, G., Tanzman, J., & Matthews, W. J. (2014). Excess success for psychology articles in the journal *Science*, *PLOS One*, **9**(12), e114255. doi:10.1371/journal.pone.0114255

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments & Computers*, **30**, 690–697.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist* 102, 460. doi:10.1511/2014.111.460

Herzog, M. H., Francis, G., & Clarke, A. (in press). *How to not lie with statistics: The essentials of statistics and experimental design for everyone with many examples from medicine and bioengineering*. New York: Springer.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, **1**(5), 658–676. doi:10.1002/wcs.72

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: Cambridge University Press.

McElreath, R. (2016). *Statistical rethinking*. Boca Raton, FL: CRC Press.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349 0.1126/science.aac4716

R core team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. www.r-project.org/

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, **1**(1), 19–26.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple study articles. *Psychological Methods*, **17**(4), 551–566.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, **9**(6) 666–681.

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, **25**(6), 2083–2101.

Stefanucci, J. K., & Geuss, M. N. (2009). Big people, little world: The body influences size perception. *Perception*, **38**, 1782–1795.

Ueno, T., Fastrich, G. M., & Murayama, K. (2016). Meta-analysis to integrate effect sizes within an article: Possible misuse and Type I error inflation. *Journal of Experimental Psychology: General*, *5*, 643–654.

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

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638.

Wolfe, J. M. (2013). Registered reports and replications in *Attention, Perception, & Psychophysics* [Editorial]. *Attention, Perception, & Psychophysics*, *75*, 781–783.

Cambridge Elements

Perception

James T. Enns

The University of British Columbia

Editor James T. Enns is Professor at the University of British Columbia, where he researches the interaction of perception, attention, emotion, and social factors. He has previously been Editor of the *Journal of Experimental Psychology: Human Perception and Performance* and an Associate Editor at *Psychological Science, Consciousness and Cognition, Attention Perception & Psychophysics,* and *Visual Cognition*.

About the Series

The modern study of human perception includes event perception, bidirectional influences between perception and action, music, language, the integration of the senses, human action observation, and the important roles of emotion, motivation, and social factors. Each Element in the series combines authoritative literature reviews of foundational topics with forward-looking presentations of the recent developments on a given topic.

Cambridge Elements

Perception

Elements in the Series

Hypothesis Testing Reconsidered
Gregory Francis

A full series listing is available at: www.cambridge.org/EPER