History, Teaching, and Public Awareness

Scientific Hypothesis-Testing Strengthens Neuroscience Research

Bradley E. Alger

https://doi.org/10.1523/ENEURO.0357-19.2020

Department of Physiology and Program in Neuroscience, University of Maryland School of Medicine, Baltimore, MD, 21201

Abstract

Science needs to understand the strength of its findings. This essay considers the evaluation of studies that test scientific (not statistical) hypotheses. A scientific hypothesis is a putative explanation for an observation or phenomenon; it makes (or "entails") testable predictions that must be true if the hypothesis is true and that lead to its rejection if they are false. The question is, "how should we judge the strength of a hypothesis that passes a series of experimental tests?" This question is especially relevant in view of the "reproducibility crisis" that is the cause of great unease. Reproducibility is said to be a dire problem because major neuroscience conclusions supposedly rest entirely on the outcomes of single, *p* valued statistical tests. To investigate this concern, I propose to (1) ask whether neuroscience typically does base major conclusions on single tests; (2) discuss the advantages of testing multiple predictions to evaluate a hypothesis; and (3) review ways in which multiple outcomes can be combined to assess the overall strength of a project that tests multiple predictions of one hypothesis. I argue that scientific hypothesis testing in general, and combining the results of several experiments in particular, may justify placing greater confidence in multiple-testing procedures than in other ways of conducting science.

Key words: estimation statistics; hypothesis testing; meta-analysis; p value; reproducibility crisis; statistical hypothesis

Significance Statement

The statistical *p* value is commonly used to express the significance of research findings. But a single *p* value cannot meaningfully represent a study involving multiple tests of a given hypothesis. I report a survey that confirms that a large fraction of neuroscience work published in *The Journal of Neuroscience* does involve multiple-testing procedures. As readers, we normally evaluate the strength of a hypothesis-testing study by "combining," in an ill-defined intellectual way, the outcomes of multiple experiments that test it. We assume that conclusions that are supported by the combination of multiple outcomes are likely to be stronger and more reliable than those that rest on single outcomes. Yet there is no standard, objective process for taking multiple outcomes into account when evaluating such studies. Here, I propose to adapt methods normally used in meta-analysis across studies to help rationalize this process. This approach offers many direct and indirect benefits for neuroscientists' thinking habits and communication practices.

Introduction

Scientists are not always clear about the reasoning that we use to conduct, communicate, and draw conclusions from our work, and this can have adverse consequences. A lack of clarity causes difficulties and wastes time in evaluating and weighing the strength of each others' reports. I suggest that these problems have also influenced perceptions about the "reproducibility crisis" that science is reportedly suffering. Concern about the reliability of

Received September 4, 2019; accepted June 9, 2020; First published July 8, 2020.
The author declares no competing financial interests.
Author contributions: B.E.A. conceived of and wrote the paper.

Acknowledgements: I thank Asaf Keller for his comments on a draft of this manuscript. I also thank the two reviewers for their useful critical comments that encouraged me to develop the arguments herein more fully.

science has reached the highest levels of the NIH (Collins and Tabak, 2014) and numerous other forums (Landis et al., 2012; Task Force on Reproducibility, American Society for Cell Biology, 2014). Many of the concerns stem from portrayals of science like that offered by the statistician, John Ioannidis, who argues that "most published research findings are false," especially in biomedical science (Ioannidis, 2005). He states "... that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of *claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p value* $<$*0.05*" (italics added).

He continues, "Research is not most appropriately represented and summarized by *p* values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on *p* values."

Additional concerns are added by Katherine Button and colleagues (Button et al., 2013), who conclude that much experimental science, such as neuroscience, is fatally flawed because its claims are based on statistical tests that are "underpowered," largely because of small experimental group sizes. Statistical power is essentially the ability of a test to identify a real effect when it exists. Power is defined as "1-β," where β is the probability of failing to reject the null hypothesis when it should be rejected. Statistical power varies from 0 to 1 and values of $\geq$0.8 are considered "good." Button et al. (2013) calculate that the typical power of a neuroscience study is $\sim$0.2, i.e., quite low.

However, these serious concerns arise from broad assumptions that may not be universally applicable. Biomedical science encompasses many experimental approaches, and not all are equally susceptible to the criticisms. Projects in which multiple tests are performed to arrive at conclusions are expected to be more reliable than those in which one test is considered decisive. To the extent that basic ("pre-clinical") biomedical science consists of scientific hypothesis testing, in which a given hypothesis is subjected to many tests of its predictions, it may be more reliable than other forms of research.

It is critical here to distinguish between a "scientific hypothesis" and a "statistical hypothesis," which are very different concepts (Alger, 2019; chapter 5). A scientific hypothesis is a putative conceptual explanation for an observation or phenomenon; it makes predictions that could, in principle, falsify it. A statistical hypothesis is simply a mathematical procedure (often part of Null Hypothesis Significance Testing, NHST) that is conducted as part of a broader examination of a scientific hypothesis (Alger, 2019, p. 133). However, scientific hypotheses can be tested without using NHST methods, and, vice versa, NHST methods are often used to compare groups when no scientific hypothesis is being tested. Unless noted otherwise, in this essay "hypothesis" and "hypothesis testing" refer to scientific hypotheses.

To appreciate many of the arguments of Ioannidis, Button, and their colleagues, it is necessary to understand their concept of positive predictive value (PPV; see equation below). This is a statistical construct that is used to estimate the likelihood of reproducing a given result. PPV is defined as "the post-study probability that [the experimental result] is true" (Button et al., 2013). In addition to the "pre-study odds" of a result's being correct, the PPV is heavily dependent on the *p* value of the result and the statistical power of the test. It follows from the statisticians' assumptions about hypotheses and neuroscience practices that calculated PPVs for neuroscience research are low (Button et al., 2013). On the other hand, PPVs could be higher if their assumptions did not apply. I stress that I am not advocating for the use of the PPV, which can be criticized on technical grounds, but must refer to it to examine the statistical arguments that suggest deficiencies in neuroscience.

To look into the first assumption, that neuroscience typically bases many important conclusions on single *p* valued tests, I analyze papers published in consecutive issues of *The Journal of Neuroscience* during 2018. For the second assumption, I review elementary joint probability reasoning that indicates that the odds of obtaining a group of experimental outcomes by chance alone are generally extremely small. This notion is the foundation of the argument that conclusions derived from multiple experiments should be more secure those derived from one test. However, there is currently no standard way of objectively evaluating the significance of a collection of results. As a step in this direction, I use two very different procedures, Fisher's method of combining results and meta-analysis of effect sizes (Cummings and Calin-Jageman, 2017) measured by Cohen's *d*, which have not, as far as I know, been applied to the problem of combining outcomes in the way that we need. Finally, in Discussion, I suggest ways in which combining methods such as these can improve how we assess and communicate scientific findings.

Materials and Methods

To gauge the applicability of the statistical criticisms to typical neuroscience research, I classified all Research Articles that appeared in the first three issues of *The Journal of Neuroscience* in 2018 according my interpretation of the scientific "modes" they represented, i.e., "hypothesis testing," "questioning," etc., because these modes have different standards for acceptable evidence. Because my focus is on hypothesis testing, I did a pdf search of each article for "hypoth" (excluding references to "statistical" hypothesis and cases where "hypothesis" was used incorrectly as a synonym for "prediction"). I also searched "predict" and "model" (which was counted when used as a synonym for "hypothesis" and excluded when it referred to "animal models," "model systems," etc.) and checked the contexts in

Correspondence should be addressed to Bradley E. Alger at balgerlab@gmail.com.

https://doi.org/10.1523/ENEURO.0357-19.2020

Copyright © 2020 Alger et al.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license, which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

which the words appeared. In judging how to categorize a paper, I read its Abstract, Significance Statement, and as much of the text, figure legends, and Discussion as necessary to understand its aims and see how its conclusions were reached. Each paper was classified as "hypothesis-based," "discovery science," (identifying and characterizing the elements of an area), "questioning" (a series of related questions not evidently focused on a hypothesis), or "computational-modeling" (where the major focus was on a computer model, and empirical issues were secondary).

I looked not only at what the authors said about their investigation, i.e., whether they stated directly that they were testing a hypothesis or not, but what they actually did. As a general observation, scientific authors are inconsistent in their use of "hypothesis," and they often omit the word even when it is obvious that they are testing a hypothesis. When the authors assumed that a phenomenon had a specific explanation, then conducted experimental tests of logical predictions of that explanation, and drew a final conclusion related to the likely validity of the original explanation, I counted it as implicitly based on a hypothesis even if the words "hypothesis," "prediction," etc. never appeared. For all hypothesis-testing papers, I counted the number of experimental manipulations that tested the main hypothesis, even if there were one or more subsidiary hypotheses (see example in text). If a paper did not actually test predictions of a potential explanation, then I categorized it as "questioning" or "discovery" science. While my strategy was unavoidably subjective, the majority of classifications would probably be uncontroversial and disagreements unlikely to change the overall trends substantially.

To illustrate use of the statistical combining methods, I analyzed the paper by Cen et al. (2018), as suggested by a reviewer of the present article. The authors made multiple comparisons with ANOVAs followed by Bonferroni *post hoc* tests; however, to make my analysis more transparent, I measured means and SEMs from their figures and conducted two-tailed *t* tests. When more than one experimental group was compared with the same standard control, I took only the first measurement to avoid possible complications of non-independent *p* values. I used the *p* values to calculate the combined mean significance level for all of the tests according to Fisher's method (see below). This is an extremely conservative approach, as including the additional tests would have further increased the significance of the combined test.

For the meta-analysis of the Cohen's *d* parameter (Cummings and Calin-Jageman, 2017; p. 239), I calculated effect sizes on the same means and SEMs from which *p* values were obtained for the Fisher's method example. I determined Cohen's *d* using an on-line calculator (https://www.socscistatistics.com/effectsize/default3.aspx) and estimated statistical power with G*-Power (http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html). I then conducted a random-effects meta-analysis on the Cohen's *d* values with Exploratory Software for Confidence Interval (ESCI) software, which is available at https://thenewstatistics.com/itns/esci/ (Cummings and Calin-Jageman, 2017).

Results

Of the total of 52 Research Articles in the first three issues of *The Journal of Neuroscience* in 2018, I classified 39 (75%) as hypothesis-based, with 19 "explicitly" and 20 "implicitly" testing one or more hypotheses. Of the remaining 13 papers, eight appeared to be "question" or "discovery" based, and five were primarily computer-modeling studies that included a few experiments (see Table 1). Because the premises and goals of the non-hypothesis testing kinds of studies are fundamentally distinct from hypothesis-testing studies (Alger, 2019; chp. 4), the same standards cannot be used to evaluate them, and I did not examine these papers further.

None of the papers based its major conclusion on a single test. In fact, the overarching conclusion of each hypothesis-based investigation was supported by approximately seven experiments (6.9 ± 1.57, mean ± SD, $n = 39$) that tested multiple predictions of the central hypothesis. In 20 papers, at least one (one to three) alternative hypothesis was directly mentioned. Typically (27/39), the experimental tests were "consistent" with the overall hypothesis, while in 19 papers, at least one hypothesis was explicitly falsified or ruled out. These results replicate previous findings (Alger, 2019; chapter 9).

As noted earlier, some science criticism rests on the concept that major scientific conclusions rest on the outcome of a single *p* valued test. Indeed, there are circumstances in which the outcome of a single test is intended to be decisive, for instance, in clinical trials of drugs where we need to know whether the drugs are safe and effective or not. Nevertheless, as the preceding analysis showed, the research published in *The Journal of Neuroscience* is not primarily of this kind. Moreover, we intuitively expect conclusions bolstered by several lines of evidence to be more secure than those resting on just one. Simple statistical principles quantify this intuition.

Provided that individual events are truly independent—the occurrence of one does not affect the occurrence of the other and the events are not correlated—then the rule is to multiply their probabilities to get the probability of the joint, or compound, event in which all of the individual events occur together or sequentially. Consider five games of chance with probabilities of winning of 1/5, 1/15, 1/20, 1/6, and 1/10. While the odds of winning any single game are not very small, if you saw someone step up and win all five in a row, you might well suspect that he was a cheat, because the odds of doing that are 1/90,000.

The same general reasoning applies to the case in which several independent experimental predictions of a given hypothesis are tested. If the hypothesis is that GABA is the neurotransmitter at a given synapse, then we could use different groups of animals, experimental preparations, etc. and test five independent predictions: that synaptic stimulation will evoke an IPSP; chemically distinct pharmacological agents will mimic and block the IPSP; immunostaining for the GABA-synthetic enzyme will be found in the pre-synaptic nerve terminal; the IPSP will not occur in a GABA receptor knock-out animal, etc. The experiments test completely independent predictions of the same hypothesis, hence the chance probability of obtaining five significant outcomes that are consistent with it by random chance alone must be much lower than that of obtaining any one of them. If the tests were done at $p \leq 0.05$, the chances would be $\leq(0.05)^5$ or $\leq 3.13^{-7}$ that they would all just happen to be consistent with the hypothesis. Accordingly, we feel that a hypothesis that has passed many tests to be on much firmer ground than if it had passed only one test. Note, however, that the product of a group of p values is just a number; it is not itself a significance level.

It can be difficult for readers to tease the crucial information out of scientific papers as they are currently written. Not only is the work intrinsically complicated, but papers are often not written to maximize clarity. A common obstacle to good communication is the tendency of scientific papers to omit a direct statement of the hypotheses that are being tested, which is an acute problem in papers overflowing with data and significance tests. An ancillary objective of my proposal for analysis is to encourage authors to be more straightforward in laying out the logic of their work. It may be instructive to see how a complex paper can be analyzed.

As an example, I used the paper of Cen et al. (2018). Although the paper reports a total of 114 p values, they do not all factor equally in the analysis. The first step is to see how the experiments are organized. The authors state that their main hypothesis is that N-cadherin, regulated by PKD1, promotes functional synapse formation in the rodent brain. It appears that the data in the first two figures of the paper provide the critical tests of this hypothesis. These figures include 43 statistical comparisons, many of which were controls to ensure measurement validity, or which did not critically test the hypothesis, e.g., 18 tests of spine area or miniature synaptic amplitude were supportive, but not critical. I omitted them, as well as multiple comparisons made to the same control group to avoid the possibility of correlations among p values. For instance, if an effect was increased by PKD1 overexpression (OE) and reduced by dominant negative (DN) PKD1, I counted only the increase, as both tests used the same vector-treated control group. In the end, six unique comparisons tested crucial, independent, non-redundant predictions (shown in Figs. 2*A2*,*B2*,*D2*,*E2*, 3*B2*,*C2* of Cen et al., 2018). I emphasize that this exercise is merely intended to illustrate the combining methods; the ultimate aim is to encourage authors to explain and justify their decisions about including or excluding certain tests in their analyses.

Cen et al. (2018) test the following predictions of their main hypothesis with a variety of morphologic and electrophysiological methods:

(1) N-cadherin directly interacts with PKD1.Test: GST pull-down.

(2) N-cadherin and PKD1 will co-localize to the synaptic region. Test: immunofluorescence images.

Predictions (1) and (2) are descriptive, i.e., non-quantified; other predictions are tested quantitatively.

(3) PKD1 increases synapse formation. Tests (Fig. 2*A2*): OE of hPKD1 increases spine density and area ($p < 0.001$ for both); DN-hPKD1 decreases spine density and area ($p < 0.001$ for both).

(4) PKD1 increases synaptic transmission. Tests (Fig. 2*B2*): OE of hPKD1 increases mEPSC frequency ($p < 0.006$) but not amplitude; DN-hPKD1 decreases mEPSC frequency ($p < 0.002$) but not amplitude.

(5) PKD1 acts upstream of N-cadherin on synapse formation and synaptic transmission. Tests (Fig. 3*B2*): DN-hPKD-induced reductions of spine density and area are rescued by OE of N-cadherin ($p < 0.001$ for both). DN-hPKD1-induced reduction in mEPSC frequency is rescued by OE of N-cadherin ($p < 0.001$ for both).

These are the key predictions the main hypothesis: their falsification would have called for the rejection of the hypothesis in its present form. The organization of these tests of the main hypothesis is illustrated in Figure 1.

Cen et al. (2018) go on to identify specific sites on N-cadherin that PKD1 binds and phosphorylates and they test the associated hypothesis that these sites are critical for the actions of PKD1 on N-cadherin. They next investigate β-catenin as a binding partner for N-cadherin and test the hypothesis that this binding is promoted by PKD1. While these subsidiary hypotheses and their tests clearly complement and extend the main hypothesis, they are distinct from it and must be analyzed separately. Whether falsified or supported, the outcomes of testing them would not affect the conclusion of the main hypothesis. The relationships among the main hypothesis and other hypotheses are shown in Figure 2. Note that Cen et al. (2018) is unusually intricate, although not unique; the diagrams of most papers will not be nearly as complicated as Figures 1, 2.

Basic probability considerations imply that the odds of getting significant values for all six critical tests in Cen et al. (2018) by chance alone are extremely tiny; however, as mentioned, the product of a group of p values is not a significance level. R.A. Fisher introduced a method for converting a group of independent p values that all test a given hypothesis into a single parameter that can be used in a significance test (Fisher, 1925; Winkler et al., 2016; see also Fisher's combined probability test https://en.wikipedia.org/wiki/Fisher's_method; https://en.wikipedia.org/wiki/Extensions_of_Fisher's_method). For convenience, I will call this parameter "p_{FM}" because it is not a conventional p value. Fisher's combined test is used in meta-analyses of multiple replications of the same experiment across a variety of conditions or laboratories but has not, to my knowledge, been used to evaluate a collection of tests of a single scientific hypothesis. Fisher's test is:

$$\chi^2 = -2\sum_{i=1}^{k} \ln(p_i),$$

where p_i is the p value of the ith test and there are k tests in all. The sum of the natural logarithms (*ln*) of the p values, multiplied by -2, is a χ^2 variable with 2*k* degrees of freedom and can be evaluated via a table of critical values for the χ^2 distribution (for derivation of Fisher's test equation, see: https://brainder.org/2012/05/11/the-logic-of-the-fisher-method-to-combine-p-values/). Applying Fisher's test to Cen et al.'s major hypothesis (k = 6; df = 12), yields

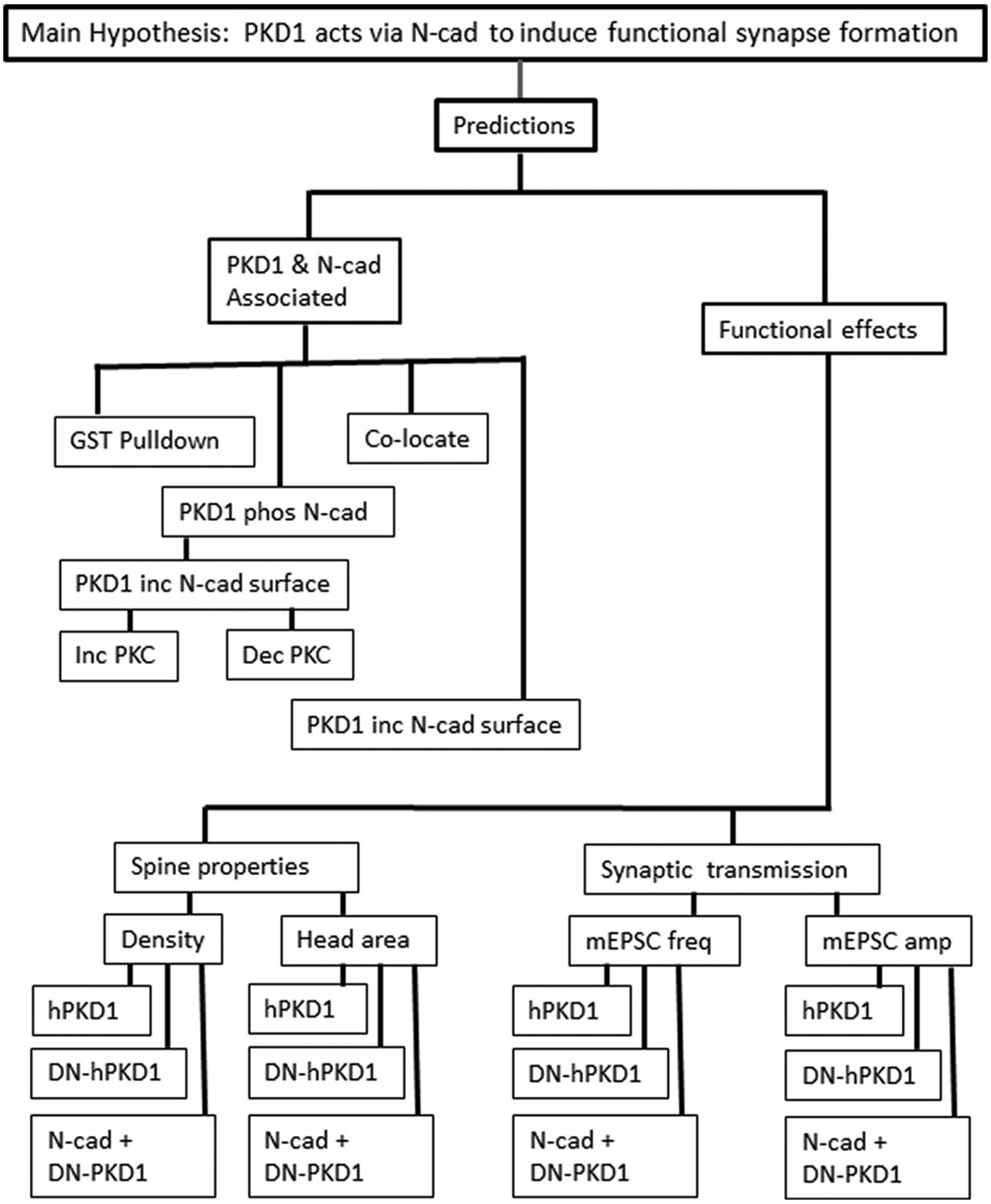


Figure 1. Diagram of the main hypothesis and predictions of Cen et al. (2018). The solid lines connect the hypothesis and the logical predictions tested. This diagram omits experimental controls tests that primarily validate techniques, include non-independent *p* values, or add useful but non-essential information. The main hypothesis predicts that PKD1 associates directly with N-cadherin, and that PKD1 and N-cadherin jointly affect synaptic development in a variety of structural and physiological ways. Separate groups of experiments test these predictions.

$$X^2 = 35.95 \text{ or } p < 0.001.$$

In other words, the probability, p_{FM}, of getting their collection of *p* values by chance alone is <0.001, and therefore, we may be justified having confidence in the conclusion. Fisher's method, or a similar test, does not add any new element but gives an objective estimate of the significance of the combined results. (Note here that the mathematical transformation involved can yield results that differ qualitatively from simply multiplying the *p* values.) I stress that Fisher's method is markedly affected by any but the most minor correlations (i.e., $r > 0.1$) among *p* values; notable correlations among these values will cause p_{FM} to be much lower (i.e., more extreme) than the actual significance value (Alves and Yu, 2014; Poole et al., 2016). Careful experimental design is required to ensure the independence of the tests to be combined.

Fisher's method is only one of a number of procedures for combining test results. To illustrate an alternative approach, I re-worked the assessment of Cen et al. (2018) as a meta-analysis (see Borenstein et al., 2007; Cummings and Calin-Jageman, 2017) of the effect sizes, defined by Cohen's *d*, of the same predictions. Cohen's *d* is a normalized, dimensionless measure of the mean difference between control and experimental values. I treated each prediction of the main hypothesis in Cen et al. (2018) as a

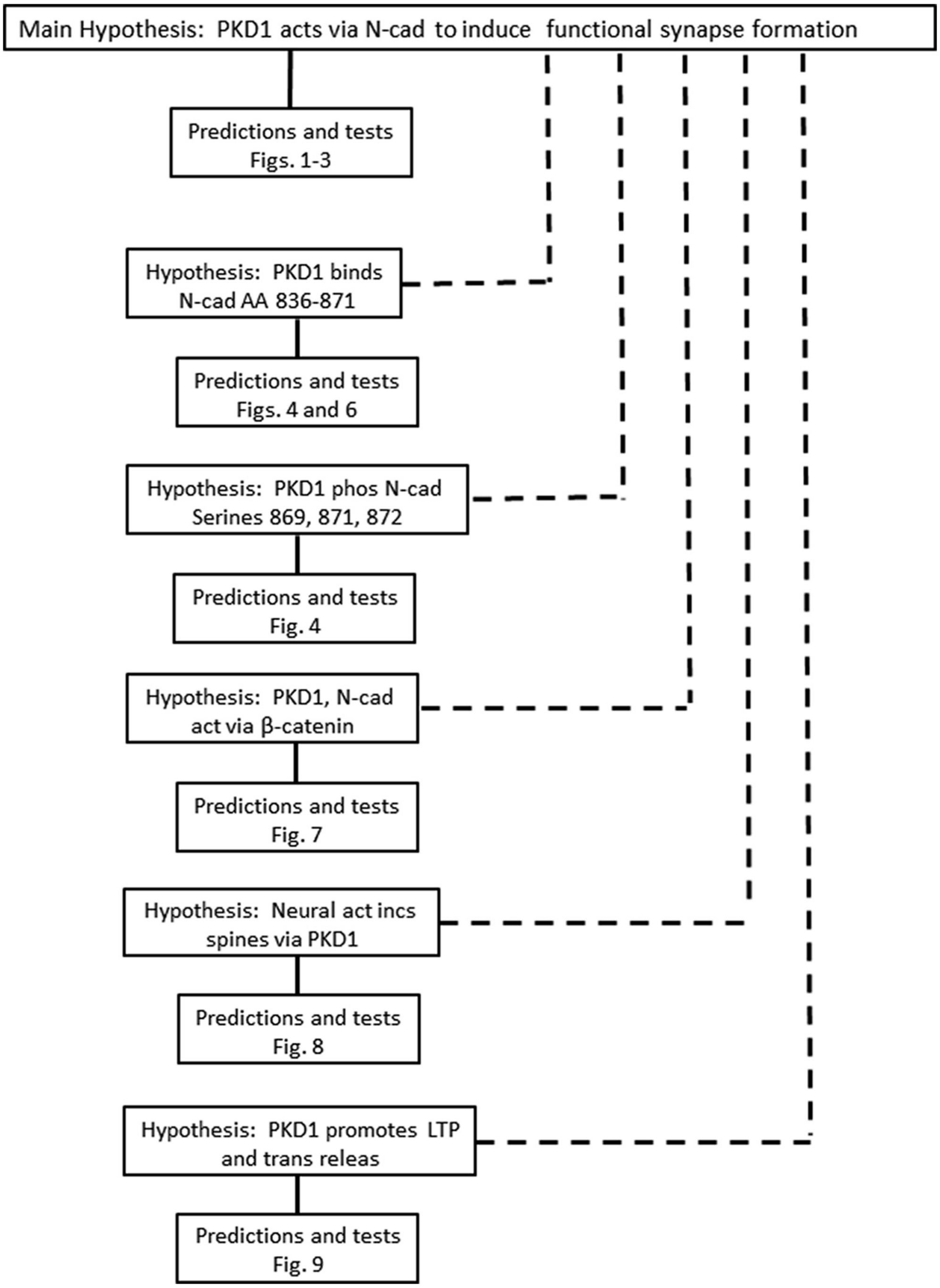
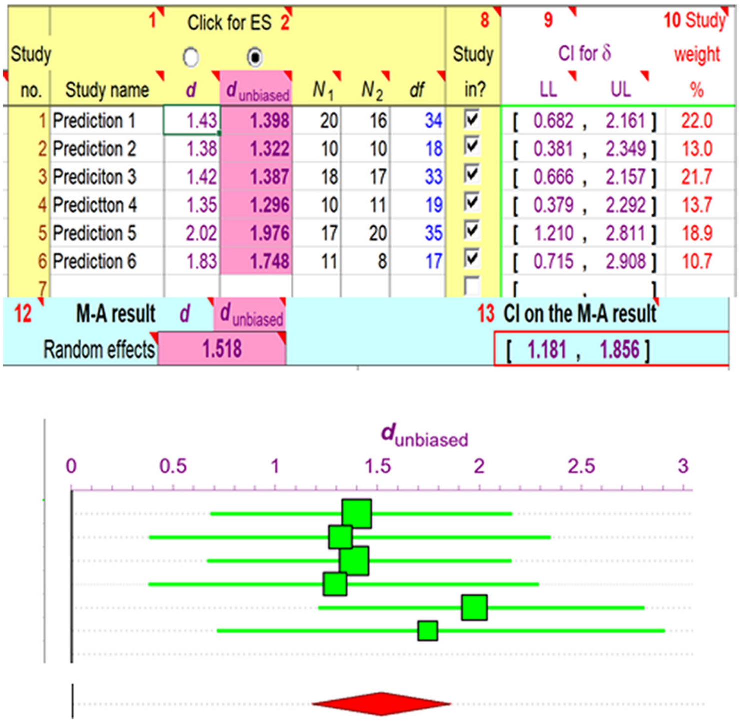

Main Hypothesis: PKD1 acts via N-cad to induce functional synapse formation
Predictions and tests Figs. 1-3
Hypothesis: PKD1 binds N-cad AA 836-871
Predictions and tests Figs. 4 and 6
Hypothesis: PKD1 phos N-cad Serines 869, 871, 872
Predictions and tests Fig. 4
Hypothesis: PKD1, N-cad act via β-catenin
Predictions and tests Fig. 7
Hypothesis: Neural act incs spines via PKD1
Predictions and tests Fig. 8
Hypothesis: PKD1 promotes LTP and trans releas
Predictions and tests Fig. 9

Figure 2. Diagram of the logical structure of Cen et al. (2018). The paper reports several distinct groups of experiments. One group tests the main hypothesis and others test subsidiary hypotheses that are complementary to the main one but are not a necessary part of it. Connections between hypotheses and predictions that are logically necessary are indicated by solid lines; dotted lines indicate complementary, but not mandatory, connections. Falsification of the logically-necessary predictions would call for rejection of the hypothesis in its present form; falsification of any of the subsidiary hypothesis would not affect the truth of the main hypothesis. The figure numbers in the boxes identify the source of major data in Cen et al., 2018 that were used to test the indicated hypothesis.

two-sample independent comparisons test, determined Cohen's *d* for each comparison, and conducted a random-effects meta-analysis (see Materials and Methods). Figure 3 shows effect sizes together with their 95% confidence intervals for each individual test, plus the calculated group mean effect size (1.518) and its confidence interval (1.181, 1.856). Effect sizes of 0.8 and 1.2 are considered "large" and "very large," respectively, hence, an effect size of 1.518 having a 95% confidence interval well above zero is quite impressive and reinforces the conclusion reached by Fisher's method, namely, that Cen et al.'s experimental tests strongly corroborate their main hypothesis.



Study	1	Click for ES	2				8	9		10 Study
		○	●				Study	CI for δ		weight
no.	Study name	*d*	$d_{unbiased}$	N_1	N_2	*df*	in?	LL	UL	%
1	Prediction 1	1.43	1.398	20	16	34	☑	[0.682 ,	2.161]	22.0
2	Prediction 2	1.38	1.322	10	10	18	☑	[0.381 ,	2.349]	13.0
3	Prediciton 3	1.42	1.387	18	17	33	☑	[0.666 ,	2.157]	21.7
4	Predictton 4	1.35	1.296	10	11	19	☑	[0.379 ,	2.292]	13.7
5	Prediction 5	2.02	1.976	17	20	35	☑	[1.210 ,	2.811]	18.9
6	Prediction 6	1.83	1.748	11	8	17	☑	[0.715 ,	2.908]	10.7
7							☐	[,	]	
12	M-A result	*d*	$d_{unbiased}$				13	CI on the M-A result		
	Random effects	1.518						[1.181 ,	1.856]	

Figure 3. Meta-analysis of the effect sizes observed in the primary tests of the main hypothesis of Cen et al. (2018; *n* = 6; shown in Fig. 1). I obtained effect sizes by measuring the published figures and calculated Cohen's *d* values with an on-line calculator: https://www.socscistatistics.com/effectsize/default3.aspx. Analysis and graphic display (screenshot) were done with ESCI (free at https://thenewstatistics.com/itns/esci/). Top panel shows individual effect sizes (corrected, $d_{unbiased}$) for the tendency of small samples to overestimate true effect sizes (see Cummings and Calin-Jageman, 2017; pp 176–177), *N*s and degrees of freedom (df) of samples compared, together with confidence intervals (CIs) of effect sizes and relative weights (generated by ESCI and based mainly on sample size) that were assigned to each sample. Upper panel also shows mean effect size for random effects model and CI for mean. Bottom panel shows individual means (squares) and CIs for $d_{unbiased}$ (square size is proportional to sample weight). The large diamond at the very bottom is centered (vertical peak of diamond) at the mean effect size, while horizontal diamond peaks indicate CI for the mean.

The findings underscore the conclusions that (1) when evaluating the probable validity of scientific conclusions, it is necessary to take into account all of the available data that bear on the conclusion; and (2) obtaining a collection of independent experimental results that all test a given hypothesis constitutes much stronger evidence regarding the hypothesis than any single result. These conclusions are usually downplayed or overlooked in discussions of the reproducibility crisis and their omission distorts the picture.

To appreciate the problem, we can re-examine the argument that the PPV of much neuroscience is also low (Button et al., 2013). PPV is quantified as:

$$\mathrm{PPV} = [(1-\beta)*\mathrm{R}]\,/\,[((1-\beta)*\mathrm{R})+\alpha)],$$

where R represents the "pre-study odds" that a hypothesis is correct, α is the *p* value, and 1-β is the power of the statistical test used to evaluate it. R is approximated as the anticipated number of true (T) hypotheses divided by the total number of alternative hypotheses in play, true plus false; i.e., R = T/(T + F). This argument depends heavily on the concept of "pre-study odds." In the example of a "gene-screen" experiment (Ioannidis, 2005) that evaluates 1000 genes, i.e., 1000 distinct "hypotheses" where only one gene is expected to be the correct one (note that these are not true hypotheses, but it is simplest to retain the statisticians' nomenclature here). R is ~1/1000, and with a *p* value for each candidate gene of 0.05, PPV would be quite low, ~0.01, even if the tests have good statistical power (≥0.8). That is, the result would have ~1/100 chance of being replicated, apparently supporting the conclusion that most science is false.

Fortunately, these concerns do not translate directly to laboratory neuroscience work in which researchers are testing actual explanatory hypotheses. Instead of confronting hundreds of alternatives, in these cases, previous work has reduced the number to a few genuine hypotheses. The maximum number of realistic alternative explanations that I found in reviewing *The Journal of Neuroscience* articles was four and that was rare. Nonetheless, in such cases, R and PPV would be relatively high. For example, with four alternative hypotheses, R would be 1/4; i.e., ~250 times greater than in the gene-screen case. Even with low statistical power of ~0.2 and *p* value of 0.05, PPV would be ~0.5, meaning that, by the PPV argument, replication of experimental science that tests four alternative hypotheses should be ~50 times more likely than that of the open-ended gene-screen example.

Furthermore, PPV is also inversely related to the *p* value, α; the smaller the α, the larger the PPV. A realistic

Table 1: Analysis of *The Journal of Neuroscience* Research Articles

Start	End	Hyp-E	Hyp-I	Alt Hyp	# Tests	Support	Reject	Disc	Ques	Comp
32	50	X			6	X				
51	59		X	2	5	X	X			
60	72		X		7	X	X			
74	92		X		7	X				
93	107		X		6	X				
108	119								X	
120	136		X		7	X				
137	148		X	3	8	X	X			
149	157	X		2	5	X	X			
158	172		X	2	8	X	X			
173	182	X		2	5		X			
183	199	X			9	X				
200	219							X		
220	231								X	
232	244	X			7	X				
245	256	X		1	7	X	X			
263	277	X		1	7	X	X			
278	290								X	
291	307		X		7	X				
308	322		X	1	7	X	X			
322	334		X		7	X				
335	346								X	
347	362							X		
363	378		X		6	X				
379	397	X		1	8	X	X			
398	408									X
409	422		X	1	4	X				
423	440	X		3	8	X	X			
441	451	X		2	9	X	X			
452	464		X		6	X				
465	473	X			8	X				
474	483	X			9	X				
484	497	X		1	9	X	X			
498	502								X	
518	529		X		8	X				
530	543									X
548	554		X	1	8	X				
555	574		X		9	X				
575	585	X			8		X			
586	594							X		
595	612	X		1	3	X	X			
613	630		X	1	7	X	X			
631	647		X		8	X				
648	658	X			6	X				
659	678	X		3	5	X				
679	690	X.			5		X			
691	709									X
710	722									X
723	732		X	1	5	X	X			
733	744									X
745	754	X		1	4		X			
755	768		X		5	X				

Classification of research reports published in *The Journal of Neuroscience*, vol. 38, issues 1–3, 2018, identified by page range ($n = 52$). An x denotes that the paper was classified in this category. Categories were: Hyp-E: at least one hypothesis was fairly explicitly stated; Hyp-I: at least one hypothesis could be inferred from the logical organization of the paper and its conclusions, but was not explicitly stated; Alt-Hyp: at least one alternative hypothesis in addition to the main one was tested; # Tests: is an estimate of the number of experiments that critically tested the major (not subsidiary or other) hypothesis; Support: the tests were consistent with the main hypothesis; Reject: at least some tests explicitly falsified at least one hypothesis; Disc: a largely "discovery science" report, not obviously hypothesis-based; Ques: experiments attempted to answer a series of questions, not unambiguously hypothesis-based; Comp: mainly a computational modeling study, experimental data were largely material for model.

calculation of PPV should reflect the aggregate probability of getting the cluster of results. Naive joint probability considerations, Fisher's method, or a meta-analysis of effect sizes, all argue strongly that the aggregate probability of obtaining a given group of *p* values will be much smaller than any one *p* value. Taking these much smaller aggregate probabilities into account gives much higher PPVs for multiple-part hypothesis-testing experiments. For example, Cen et al. (2018), as is very common, do not specify definite alternative hypotheses; they simply posit and test their main hypothesis, so the implied alternative hypothesis is that the main one is false; hence, $R = 1/2$. Applying $p < 0.001$, as suggested by both Fisher's method and meta-analysis, to Cen et al.'s main hypothesis implies a PPV of 0.99, that is, according to the PPV argument, their primary conclusion regarding N-cadherin and PKD1 on synapse formation should have a 99% chance of being replicated.

Finally, we should note that these calculations incorporate the low statistical power reported by Button et al. (2013), i.e., 0.2, whereas actual power in many kinds of experiments may be higher. Cen et al. (2018) did not report a pre-study power analysis, yet *post hoc* power (as determined by G∗ Power software) for the six tests discussed earlier ranged from 0.69 to 0.91 (mean = 0.79), which, although much higher than the earlier estimate, is still underestimated. Power depends directly on effect size, which for the results reported by Cen et al. (2018) ranged from 1.38 to 2.02, and the version of G∗ Power that I used does not accept effect sizes >1.0. Thus, the higher levels of statistical power achievable in certain experiments will also make their predicted reliability dramatically higher than previously calculated.

Discussion

To determine the validity and importance of a multifaceted, integrated study, it is necessary to examine the study as a whole. Neuroscience has no widely accepted method for putting together results of constituent experiments and arriving at a global, rational assessment of the whole. Since neuroscience relies heavily on scientific hypothesis testing, I propose that it would benefit from a quantitative way of assessing hypothesis-testing projects. Such an approach would have a number of benefits. (1) Typical papers are jammed full of experimental data, and yet the underlying logic of the paper, including its hypotheses and reasoning about them, is frequently left unstated. The use of combining methods would require authors to outline their reasoning explicitly, which would greatly improve the intelligibility of their papers, with concomitant savings of time and energy spent in deciphering them. (2) The reliability of projects whose conclusions are derived from several tests of a hypothesis cannot be meaningfully determined by checking the reliability of one test. The information provided by combining tests would distinguish results expected to be more robust from those likely to be less robust. (3) Criticisms raised by statisticians regarding the reproducibility of neuroscience often presuppose that major scientific conclusions are based on single tests. The use of combining tests will delineate the limits of this criticism.

Fisher's method and similar meta-analytic devices are well-established procedures for combining the results of multiple studies of the "same" basic phenomenon or variable; however, what constitutes the "same" is not rigidly defined. "Meta-analysis is the quantitative integration of results from more than one study on the same or similar questions" (Cummings and Calin-Jageman, 2017; p. 222). For instance, it is accepted practice to include studies comprising entirely different populations of subjects and even experimental conditions in a meta-analysis. If the populations being tested are similar enough, then it is considered that there is a single null hypothesis and a fixed-effects meta-analysis is conducted; otherwise, there is no unitary null-hypothesis, and a random-effects meta-analysis is appropriate (Fig. 3; Borenstein et al., 2007; Cummings and Calin-Jageman, 2017). Combining techniques like those reviewed here have not, as far as I know, expressly been used to evaluate single hypotheses, perhaps because the need to do so has not previously been recognized.

Meta-analytic studies can reveal the differences among studies as well as quantify their similarities. Indeed, one off-shoot of meta-analysis is "moderator analysis" to track down sources of variability ("moderators") among the groups included in an analysis (Cummings and Calin-Jageman, 2017, p. 230). Proposing and testing moderators is essentially the same as putting forward and testing hypotheses to account for differences. In this sense, among others, the estimation approaches and hypothesis-testing approaches clearly complement each other.

I suggest that Fishers' method, meta-analyses of effect sizes, or related procedures that concatenate results within a multitest study would be a sensible way of assessing the significance of many investigations. In practice, investigators could report both the *p* values from constituent tests as well as an aggregated significance value. This would reveal the variability among the results and assist in the interpretation of the aggregate significance value for the study. Should the aggregate test parameters themselves have a defined significance level and, if so, what should that be? While ultimately the probability level for an aggregate test that a scientific community recognizes as "significant" will be a matter of convention, it might make sense to stipulate a relatively stringent level, say ≤0.001 or even greater, for any parameter, e.g., p_{FM}, that is chosen to represent a collection of tests.

It is important that each prediction truly follow from the hypothesis being investigated and that the experimental results are genuinely independent of each other for the simple combining tests that I have discussed. (More advanced analyses can deal with correlations among *p* values; Poole et al., 2016.) There is a major ancillary benefit to this requirement. Ensuring that tests are independent will require that investigators plan their experimental designs carefully and be explicit about their reasoning in their papers. These changes should improve both the scientific studies and the clarity of the reports; it would be good policy in any case and the process can be as transparent as journal reviewers and editors want it to be.

Besides encouraging investigators to organize and present their work in more user-friendly terms, the widespread adoption of combining methods could have additional benefits. For instance, it would steer attention away from the "significance" of *p* values. Neither Fisher's method nor meta-analyses require a threshold *p* value for inclusion of individual test outcomes. The results of every test of a hypothesis should be taken into account no matter what its *p* value. This could significantly diminish the unhealthy overemphasis on specific *p* values that has given rise to publication bias and the "file drawer problem" in which statistically insignificant results are not published.

Use of combining tests would also help filter out single, significant-but-irreproducible results that can otherwise stymie research progress. The "winner's curse" (Button et al., 2013), for example, happens when an unusual, highly significant published result cannot be duplicated by follow-up studies because, although highly significant, the result was basically a statistical aberration. Emphasizing the integrated nature of most scientific hypothesis-testing studies will decrease the impact of an exceptional result when it occurs as part of a group.

Of course, no changes in statistical procedures or recommendations for the conduct of research can guarantee that science will be problem free. Methods for combining test results are not a panacea and, in particular, will not curb malpractice or cheating. Nevertheless, by fostering thoughtful experimental design, hypothesis-based research, explicit reasoning, and reporting of experimental results, they can contribute to enhancing the reliability of neuroscience research.

Recently, a large group of eminent statisticians (Benjamin et al., 2018) has recommended that science "redefine" its "α" (i.e., significance level), to $p < 0.005$ from $p < 0.05$. These authors suggest that a steep decrease in *p* value would reduce the number of "false positives" that can contribute to irreproducible results obtained with more relaxed significance levels. However, another equally large and eminent group of statisticians (Lakens et al., 2018) disagrees with this recommendation, enumerating drawbacks to a much tighter significance level, including an increase in the "false negative rate," i.e., missing out on genuine discoveries. This second group argues that, instead of redefining the α level, scientists should "justify" whatever α level they choose and do away with the term "statistical significance" altogether.

I suggest that there is third way: science might reserve the more stringent significance level for a combined probability parameter, such as p_{FM}. This would provide many of the advantages of a low *p* value for summarizing overall strength of conclusions without the disadvantages of an extremely low *p* value for individual tests. A demanding significance level for conclusions derived from multiple tests of a single hypothesis would help screen out the "false positives" resulting from single, atypical test results. At the same time, a marginal or even presently "insignificant" result, would not be discounted if it were an integral component of a focused group of tests of a hypothesis, which would help guard against both the problem of "false negatives" and an obsession with *p* values.

References

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

Alves G, Yu YK (2014) Accuracy evaluation of the unified p-values from combining correlated p-values. PLoS One 9:e91225.

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, et al. (2018) Redefine statistical significance. Nat Hum Behav 2:6–10.

Borenstein M, Hedges L, Rothstein H (2007) Meta analysis: fixed effects vs. random effects. Available from www.meta-analysis.com.

Button KS, Ioannidis JP, 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.

Cen C, Luo LD, Li WQ, Li G, Tian NX, Zheng G, Yin DM, Zou Y, Wang Y (2018) PKD1 promotes functional synapse formation coordinated with N-cadherin in hippocampus. J Neurosci 38:183–199.

Collins FS, Tabak LA (2014) Policy: NIH plans to enhance reproducibility. Nature 505:612–613.

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

Fisher RA (1925) Statistical methods for research Workers. In: Biological monographs and manuals series. Edinburgh: Oliver and Boyd.

Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2:e124.

Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, Argamon SE, Baguley T, Becker RB, Benning SD, Bradford DE, Buchanan EM, Caldwell AR, Van Calster B, Carlsson R, Chen SC, Chung B, Colling LJ, Collins GS, Crook Z, Cross ES, et al. (2018) Justify your alpha. Nat Hum Behav 2:168–171.

Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, Finkelstein R, Fisher M, Gendelman HE, Golub RM, Goudreau JL, Gross RA, Gubitz AK, Hesterlee SE, Howells DW, Huguenard J, et al. (2012) A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490:187–191.

Poole W, Gibbs DL, Shmulevich I, Bernard B, Knijnenburg TA (2016) Combining dependent P-values with an empirical adaptation of Brown's method. Bioinformatics 32:i430–i436.

Task Force on Reproducibility, American Society for Cell Biology (2014) How can scientists enhance rigor in conducting basic research and reporting research results. Available at http://www.acsb.org/reproducibility.

Winkler AM, Webster MA, Brooks JC, Tracey I, Smith SM, Nichols TE (2016) Non-parametric combination and related permutation tests for neuroimaging. Hum Brain Mapp 37:1486–1511.