Dual Perspectives

Dual Perspectives Companion Paper: Better Inference in Neuroscience: Test Less, Estimate More, by Robert J. Calin-Jageman et al.

Neuroscience Needs to Test Both Statistical and Scientific Hypotheses

Bradley E. Alger
Department of Physiology, Program in Neuroscience, University of Maryland School of Medicine, Baltimore, Maryland 21201

Experimental neuroscience typically uses "*p*-valued" statistical testing procedures (null hypothesis significance testing; NHST) in evaluating its results. The rote, often misguided, application of NHST (Gigerenzer, 2008) has led to errors and "questionable research practices." Although the problems could be avoided with better statistics training (Lakens, 2021), there have been calls to abandon NHST altogether. One suggestion is to replace NHST with "estimation statistics" (Cumming and Calin-Jageman, 2017; Calin-Jageman and Cumming, 2019). Estimation statistics emphasizes the uncertainty inherent in scientific investigations and uses metrics, e.g., confidence intervals (CIs), that draw attention to uncertainty. Besides procedural steps and methods, the Estimation Approach prefers expressing "quantitative," rather than "qualitative" conclusions and making generalizations, rather than testing scientific hypotheses. The Estimation Approach embodies a philosophy of science—its ultimate goals, experimental mindset, and specific aims—that diverges unhelpfully from what laboratory-based neuroscience needs. The Estimation Approach meshes naturally with, e.g., clinical neuroscience, drug development, human psychology, and social sciences. It fits less well with much of the neuroscience published in the *Journal of Neuroscience*, for example. In contrast, the philosophy behind NHST fits naturally with traditional, evaluative testing of scientific hypotheses. Finally, some Estimation Approach remedies, e.g., replication, ideally with "preregistration," are incompatible with much experimental neuroscience. This Dual Perspective essay argues that, while neuroscience can benefit from practical aspects of estimation statistics, entirely replacing conventional methods with the Estimation Approach would be a mistake. NHST testing should be retained and improved.

Significance Statement

Experimental neuroscience relies on statistical procedures to assess the meaning and importance of its research findings. Optimal scientific communication demands a common set of assumptions for expressing and evaluating results. Problems arising from misuse of conventional significance testing methods have led to a proposal to replace significance testing with an Estimation Statistics Approach. Practical elements of the Estimation Approach can usefully be incorporated into conventional methods. However, the prevailing philosophy of the Estimation Approach does not address certain important needs of much experimental neuroscience. Neuroscience should adopt beneficial elements of the Estimation Approach without giving up the advantages of significance testing.

Introduction

Like other experimental sciences, neuroscience needs to test its hypotheses, determine the efficacy of its experimental treatments and assess the validity of its claims. In the 1800s, scientific statistics consisted mainly of comparing the means of group values and trying to estimate whether they were really different (Bernard, 1865/1957; Gigerenzer et al., 1989). There was no appreciation of variance or sample size and no way of rigorously determining, say, which fertilizer produced the best crop yield. In the 1930s, Ronald Fisher invented basic "significance testing," which grew into null hypothesis significance testing (NHST; Perezgonzalez, 2015a). Because it was a useful tool for decision-making, many branches of science adopted NHST. However, through misuse, misunderstanding, and abuse, NHST became a "mindless ritual" (Gigerenzer, 2008), which led to calls for its

Received June 10, 2022; revised Aug. 18, 2022; accepted Aug. 23, 2022.
I thank Tom Abrams, Bruce Krueger, and Brian Polster for their valuable comments on a draft of this manuscript. I also thank Bob Calin-Jageman for proposing this interesting discussion.
B.E.A. has published a book defending the scientific hypothesis.
Correspondence should be addressed to Bradley E. Alger at balgerlab@gmail.com.
https://doi.org/10.1523/JNEUROSCI.1134-22.2022

abandonment. Instead, this Dual Perspective essay advocates improving NHST by supplementing it with practical elements of the Estimation Approach. The inclusion of effect sizes, confidence intervals (CIs), and a heightened awareness of uncertainty could enhance neuroscience practice and communications. However, certain philosophical aspects of the Approach make it less appropriate for many areas of experimental neuroscience. The discussion assumes the frequentist statistics framework that most neuroscientists learn. In frequentist statistics, probability is an objective property of the world that is determined by the long-run frequency of occurrence of phenomena. Other statistical frameworks evaluate probability in subjective terms based on prior knowledge.

This essay will concentrate on advantages of NHST and shortcomings of the Estimation Approach. A critic of NHST, Ioannidis (Szucs and Ioannidis, 2017) focuses attention on the misuse of NHST in "...psychological science, cognitive neuroscience, and biomedical research," yet acknowledges that NHST "may have legitimate uses when there are precise quantitative predictions..." In much neuroscience, precise predictions of scientific hypotheses are possible, and NHST procedures will continue to be valuable.

The Estimation Approach recommends two main guidelines for neuroscience. It should, "pose quantitative research questions..." (i.e., as opposed to "qualitative" ones, see below), and "countenance uncertainty in all statistical conclusions..." (Calin-Jageman and Cumming, 2019). Moreover, the Approach holds that "Inference is at the heart of the scientific method: we collect finite datasets and then try to make reasonable generalizations about how the world works." These principles fail to recognize certain essential requirements of scientific reasoning.

Successful predictions and accurate descriptions of nature may come from reasonable generalizations without providing deeper mechanistic understanding of how the world works. Moreover, a central feature of the scientific method is the scientific hypothesis (Bernard, 1865/1957; Popper, 1959/2002), which the Estimation Approach almost entirely ignores. Scientific understanding is advanced by proposing and rigorously testing possible explanations (scientific hypotheses) for phenomena. Progress is made when scientific hypotheses either pass rigorous tests (and are "corroborated") or are rejected as false. This scientific-hypothesis testing process requires making "qualitative" decisions, which the Estimation Approach prefers to avoid. Accepting outcomes expressed in "quantitative" terms detracts from the need to reach definite conclusions about the soundness of our ideas. The significance-testing approach facilitates qualitative decision-making. Emphasizing uncertainty can provide valuable balance to scientific reports, but it must not detract from the search for truth.

Scientific versus statistical hypotheses

The term "hypothesis" refers to two quite different concepts: the scientific hypothesis and the statistical hypothesis. The distinction is widely overlooked (Box 1). The Estimation Approach implicitly deals almost exclusively with statistical hypotheses, not with the scientific hypotheses that engage many neuroscientists.

The roots of the scientific hypothesis are centuries old (Bernard, 1865/1957; Alger, 2019). The modern scientific hypothesis is fundamentally a proposed explanation for some phenomenon or property of the world being the way it is. It provides a tentative answer to the question "why?" For example, to explain the extinction of the dinosaurs, Luis and Walter Alvarez (Alvarez et al., 1980) proposed that ~66 million years ago, an enormous asteroid hit the earth and caused cataclysmic geological and meteorological reactions that killed off the dinosaurs. The Alvarez hypothesis has been corroborated by many tests and is considered a good explanation for the extinctions.

The statistical hypothesis is a 20th century invention. It is part of a mathematical testing procedure and is, itself, nonexplanatory. In fact, what we now call "significance testing" or "null-hypothesis significance testing," has been called, simply, "data-testing"; the notion of NHST was "concocted" in 1940 (Perezgonzalez, 2015a). A statistical hypothesis is commonly used to estimate whether sample groups are likely to have come from the same larger population. For example, a statistical null hypothesis, H_o, might state that the mean heights of men and women do not differ "significantly." (Two sample means will always differ at some level of measurement resolution. Scientists are only interested in differences that are "big enough" to be "significant." What counts as significant is a matter of convention.) A statistical test, say a *t* test, is applied to randomized samples of people's heights to determine whether they are likely to differ within probability limits. That is all that the test provides: a numerical comparison associated with a probability which must be interpreted. The outcome of statistical null-hypothesis testing says nothing about why the groups do or do not differ, or whether a "significant" difference reveals anything meaningful about the world. Statistical-hypothesis testing is merely a mechanical tool, albeit a very useful one, for decision-making.

The scientific hypothesis is an idea about the real world. It benefits science in many ways. It helps organize thinking and communication (Alger, 2019). Testing multiple predictions of a hypothesis improves experimental design and strengthens the reliability of conclusions about that hypothesis (Alger, 2020). Developing and testing multiple alternative hypotheses to explain a given phenomenon sharpens analytic reasoning and helps to reduce biases ("confirmation bias," "publication bias"). Creative, detailed thinking about a problem encourages considerations of "other interpretations" for data and decreases fallacies (Bernard, 2020). Scientific hypothesis formation and testing comes naturally to the human mind. People have powerful cognitive drives to understand the world (Kahneman, 2011). A primary inferential reasoning tool is the use of "counterexamples" (Johnson-Laird, 2010). We reflexively assess the soundness of rules (hypotheses) by trying to find cases which, if true, would disprove them. Use of the scientific hypothesis thus builds on natural cognitive tendencies. However, it is essential to state scientific hypotheses explicitly to protect against unconscious biases that can arise from tacit hypotheses. The appropriate use of scientific hypotheses requires making distinctions and drawing conclusions, such as whether experimental data are consistent or inconsistent with the hypothesis. For these reasons and more scientists need rational tools for making decisions. Conventional significance testing, correctly understood and cautiously applied, is such a tool.

Box 1. Comparison of scientific and statistical hypotheses

Similarities

1. Scientific and statistical hypotheses are frequently used together in scientific work.
2. Both are tested with empirical evidence, and the tests can only show that the hypotheses are false or that the results are consistent with them. Evidence cannot prove that the hypotheses are true.

Differences

1. Scientific hypotheses attempt to explain phenomena in terms of unobservable objects and processes ("mechanisms") that cannot ordinarily be directly measured. Scientific hypotheses are tested indirectly by deducing and testing measurable predictions that follow from the hypotheses. Unlike hypotheses, predictions can be tested directly and determined to be true or false by making measurements. Test results are compared to the predicted results: if the predictions are false, the hypothesis is false. If the predictions are true, the hypothesis might or might not be true. Confirmed predictions "corroborate" the scientific hypothesis, but do not "confirm" it (Popper, 1959/2002). Scientific hypotheses are tested via an enormous variety of empirical measures.

Note: scientific predictions and scientific hypotheses can be related to each other as "if...then" statements. "**If** [this scientific hypothesis is true] **then** [this prediction that it makes must also be true]." This is a deductive relationship; we deduce predictions from hypotheses. This is why falsifying a valid prediction can falsify a hypothesis that makes it. The "if...then" statement itself is not a hypothesis; it only puts a scientific hypothesis together with one or more of its predictions so that their relationships are obvious. The prediction part of the "if...then" statement generally points to a measurement that can be made to test the scientific hypothesis.

2. Statistical hypotheses are often used to evaluate the numerical results obtained by experimentally testing predictions of scientific hypotheses. Statistical hypotheses themselves do not make predictions and are tested only mathematically.
3. The two kinds of hypothesis are independent. Statistical hypotheses can be formulated and tested outside of science. And scientific hypotheses can be tested without using statistical hypotheses. The Alvarez hypothesis predicted that somewhere on earth there would be a huge impact crater with the same age as the dinosaur extinction. This predicted impact was a one-time event. It was confirmed by finding such a crater in the Yucatan peninsula of Mexico. No statistics were involved.

For instance, given the observation that rats treated with drug X lose weight, a scientific hypothesis might be that drug X depresses appetite. One prediction of this hypothesis is that rats treated with X will eat less food than nontreated rats. We could test this prediction by setting up a null statistical hypothesis, H_o, that the amount of food eaten by the two groups of rats will not differ significantly. A statistical test of food consumption data would determine whether any difference in amounts eaten is likely to have occurred by random chance alone.

Science tries to find real causes and effects and does not account for reliable effects by saying they "happened by chance." In neuroscience, the conventional level for a statistically significant difference is "$p \leq 0.05$," i.e., if the probability of getting a result by chance alone is less than or equal to 1/20, then the result is tentatively considered to be real. If X-treated rats ate significantly less food than untreated rats, then we would reject H_o (which said that food consumption of the two groups would not differ). This test would confirm the prediction of the scientific hypothesis that X-treated rats would eat less. Since the prediction was true, we conclude the scientific hypothesis that made it might be true.

Ideally, investigators would then go back to their scientific hypothesis, derive another prediction, perhaps that X-treated rats would not work as hard for food as nontreated rats, set up another statistical hypothesis, test that one, and so on. Their objective would be to discover the reason that X-treated rats lost weight. This is basic reasoning within the significance-testing NHST framework. By contrast, assuming this was an "exploratory" study operating under the Estimation Approach, the investigators might opt to design a larger, more tightly controlled "confirmatory" study, preferably preregistering it, and submit their plan for review. Their objective would be to determine more precisely the quantitative effect of X on rat food consumption.

Beneficial practical aspects of the Estimation Approach

Estimation statistics can readily be "translated" into conventional *p*-values, significance levels, etc. and could, in principle, be used for the same purposes as NHST is (Perezgonzalez, 2015b; Cumming and Calin-Jageman, 2017). One limit of a 95% CI is the *p*-value, $p = 0.05$. Values within the CI are judged nonsignificant and values outside it, significant. We could have compared food consumption of drug X-treated and untreated rats in this way. The CIs would highlight the range of uncertainty within the significance limits but would not fundamentally alter the strategy of testing a scientific hypothesis. The final result would still be a dichotomous judgment regarding the truth of a scientific prediction.

Box 2. How CIs could improve conventional practice: an example

A neuroscientific dispute lasting decades centered on whether long-term potentiation (LTP), the major proposed neural mechanism of learning and memory, was mediated by a presynaptic or a postsynaptic mechanism (for review, see Nicoll, 2017). There was universal agreement that LTP was induced by calcium entering the postsynaptic cell through NMDA receptors. Therefore, if LTP was expressed presynaptically, a "retrograde messenger" would have to go from the postsynaptic cell back to the presynaptic nerve terminal. Schuman and Madison (1991) and O'Dell et al. (1991) found that manipulating the nitric oxide, NO, system affected LTP and proposed the hypothesis that NO was the retrograde messenger in LTP. NO production seemed to be an "absolute requirement" for LTP (Hardingham et al., 2013). These findings quickly became controversial. Some laboratories found that manipulating NO systems affected LTP while others did not.

To see how the Estimation Approach could have impacted this area, I examined the figures in O'Dell et al. (1991). Their graphs plot the data as mean percentage of control response amplitudes, plus SEMs at sequential time points. As shown in their Figure 3, NO inhibitors markedly reduced mean LTP. However, rough measurements from the graphs suggest that the 95% CIs, calculated from estimated SEMs of ~25% about the mean responses, were rather large. The LTP magnitude (raw effect size) was a

mean ~120% increase over control levels, with a CI extending from ~71% to 169% ($n = 21$). Two different NO synthase inhibitors significantly reduced LTP to ~17% (CI = −24%, 58%; $n = 11$) and 42% (CI = −7%, 91%; $n = 10$), respectively. Hence, antagonists of the NO system did not block LTP in an all-or-none fashion, which might mean that the role of NO on synaptic transmission was complicated.

Subsequent demonstrations that LTP is not expressed as an increase in presynaptic transmitter release (Nicoll, 2017) indirectly ruled out the proposed role for NO in LTP. However, success in LTP production could depend on initial transmitter release probability (Larkman et al., 1992) and several presynaptic targets of NO have been reported (Hardingham et al., 2013). Possibly, under some conditions, transmitter release affected by NO contributed to the variability in the O'Dell et al. (1991) results and the ensuing controversy. A fuller appreciation of the uncertainty in the initial experiments, as conveyed by CIs, might have steered research into a more profitable direction earlier.

The Estimation Approach usefully warns against misinterpreting "statistical significance," as meaning "important," whereas, as noted above, it simply indicates a result is unlikely to occur by chance if H_o had been true. Likewise, the Estimation Approach emphasizes that a "failure to reject H_o" is not the same as "accepting H_o" and concluding the groups do not differ. Paying more attention to these concepts should improve standard neuroscience practices. Nevertheless, this would not call for abandoning significance testing and replacing it with the Estimation Approach. The impetus for getting rid of significance-testing comes primarily from philosophical aspects of the latter approach.

Shortcomings of the Estimation Approach as applied to neuroscience

Specific shortcomings of the Estimation Approach philosophy include: (1) it fails to distinguish between scientific and statistical hypotheses; (2) it emphasizes "quantitative," rather than "qualitative," research questions; (3) it dwells on cases in which single *p*-valued tests are said to provide definitive answers; (4) it does not recognize the extent to which neuroscience subfields differ from each other; and finally, (5) it relies too heavily on "preregistration" procedures to resolve controversies.

Quantitative versus qualitative (dichotomous) differences

One of the core principles of the Estimation Approach is that dichotomous—either/or, yes/no—thinking, asking, e.g., "whether or not an effect is present," is to be avoided. Instead, neuroscientists should ask, "to what extent" an effect is present. This approach does not acknowledge that either/or thinking is essential for science. Scientists must act and taking action requires making the dichotomous decision whether to act or not. Moreover, age-old values and practices of science are bound up in dichotomous thinking.

Truth/falsehood is a key dichotomy

According to Claude Bernard (Bernard, 1865/1957), "The truth must be the goal of our studies. Being satisfied by plausibility or likelihood is the true pitfall."

The ultimate goal of science is a true understanding of nature. Although scientific hypothesis testing is not the only way to do research (see below), it is a principal method and is pervasive. In this mode, the search for knowledge is conducted by advancing and testing scientific hypotheses and rejecting the false ones (Popper, 1959/2002; Platt, 1964). Decisions to accept or reject hypotheses are dichotomous. The tested and corroborated ones constitute our current body of scientific knowledge. Falsified hypotheses constitute valuable "negative information" in scientific knowledge. The distinction between falsified and corroborated is qualitative.

Quantitative conclusions are like "plausible" or "likely" outcomes and being satisfied with them is a "real pitfall." Such outcomes cannot be rigorously tested and falsified, but only found to be more-or-less-plausible or likely. Conclusions based on quantitative thinking mainly stimulate the formation of scientific hypotheses that must then be qualitatively tested.

Logically valid/invalid

Fallacies are arguments that "seem to be valid but [are] not..." (Bernard, 2020). Fallacies distort data interpretation and conclusions. Errors caused by fallacies corrupt the scientific literature, undermine credibility in science, create irreproducible results, and waste time and money in correcting them. Preventing fallacious arguments requires differentiating them qualitatively from valid ones. An approach that blurs crucial distinctions between valid and invalid conclusions, even indirectly, diminishes the force of scientific reasoning.

Action/inaction

Experimental science is an active endeavor. Scientists must do things that demand making dichotomous choices. Do you commit the time, effort, and resources to conducting a particular experiment or not? Such yes/no decisions depend, in turn, on other decisions. Is the current explanation for some phenomenon justified? If not, is there a feasible experiment to test it? Will doing so require obtaining a new knock-out mouse? You cannot arrange to obtain a knock-out mouse, "to an extent"; you either do it or not. Scientists continually confront such dichotomous choices. They require rational (even if conventional), objective grounds for making them. Significance testing procedures such as NHST serve this function.

Neuroscience is not a unitary field

Scientists use NHST for a variety of reasons. A leading statistician, Daniel Lakens (Lakens, 2021), argues that, properly used, NHST methods can be valid and valuable in making the dichotomous decisions that scientists must make. There are caveats. Scientific fields differ in many ways, including their degree of control over the variables they study, the maturity of the concepts they work with, and the rigor of the logic that holds their arguments together. In some fields, e.g., psychology, weakness in these areas have led to concerns that NHST testing procedures are overused (Scheel et al., 2021). The same critics acknowledge the utility of significance testing in biology and related fields. Subfields of neuroscience run the gamut from some that are akin to psychology, to others, e.g., structural neurobiology, in which tools and variables are extremely well defined and well controlled. Significance testing will, accordingly, be more necessary and justified in some neuroscience subfields than others. Also consider that the effect sizes characteristic of particular neuroscience-related areas can vary over two orders of magnitude depending on the area (Cumming and Calin-Jageman, 2017; pp.

178–179). Effect size is a critical parameter for calculation of statistical power (Lakens, 2013). Accordingly, variability in the magnitude of effect sizes ought to influence concerns about statistical power, yet this issue is sometimes overlooked by neuroscience critics. These general issues highlight a few of the drawbacks to the blanket proposal to replace NHST with the Estimation Approach throughout neuroscience.

One "definitive" test versus many corroborating tests

Descriptions of the Estimation Approach often analyze examples in which a single statistical test result is proposed to be decisive or "definitive" in establishing an important conclusion. For instance, Calin-Jageman and Cumming (2019) critique studies that report beneficial effects of caffeine on memory (Borota et al., 2014) or whether intranasal oxytocin enhances human trust behavior (Kosfeld et al., 2005). In the studies, drug-treated and drug-naive groups of people were compared. The major conclusions rested on single *p*-valued tests that detected marginally statistically significant positive treatment effects, with wide variability in the CIs. Possible effect sizes ranged from "negligible" to moderate. Follow-up studies largely failed to replicate the results. Calin-Jageman and Cumming (2019) remark that, "this trajectory from a single significant result to wide-spread acceptance of a categorical claim is the norm in our field." They do not specify "our field," but such a trajectory is not characteristic of all neuroscience.

In fact, in much laboratory-based experimental neuroscience, a single statistical test does not convey the main message of an investigation. A review of over 52 sequential articles in the *Journal of Neuroscience* (Alger, 2020), found that most (75%) were small laboratory investigations that used a variety of experimental approaches to test several predictions of one or more scientific hypotheses. A single *p*-value was never "widely accepted" as being definitive. Indeed, while positive results were generally reported, in many cases the multiple testing process led to the explicit rejection of alternative hypotheses, which is evidence that many investigators do take an appropriately critical approach to evaluating their data.

The Estimation Approach labels such laboratory studies as "exploratory," which implies that they are merely preliminary to a more definitive ("confirmatory") study. (Note: exploratory studies are not "pilot studies." Exploratory studies are small-scale investigations that are published, whereas pilot studies are informal trials within a laboratory and are unpublished.) The exploratory-confirmatory dichotomy (Wagenmakers et al., 2012) is widely recognized in psychology and social science, although some authorities in those areas consider it a "false dichotomy," that has mislead researchers (Scheel et al., 2021). Exploratory studies are considered to have low "evidential status" (Allen and Mehler, 2019), i.e., to carry less weight, than confirmatory studies, which are intended to be decisive tests of statistical hypotheses. It is argued by some that exploratory studies, which are often small and lack statistical power, should not be used to test hypotheses (Button et al., 2013; Calin-Jageman and Cumming, 2019). Yet exploratory studies play vital roles that cannot be filled by confirmatory studies (Kimmelman et al., 2014; Lakens, 2021).

In the realm of laboratory research in neuroscience, extensive scientific-hypothesis testing work is often on the cutting edge. Cellular, molecular, and genetic studies of LTP, for instance, have been going on for nearly 50 years and have enormously advanced the neuroscience of learning and memory, synaptic physiology, neuronal structure, and behavior. The great majority of such studies are multi-part tests of scientific hypotheses, conducted on small samples in small laboratories. To my knowledge, no large "confirmatory" studies have been conducted in LTP. Dismissing research on LTP as merely "exploratory" misses the point that, even when contentious, these studies are demonstrably pushing back the frontiers of neuroscience. The Estimation Approach philosophy cannot deal effectively with wide swaths of neuroscience research.

In broad terms, the Estimation Approach does not take into account the common strategy of employing multiple tests of scientific hypotheses. Multiple testing of a single hypothesis makes for more robust conclusions than single testing does. Consider an investigation to test the hypothesis that GABA is the transmitter at a particular synapse. Nowadays, investigators would typically combine electrophysiological, pharmacological, anatomic, and genetic experiments in testing such a hypothesis. For each method used, one or more significance tests is reported. No one test is given decisive weight. Instead, the overall conclusion of the paper emerges from an intellectual synthesis of all of the results.

As argued before (Alger, 2020), it is intuitively clear that the reliability of an investigation that tests several distinct predictions of a scientific hypothesis should be greater than that provided by any one test. As an analogy, consider independent games of chance offering individual odds of 1/5, 1/10, 1/12, 1/15, and 1/20 to win. The odds of winning any one game are not bad, but the odds of winning them all in succession are quite low: 1/180,000. This is because the joint probability of a collection of independent events is the product of the individual probabilities, the number we get from multiplying them together.

A *p*-value is a probability. Therefore similar, although not identical, reasoning can be applied to a collection of tests that yield *p*-values. It would be highly unlikely for a group of independent, *p*-valued tests of a hypothesis to come out significant by chance alone. Thus, the conclusion of a multipart test of a scientific hypothesis should be more robust than the results of any one experiment. Authors of such investigations essentially make this point in their Discussions. Currently, no quantitative method for combining disparate *p*-valued results is in general use, although such methods exist. The Fisher Method described by R. A. Fisher sums the lns of a collection of *p*-values of independent tests of a given hypothesis to yield an aggregate significance value for the collection (Alger, 2020). An alternative approach would combine results via a meta-analysis (Cumming and Calin-Jageman, 2017; Alger, 2020). Such a combined test could be useful in evaluating studies that test scientific hypotheses in neuroscience.

A combined significance test offers several advantages: (1) it diminishes the influence of any one *p*-value; (2) there is no "cutoff" significance level for inclusion of *p*-values; all *p*-values of whatever magnitude go into the calculation. This should decrease the unhealthy focus on specific *p*-values and associated questionable practices; (3) its proper use requires careful thinking about the logic of the experiments. The scientific hypothesis must be stated explicitly because only genuine, independent tests of predictions can be included; and (4) it would summarize the strength of the overall conclusions relating to the hypothesis.

The key point is that qualitative approaches to improving scientific-hypothesis testing, such as a combined significance test, fit naturally with conventional NHST, but are outside the preferred scope of the Estimation Approach with its emphasis on quantitative problems, single *p*-valued tests, and lack of distinctions between scientific and statistical hypotheses.

Additional concerns raised by the Estimation Approach philosophy

Replication (or "reproducibility") is a foundational virtue of science—true results should be reproducible. (Reproducible results are not necessarily true, of course, and thinking they are is a fallacy; Bernard, 2020.) The Estimation Approach holds up reproducibility of specific results as, perhaps, the highest scientific standard. However, neuroscience is exceedingly complex and there are countless legitimate reasons that a published finding might be irreproducible by another laboratory. The effects of biological sex of experimenters on rodent behavior (Sorge et al., 2014), or of a contaminant in a commercially supplied drug on synaptic transmission (Lafourcade et al., 2009) are among the innumerable subtle and unforeseeable reasons for irreproducibility. Therefore, it is not a simple matter to say just how the quest for reproducibility should come into the daily practice of neuroscience.

Initially, there is very little information in an irreproducible result. If study B fails to reproduce study A, it is unknown which is right and, of course, they might both be right (or wrong). To settle the issue, the Estimation Approach recommends more replication studies accompanied by a meta-analysis of the data. Ideally, there will be large, meticulously planned confirmatory studies that formally state their methods in advance and "preregister" them. Preregistration is basically a promise to carry out a study exactly as described and to report the results exactly as obtained. In return, approved preregistered studies are guaranteed publication of whatever results are obtained. The preregistration strategy can be invaluable in the right context (e.g., clinical trials). Preregistration, although not an integral component of the Estimation Approach, nevertheless epitomizes its primary concern with replication. Preregistration is not applicable throughout experimental neuroscience, however, and there are problems with the preregistration strategy itself.

Replication

An investigator whose laboratory pilot study has failed to reproduce a published result must decide whether to try to pinpoint the problem and pursue replication studies, or to set aside the irreproducible results, derive other predictions from the scientific hypothesis at stake and test them. Conducting a full-scale replication study may demand substantial time, effort, and money (Allen and Mehler, 2019). The intranasal oxytocin study (Kosfeld et al., 2005) has been the subject of numerous attempted replications and its findings have not, so far, been reliably reproduced (Nave et al., 2015). The failures have been attributed to everything from inadequate research practices (Calin-Jageman and Cumming, 2019) to a variety of "person-specific" variables, e.g., personality traits, genetic factors, etc. (Kurokawa, et al., 2021). It is an open question whether more progress has been made by investigations of other predictions of the oxytocin hypothesis, or by repeating the original protocol many times. It may be that replication studies serve more to determine the degree of reproducibility of published observations than to investigate the validity of the ideas they are connected to.

Preregistration

Preregistration will not work for laboratory scientists who are actively and continually engaged in a "dialogue" with nature (cf. Bernard, 1865/1957). That is, where scientists test predictions that falsify the hypothesis, come up with another hypothesis, or make an unanticipated observation, and pursue the new line of research, etc. Many such factors militate against preregistration in highly competitive, fast-moving research fields. Preregistration would not have been appropriate for the labs racing to figure out the mechanisms of LTP, for instance. In such fields, too little is known to be able to specify in advance all of the relevant variables or the likely outcomes of the manipulations to make preregistration a viable option. The risk that a clever, original idea could become known to one's competitors during preregistration review may also be a disincentive. More generally, the constraints of preregistration will make this method unappealing to a large group of neuroscientists.

While preregistration is intended to eliminate investigator bias and certain questionable practices, the preregistration system itself is not immune to abuse (Yamada, 2018; Allen and Mehler, 2019). One scheme is called PARKing (preregistration after results are known; Yamada, 2018). In PARKing, investigators "propose" to do a well-planned and tightly-controlled study that they have in fact already done. Knowing how their experiments turned out, they can describe their methods, "expected results," analytic procedures, etc., for preregistration review. Once their sham proposal has been accepted, the investigators report their previously gathered data with publication guaranteed. This integrity of preregistration can be further undermined if investigators use selective reporting of data, i.e., withholding data that does not support their ideas. After a study is reviewed and registered, there is no control over the data that are reported. Selective reporting is of course not a unique problem for preregistration, but its possibility shows how a system that is expressly intended to guarantee the trustworthiness of results may still be undermined.

Concern has been raised about the increasing number of negative results from preregistered studies (Warren, 2018). While this may reflect honest data reporting, it could also reflect a ploy by investigators wondering what to do with negative data that might be hard to publish via traditional peer-review. Investigators could "propose" the work for preregistration, without disclosing that they have already done it. Once their proposal is accepted and publication is guaranteed, their problem is solved.

Most disappointingly, perhaps, for preregistration advocates, Claesen et al. (2021) found that, of 27 properly vetted, preregistered studies, 25 of them did not fully adhere to the promised protocol. Of the 25, only one fully disclosed all of its deviations, and nine failed to disclose any of their deviations. These violations of trust, together with the questionable practices mentioned above, undercut the purpose and benefits of preregistration.

Preregistration is accepted as a high standard for ensuring the reliability of published results. The pitfalls associated with preregistration are relevant here because of the central place that replication holds in the Estimation Approach philosophy. Other less rigorous forms of replication studies must have the same potential for abuse that preregistration has.

Discussion

The Estimation Approach recommends that neuroscience "countenance uncertainty," and advocates large, well-controlled replication studies to reduce uncertainty. Its goal is to determine the extent to which variability influences final conclusions. In contrast, the conventional approach, aided by significance testing, uses explanatory scientific hypotheses and attempts to eliminate variability by correctly accounting for and explaining the disparate factors and arriving at a true understanding of nature.

Plainly, neuroscience needs standards of truth and ways of evaluating its results to know whether its hypotheses have been corroborated or falsified. It needs clear standards, expressed in dichotomous thinking, for separating truth and falsehood; fallacious from valid reasoning; the logical relationship between

scientific hypotheses and predictions, etc. Neuroscience will benefit from adopting practical elements of the Estimation Approach and correcting the misuse of NHST-based significance testing (Lakens, 2021). Yet much will be lost if neuroscientists become less attentive to the rigors of genuine scientific hypothesis testing and the reasoning that supports and sustains it.

In an essay evaluating estimation statistics in *eNeuro* papers, Christophe Bernard (Bernard, 2021) concludes, "However, it is important to keep in mind that estimation statistics and significance testing can complement each other. They provide different types of information. One just has to know how to use these techniques and be aware of their limitations."

Neuroscience can improve its current procedures by taking advantage of both approaches without "replacing" one with the other.

References

Alger BE (2019) Defense of the scientific hypothesis: from reproducibility crisis to big data. New York: Oxford University Press.

Alger BE (2020) Scientific hypothesis-testing strengthens neuroscience research. eNeuro 7:ENEURO.0357-19.2020.

Allen C, Mehler DMA (2019) Open science challenges, benefits and tips in early career and beyond. PLoS Biol 17:e3000246.

Alvarez LW, Alvarez W, Asaro F, Michel HV (1980) Extraterrestrial cause for the cretaceous–tertiary extinction. Science 208:1095–1108.

Bernard C (1865/1957) An introduction to the study of experimental medicine, translated by Henry Copley Greene. New York: Dover Publ, Inc.

Bernard C (2020) On fallacies in neuroscience. eNeuro 7:ENEURO.0491-20.2020–3.

Bernard C (2021) Estimation statistics, one year later. eNeuro 8: ENEURO.0091-21.2021.

Borota D, Murray E, Keceli G, Chang A, Watabe JM, Ly M, Toscano JP, Yassa MA (2014) Post-study caffeine administration enhances memory consolidation in humans. Nat Neurosci 17:201–203.

Button KS, Ioannidis J, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376.

Calin-Jageman RJ, Cumming G (2019) Estimation for better inference in neuroscience. eNeuro 6:ENEURO.0205-19.2019–11.

Cumming G, Calin-Jageman R (2017) Introduction to the new statistics: estimation, open science, and beyond. New York: Routledge.

Claesen A, Gomes S, Tuerlinckx F Vanpaemel (2021) Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies. Roy Soc Open Sci 8:211037.

Gigerenzer G (2008) Rationality for mortals: how people cope with uncertainty, Chapter 11. New York: Oxford University Press.

Gigerenzer G, Swijtink Z, Porter T, Daston L, Beatty J, Krüger L (1989) The empire of chance: how probabiotiy changed science and everyday life. New York: Cambridge University Press.

Hardingham N, Dachtler J, Fox K (2013) The role of nitric oxide in synaptic plasticity and homeostasis. Front Cell Neurosci 7:190.

Johnson-Laird PN (2010) Mental models and human reasoning. Proc Natl Acad Sci U S A 107:18243–18250.

Kahneman D (2011) Thinking, fast and slow, Chapter 1. New York: Farrar, Straus and Giroux.

Kimmelman J, Mogil JS, Dirnagl U (2014) Distinguishing between exploratory and confirmatory preclinical research will improve translation. PLoS Biol 12:e1001863.

Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E (2005) Oxytocin increases trust in humans. Nature 435:673–676.

Kurokawa H, Kinari Y, Okudaira H, Tsubouchi K, Sai Y, Kikuchi M, Higashida H, Ohtake F (2021) Oxytocin-trust link in oxytocin-sensitive participants and those without autistic traits. Front Neurosci 15:659737.

Lafourcade C, Zhang L, Alger BE (2009) Novel mGluR- and CB1R-independent suppression of GABA release caused by a contaminant of the group I metabotropic glutamate receptor agonist, DHPG. PLoS One 4: e6122.

Lakens D (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 4:863.

Lakens D (2021) The practical alternative to the p value is the correctly used p value. Perspect Psychol Sci 16:639–648.

Larkman A, Hannay T, Stratford K, Jack J (1992) Presynaptic release probability influences the locus of long-term potentiation. Nature 360:70–73. 10.1038/360070a0

Nave G, Camerer C, McCullough M (2015) Does oxytocin increase trust in humans? A critical review of research. Perspect Psychol Sci 10:772–789.

Nicoll RA (2017) A brief history of long-term potentiation. Neuron 93: 281–290.

O'Dell TJ, Hawkins RD, Kandel ER, Arancio O (1991) Tests of the roles of two diffusible substances in long-term potentiation: evidence for nitric oxide as a possible early retrograde messenger. Proc Natl Acad Sci U S A 88:11285–11289.

Perezgonzalez JD (2015a) Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Front Psychol 6:223.

Perezgonzalez JD (2015b) Confidence intervals and tests are two sides of the same research question. Front Psychol 6:34.

Platt JR (1964) Strong Inference: certain systematic methods of scientific thinking may produce much more rapid progress than others. Science 146:347–353.

Popper K (1959/2002) The logic of scientific discovery. New York: Routledge Classics.

Scheel A, Tiokhin L, Isager P, Lakens D (2021) Why hypothesis testers should spend less time testing hypotheses. Perspect Psychol Sci 16:744–755.

Schuman EM, Madison DV (1991) A requirement for the intercellular messenger nitric oxide in long-term potentiation. Science 254:1503–1506.

Sorge RE, et al. (2014) Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nat Methods 11:629–632.

Szucs D, Ioannidis JPA (2017) When null hypothesis significance testing is unsuitable for Research: a reassessment. Front Hum Neurosci 11:390.

Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL, Kievit RA (2012) An agenda for purely confirmatory research. Perspect Psychol Sci 7:632–638.

Warren M (2018) First analysis of 'pre-registered' studies shows sharp rise in null findings. Logging hypotheses and protocols before performing research seems to work as intended: to reduce publication bias for positive results. Nature. Available at https://www.nature.com/articles/d41586-018-07118-1.

Yamada Y (2018) How to crack pre-registration: toward transparent and open science. Front Psychol 9:1831.