UNIVERSITY of CALIFORNIA PRESS | Collabra

Rouder, J N et al 2016 The Interplay between Subjectivity, Statistical Practice, and Psychological Science. *Collabra,* 2(1): 6, pp. 1–12, DOI: http://dx.doi.org/10.1525/collabra.28

REVIEW ARTICLE

The Interplay between Subjectivity, Statistical Practice, and Psychological Science

Jeffrey N. Rouder*, Richard D. Morey† and Eric-Jan Wagenmakers‡

Bayesian inference has been advocated as an alternative to conventional analysis in psychological science. Bayesians stress that subjectivity is needed for principled inference, and subjectivity by-and-large has not been seen as desirable. This paper provides the broader rationale and context for subjectivity, and in it we show that subjectivity is the key to principled measures of evidence for theory from data. By making our subjective elements focal, we provide an avenue for common sense and expertise to enter the analysis. We cover the role of models in linking theory to data, the notion that models are abstractions which are neither true nor false, the need for relative model comparison, the role of predictions in stating relative evidence for models, and the role of subjectivity in specifying models that yield predictions. In the end, we conclude that transparent subjectivity leads to a more honest and fruitful analyses in psychological science.

Keywords: Inference; model selection; Bayesian analysis; Bayes factors

Psychological science is facing a crisis of confidence. This crisis is fueled by the suspicion that many empirical phenomena may not replicate robustly [1–5], by the recent publication of implausible findings on extra-sensory perception [6–7], and by several instances of outright fraud (for an overview, see [8]). The crisis is not new (e.g., [9–12]), but the current incarnation has led to much reconsideration of methodological practice. Indeed, we have a new set of terms to consider, such as *p-hacking, researcher's degrees of freedom* [13], and *questionable research practices* [14].

One benefit of the crisis is a call for increased transparency throughout the research process, as reflected for instance in the preregistration of experiments and the public sharing of data [15–19]. Another is a call to re-examine how statistical evidence is reported. The target of this critique is the practice of performing significance tests and reporting associated *p*-values. This call to re-examine significance tests is widespread: it has been made in hundreds of publications (selective examples include [9, 20–22]) and a number of journal editorials [23–26]. Indeed one journal, *Basic and Applied Social Psychology* has recently banned significance tests altogether [27]. There are a number of critiques, but one common one is that in practice people can use significance test in a mindless fashion and, as a consequence, are prone to over-interpret the results (e.g., [28]). This general critique has spawned a series of concrete recommendations for change in design and analysis. Some of these recommendations are to run experiments with better power [29, 9], to report effect sizes and confidence intervals [30], and to provide informative graphical displays about the variabilities in data [12].

In contrast, Bayesian psychologists have provided far more disruptive critiques and recommendations (e.g., [22, 31–34]). The Bayesian approach requires an overhaul of the relations between models, data, and evidence. A key difference between Bayesian and conventional testing is the role of subjectivity. A fully Bayesian approach centers subjectivity as essential for principled analysis. This view of subjectivity runs counter to trends in psychological science where it is viewed as undesirable and unecessary. For example, Simmons et al. [13] write, "Bayesian statistics require making additional judgments (e.g., the prior distribution) on a case-by-case basis, providing yet more researcher degrees of freedom." (p. 1365). Our goal here is to show just the opposite. The subjectivist perspective provides a principled approach to inference that is transparent, honest, and productive.

The Bayesian-subjective approach advocated here has been 250 years in the making. It reflects the insights of statisticians, mathematicians, and physicists including Thomas Bayes, Simon Laplace, Bruno de Finetti, Harold

* University of Missouri, US
rouderj@missouri.edu
† University of Cardiff, GB
‡ University of Amsterdam, NL
Corresponding author: Jeffrey N. Rouder

Jeffreys, Leonard Savage, Dennis Lindley, Edwin Jaynes, and James Berger. We discuss the core elements and sketch the implications for psychological science. We have broken the perspective into six basic elements that follow from one to the next.

The Elements

Element #1: Models Connect Theory to Data

The goal of a scientific theory is to explain phenomena. In psychology, scientific theories are usually expressed in a verbal form, and they lead to verbal statements about the world. For example, the theory of automatic stereotype activation [35] leads to the statement that people will identify weapons more quickly and accurately following the presentation of an African American face than a Caucasian one [36–38]. Such a statement is plausible, but so is the null statement that object identification is unaffected by such face primes. It is desirable to assess the evidence for such statements with data. Yet, these statements make no reference to noise, variability, or probability. Without reference to variability or probability, the theories cannot be tested, and they remain disconnected from data. The same disconnect holds for theories that are expressed as mathematical equations. The Fechner-Weber law (see [39]), for example, is a simple mathematical law relating the visibility of a stimulus to its brightness and the brightness of the background. Missing from the Fechner-Weber law, however, are considerations of noise, that is, how real-world data may deviate from the idealized law. The disconnect between theory and data is not problematic. In fact, it is appropriate and reasonable—theories should be general and abstract, and they should not be designed to account for sources of noise in measurement in specific contexts.

Models are devices that connect theories to data. A model is an instantiation of a theory as a set of probabilistic statements. Only after theories have been instantiated into models does statistical analysis becomes possible. If the models do a good job of instantiating theory, then evidence for or against a model may interpreted as evidence for or against a theory. This notion that theories are instantiated as models and that inference on models is interpreted as inference on theories places an under-appreciated responsibility on researchers. Models must be carefully constructed to appropriately capture theoretically-meaningful constraint. If the models do a good job of instantiating theory, then interpreting inference on models as statements about theories is justified. Conversely, if the models do a poor job, then the interpretation may be unjustified. Through consideration of models, we encounter the first element of subjectivity. Subjectivity enters through model specification [40, 41], and as we will show, this subjectivity plays a central role in analysis.

Element #2: Use Relative Model Comparisons Rather Than Absolute Statements

Once models are specified, there remains the thorny problem of evaluating the evidence for them from the data. In significance testing, a single model –the null hypothesis– is specified, and its predictions are used to quantify, by means of the infamous p value, the extremeness or unusualness of the observed data under the null hypothesis. Small p values are then interpreted as evidence against the null hypothesis [42], following what is known as Fisher's disjunction: confronted with a low p value, there are two possibilities: either an exceptionally rare event has occurred, or the null hypothesis is false.

Unfortunately, Fisher's disjunction does not provide a logical basis for casting doubt on the null hypothesis [43]. Consider, for instance, the tragic case of Sally Clark whose two children died as babies. Clark was tried for murder in part as follows: The prosecutor took the death of two babies by natural causes as the null hypothesis. He then argued that two such deaths by natural causes were exceedingly rare, hence, we should reject the hypothesis. The jury did indeed convict her. Yet, several academic statisticians argued that this rejection was logically fallacious and a miscarriage of justice [44–46]. The argument was that murder-by-mother is also quite rare. So, we cannot just reject one hypothesis without considering the likelihood of its complement. Using the same logic of the prosecutor, it could be argued that mothers very rarely murder their children, and had this hypothesis served as the null, we could have concluded that the deaths must be from natural causes. Instead, we must consider the relative rarity of the events under both the murder-by-mother and the natural-causes hypotheses. Sally Clark was later freed on appeals, and passed away shortly thereafter.[1]

The mistaken idea that models may be rejected without considering alternatives stems in part from the notion that models are either true or false. To us, however, models are neither true nor false [47]. They are designed to capture theoretically important positions without necessarily capturing the totality of the situation. An analogy for thinking about models is a subway map. Subway maps are useful abstractions because they capture the ordering of stations and the intersection of different subway lines. These are the key constraints that are critical for subway riders to navigate from one part of the city to another. Subway maps, however, do not capture the distances between stops, indeed these distances are often distorted (to make the maps easier to read). Likewise, subway maps distort the color of the lines; the physical tracks are not bright red, brilliant green, or intense purple. Though maps are useful and capture important constraint, the subway map itself is neither true nor false. The key point here is that models are mathematical abstractions designed to capture limited sets of constraint rather than true-or-false real-world objects [48]. Whether the constraints captured in a model holds empirically is the critical question, but addressing this question is more nuansced than deciding if a model true or false.

The important question in statistical analysis is whether the critical relations specified by the model hold for the data in hand. Assessing whether the critical relations hold, however, requires having alternative models where the critical relations are violated. Consider a particular subway map that specifies a certain ordering of stations and certain intersection points and colors the tracks with certain colors. We can construct two alternatives: one has

a different ordering of stations and different intersection points but the same colors, and the other has the same orderings and intersections but different colors. Of course, the first alternative is the correct one for assessing the suitability of the subway map because the critical relations are about the orderings and intersection and not about the colors. Statistical analysis entails the comparison of a model against one or more carefully specified alternative models. Good inferential statements are about the improved quality or usefulness of some model over alternatives, and it is this comparison that is then interpreted as evidence for theoretical positions. By appropriately choosing alternative models, researchers can highlight those relations in the data which are theoretically important. Again, subjectivity enters the process of choosing appropriate alternative models, and it provides a way of specifying what parts of the model are theoretically important.

Element #3: Use Predictions To Quantify Evidence

A useful aspect of significance testing is the emphasis on prediction. There are many uses of the word prediction, but perhaps the most intuitive is that if a model predicts data, it provides a set of probability statements about where the data will lie, and does so before the data are observed. In significance testing, the predictions of one model –the null hypothesis– are specified, and these predictions are compared to the observed data. Although the logic of using a single model or hypothesis is faulty, the consideration of a hypothesis' predictions is logical and desirable. Such consideration may be expanded as follows: Suppose two competing models each make predictions about where the data will lie. If so, data may be collected, and the question of relative evidence is answered by seeing how much better one model predicts the observed data than another. Hence, comparing models consists of two steps: first, gathering the predictions for the data from each model; and second, assessing the relative accuracy of these predictions.

The approach may be illustrated by considering the evidence for race-based face priming on weapons identification. Suppose the question is considered by four different research teams, called Team A through Team D, who instantiate four different models to address the question. All four teams start with the same general setup in which a set of 20 participants identify weapons and tools following African American and Caucasian faces, and all four teams define the same measure of weapons bias. Let x_i denote the race-based difference in identification bias for the ith person, where positive x_i denotes a greater weapons bias following African-American face primes than following Caucasian face primes. All four teams model x_i as being normally distributed; i.e., $x_i \sim \text{Normal}(\mu,\sigma^2)$, where μ and σ^2 denote a theoretical mean and variance for the data. The teams follow the common path of considering effect size, denoted δ, where $\delta = \mu/\sigma$.

Each team considers two statements: that there is a race-based effect and there is no race-based effect. The statement that there is no effect is the easier one to instantiate. All four teams do so by setting $\delta = 0$.[2] This model is denoted Model M_0 and is shown in **Figure 1A** as the arrow. This model makes predictions for the observed effect size, the sample mean divided by the sample standard deviation. This prediction is shown as the solid line in **Figure 1B**, and all that is needed is the sample size of the intended experiment, which in this case was set to 20 participants. This prediction leads to probabilistic statements, for example the probability that the observed effect size is between −1 and 1 may be computed as the area under the curve on this interval, and in this case, it is .5 in value.

The next step is to specify an alternative model that instantiates an effect. One may be tempted to specify that $\delta \neq 0$, which is the conventional alternative hypothesis. The problem with this specification, however, is that it makes no predictions. Before seeing the data, the value of δ is unconstrained, and so is the prediction. Indeed, in the conventional approach, to make predictions the data must first be used to estimate δ. And this fact stretches the notion of prediction. The solution is to specify alternatives that also predict data. Good alternatives should be able to make probability statements about intervals where the sample effect lies. All four teams in our example follow this advice to specify models that make bona-fide predictions. They indeed instantiate the theory that there is an effect, but they do so differently. We consider the approach of Team A first:

Team A used three criteria in choosing a model for a race-based effect. First, they were agnostic as to the direction of the effect. Second, they thought effects larger than 2.0 in magnitude were highly unlikely. Third, they thought smaller effect sizes were more likely than larger effect sizes, which in fact is typical of effect sizes in the literature. After careful consideration, Team A chooses to instantiate a model with a race-based effect by placing a distribution on δ as shown in **Figure 1A**, dashed lines. This notion of placing a distribution on parameters such as δ arises from the Bayesian framework, and is helpful in generating flexible models that predict data. We refer to the model with this distribution as M_a, where the a is for Team A. Model M_a is a the *default alternative* introduced by Jeffreys [49], discussed by Rouder et al. [34], and implemented by Morey & Rouder [50]. The three criteria are met as follows: First, the agnosticism about the direction of the effect is reflected in the symmetry of the model around zero. Second, effect sizes larger than 2.0 are attenuated through the choice of a characteristic scale. Third, the criterion that smaller effect sizes are more likely than larger ones is reflected in the shape of the model. By discussing these criteria, Team A can document how and why their specification is judicious.

The predictions for this model for an experiment with a sample size of 20 people are shown as the dashed-line **Figure 1B**. As can be seen, this model predicts lower probability of very small effect sizes than the null model and higher probability of more moderate and even large effect sizes. By instantiating theory as specific models, Team A is adding expertise to the research endeavor. Of course, there are other instantiations of the theoretical statements, and we shall discuss them subsequently.

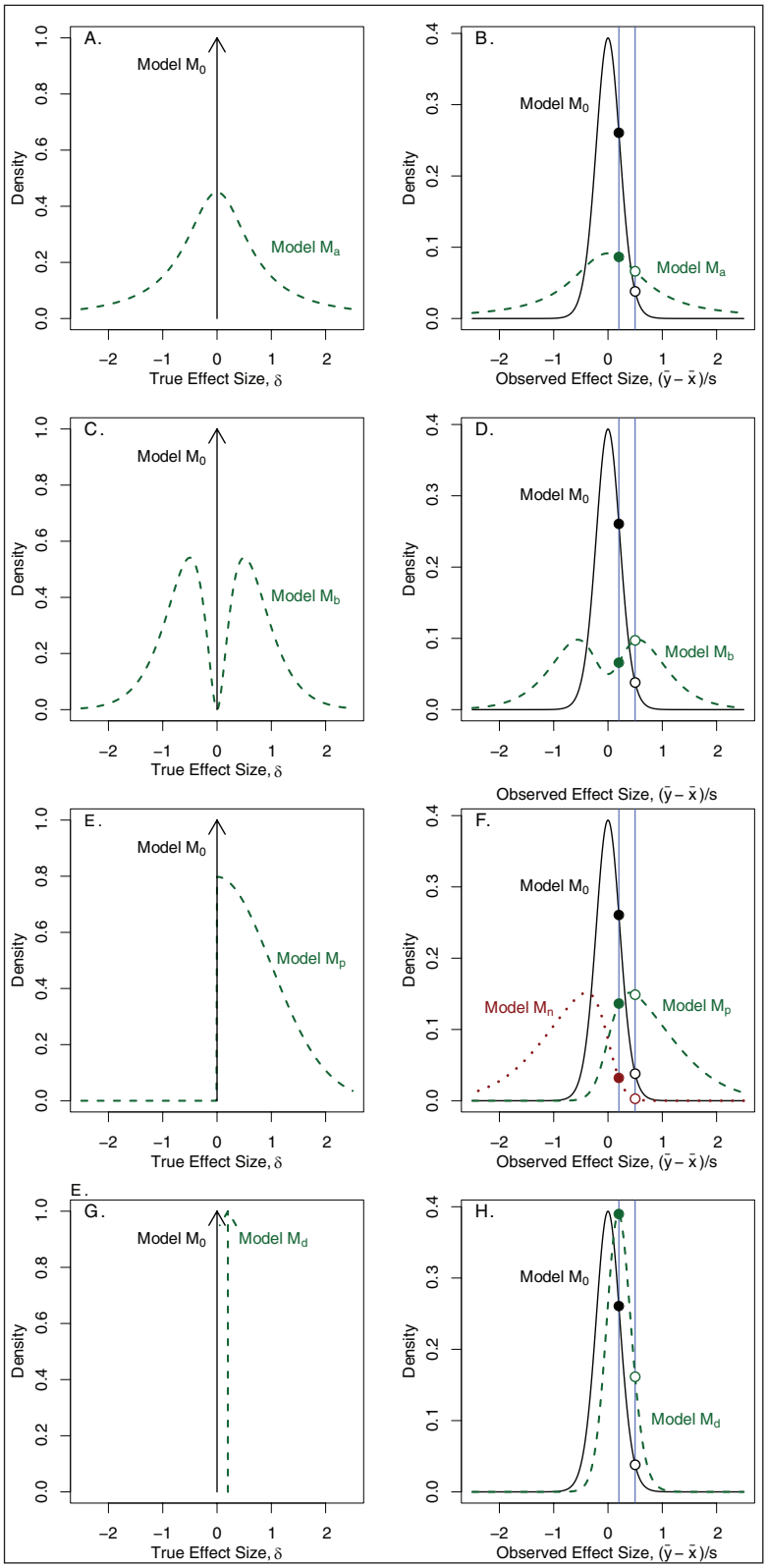

Figure 1: Models, predictions, and evidence. **A.** The models on true effect size for Team A. The arrow on $\delta = 0$ depicts an invariance of performance across the race of the face prime. The dashed line shows an alternative with mass on small and large effect sizes. **B.** The corresponding predictions of the two models for an experiment with a sample size of 20 observations. The filled and open points show the densities for observed effect-size values of .2 and .5, respectively. The ratio of the densities serve as the relative evidence for the models from the data. **C–D.** The models and predictions, respectively, for Team B. The alternative model stresses moderate effect sizes more so than small or large values. **E–F.** The models and predictions, respectively, for Team C. Team C used two alternatives, one in which participants consciously overcompensate (M_n) and second with automatic priming in the expected direction (M_p). The model M_n is not shown in Panel E to reduce clutter. **G–H.** The models and predictions, respectively for Team D. The alternative model in this case is too assumptive and arbitrary to be persuasive.

Once models make predictions, it is surprisingly easy to compare them. We illustrate the comparison for Team A, whose models and corresponding predictions are shown in **Figure 1A–B**. Suppose an experiment yielded an observed effect size of .2. The filled circles in **Figure 1B** show how well each model predicted this result. The predicted density of this result is .261 for the no-effects model (M_0), and it is .086 for the effects model (M_a). The ratio of these value is 3-to-1, and, consequently, we may state the evidence from the data as 3-to-1 in favor of M_0 relative to M_a. The panels show the case for a second, different hypothetical experimental result, one with an observed effect size of .5 (see the open circles). For this effect size, the predicted density values are .038 and .066, and the ratio is 1.7-to-1 in favor of Model M_a.

These evidence ratios are simply the probability of the data under one model relative to that under another, and are called the Bayes factor [49, 51, 52]. Following W. Edwards et al. [20], many authors have advocated Bayes factors for inference in psychological research (e.g., [22, 32, 34, 53]). Bayes factors can be interpreted as providing an updating factor for relative beliefs (e.g., [54, 55]). Here we see that the Bayes factor may be interpreted without reference to specific beliefs about the plausibility of one model or the other, as the evidence from data for two competing models.

Perhaps the least familiar element in this approach is the notion that models predict data [56]. Conventional models treat parameters as fixed quantities that are unknown before data collection. Consequently, there are no predictions before data are collected, and model comparison is complicated. With predictions, in contrast, model comparison is simple. Models yield predictions when appropriately constrained. This constraint, like all constraint in model specification, is subjective to some degree. Bayesians simply follow the subjective viewpoint to a logical endpoint—the point at which predictions are possible. Without this subjectivity, evidence remains an informal and difficult concept.

Element # 4: State Evidence

Much of today's statistical landscape derives from the ideas of Neyman and Pearson, who framed statistical analysis as an exercise in decision making [57, 58]. In their view, analysts make decisions according to certain rules, and readers, reviewers, and editors are passive bystanders who check that the pro-forma rules for decision-making were followed. However, the whole of the decision-making enterprise may not be necessary or helpful. As an alternative, researchers should be able to state finely graded measures of evidence from data for or against models. This view that evidence should be measured and reported without recourse to decisions has a long history in analysis. Perhaps the most famous proponent was Fisher [59] who considered the size of the *p*-value a measure of evidence against null. Hacking [60], A. Edwards [61], and Royall [62] have argued that likelihood ratios be used to measure the relative support for one model vs. another. In Bayesian statistics, the evidentiary view comes from Laplace (see [63, 64]) who argued that researchers should hold beliefs about the plausibility of models in the form of odds ratios, for example, a 10-to-1 odds ratio meant that the researcher believed that one model was 10 times as plausible as the other. Rozenboom [11] clearly lays out the case for the evidential viewpoint in psychology, writing:

> "The null-hypothesis significance test treats 'acceptance' or 'rejection' of a hypothesis as though these were decisions one makes. But a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a degree of believing or disbelieving which, if rational, is not a matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true." (p. 422–423)

If evidence were stressed over decisions, researchers would be free of pressure to reach particular decisions, and instead could concentrate on reporting the continuous evidence that the data provide for the hypotheses under consideration. Indeed, in this manner, all resulting evidence values from well run experiments could potentially be of interest. Those researchers who are interested in decision making can combine the measure of evidence in the data with utilities that are specified for expected action-outcome pairs, choosing the action that maximizes the expected utility [65].

Consider Team A's analysis of the two different hypothetical experimental outcomes. For each case, the team may simply state their assessment of the evidence, which is 3-to-1 for the null and 1.7-to-1 for an effect, respectively. Most researchers would not be impressed with the 1.7-to-1 support for an effect.

Team A's equivocal result, 1.7-to-1, may seem problematic if only because the researchers went through all the effort of running an experiment and did not obtain a firm answer. In conventional frequentist frameworks, if Team A wishes a more decisive answer, they would need to start from scratch and run a new experiment. In fact, collecting data and then deciding whether to continue or stop—known as *optional stopping*—is known to artificially increase Type I error rates and is a prime example *p*-hacking [66]. This unfortunate property, however, does not hold for the Bayesian evidence approach advocated here. In fact, it is a straightforward consequence of Bayes' rule that the Bayes factor interpretation of evidence does not depend on researchers intent [20, 67]. Rouder [68] explored this property in simulation and examined how Bayes factors changed with optional stopping. He showed that the nominal value of the Bayes factor tracked the true value across several different optional stopping scenarios. The interpretation of evidence as the ratio of predictive accuracy holds regardless of the sampling plan (or lack thereof). Therefore, Team A is free to run more subjects and combine them with the original sample. Matzke et al. [69] leverage this property when they recommend monitoring evidence on an observation-by-observation basis

and stopping when the evidence is deemed sufficient to draw a conclusion.

Element #5: Analysis Requires Substantive Expertise

For most researchers, statistical analysis stands apart from other aspects of science in two important ways. First, substantive researchers are inclined to view analysis as the domain of experts with special statistical skills. This treatment stands in contrast to other aspects of scholarship in psychological science including the evaluation of theory, the choice of experimental manipulations, or the interpretation of results, which are thought of activities for substantive psychologists. Second, analysis is seen as objective, an activity that must be carried out in accordance with the fixed rules set forth by the experts. In contrast, other aspects of research are seen as creative, insightful, synergistic, and scholarly. It is in this context that analysis is often done *pro forma* as a set of bureaucratic regulations that must be followed en route to publications [28, 70]. This view that analysis is a set of objective procedures developed by specialists without regard to substantive issues is counterproductive and unprincipled.

To perform productive and principled analyses, researchers should approach statistics much as they do the other research activities in psychological science. Analysis is a subjective activity that must be thoughtful and transparent. Researchers add value through the specification and interpretation of models and model comparisons. It takes substantive expertise to choose models and their alternatives and to interpret evidence for these models as evidence for theoretical statements. Although substantive researchers may consult with statisticians on how to analyze various models, assuredly the bulk of the responsibility for statistical thinking must reside with the substantive researchers themselves, much as it does for all other aspects of research. Moreover, as there are many different ways of specifying models, the research community needs to embrace intellectual diversity in analysis.

A concrete example is helpful to show how intellectual diversity might be manifest in practice. Suppose that three additional teams, Team B, Team C, and Team D are analyzing the same race-based priming data. They seek to add expertise by carefully specifying models, but they do so differently than Team A as follows:

Team B: After careful consideration, Team B comes up with one additional criteria, a stress on moderate effect sizes. Team B first studied comparable priming effects where they are known to occur, such as in other object-priming effects (e.g., [71; 72]), and race effects in similar implicit tasks (e.g., [73]). This study revealed that effect sizes tended to cluster between .2 and .8. They specify the distribution shown in **Figure 1C** on δ as an alternative to the no-effect model. This model, referred to as M_b, makes the same *a priori* commitment that it is equally likely for positive and negative race-based effects, but it makes an additional commitment that true effect-sizes are more likely to be moderate rather than big or small. The predictions for this model are shown as the dashed-line **Figure 1D**; more moderate values of effect sizes are emphasized.

Team C: Team C has a dilemma; the members do not agree on the appropriate model. One member of Team C worries that people will consciously overcompensate for any automatic effects. The result, according to this member, may be a reverse bias where tools are identified more accurately following African American faces. After much back-and-forth, the team decides on using three rather than two models: Model M_0, the equal-performance model, as well as Models M_n and M_p, models that instantiate the overcompensated and undercompensated priming effects, respectively. Model M_p is shown in **Figure 1E**, and here the effect size is constrained to be positive. Model M_n mirrors that on M_p for negative effect sizes, but is omitted in the plot for clarity. The predictions for all three models are shown in **Figure 1F**.

Team D: Team D decides to use Cohen's notion of a small effect size to set the alternative. They assume that if there is an effect, it would be positive and small, and consequently set $\delta = .2$. The specifications and predictions of this model are shown as dashed lines in **Figure 1G–H**, respectively.

All four teams are now in the position to state evidence for their models. They may compare predicted densities at observed effect sizes and compute evidence ratios as Team A did. **Table 1** provides the results for the two hypothetical experimental outcomes.

Figure 1 and **Table 1** show a diversity of models and results. Different teams reach different numerical values of the evidence they provide for theoretical statements and these vary appreciably. Some readers may understandably feel such variation is undesirable. To see why this variation is reasonable, and indeed desirable, consider the following interrelated points: 1. It is necessary to instantiate theory as a set of competing models, and this instantiation is a creative, innovative, value-added activity. A diversity of models even for the same theory should be embraced as part of an intellectual richness rather than be the subject of some arbitrary homogenization under the euphemism of convention. If we accept such diversity in modeling, then model-comparison results will also necessarily vary, much as they do in **Table 1**. There is no reason to fear such diversity so long as researchers are explicit and transparent about their modeling choices. 2. Evidence, while interpreted as assessment of theory, must be understood in the context of specific models. Evidence is not an objective property of data; rather it is also a function of the models, which in turn depend on the questions being asked and the context in which they are considered [53]. Evidence is neither unitary nor objective, and expectations to this effect are fruitless. 3. Variation in modeling

	Evidence for M_0 to alternative model			
	Team A	**Team B**	**Team C**	**Team D**
$\hat{\delta} = .2$	3.0-to-1	4.0-to-1	1.9-to-1*	1-to-1.5
$\hat{\delta} = .5$	1-to-1.7	1-to-2.6	1-to-3.9*	1-to-4.2

Table 1: Bayes factors (evidence ratios) from the four teams for two hypothetical data sets.

* Evidence is M_0 to M_p, the best alternative.

is no different in kind than variation in all other aspects of research. For example, it is expected that no two labs would operationalize a theory exactly the same way or select the same exact design or even test the same number of participants. This variation is not considered problematic in general, and we as a scientific community continually evaluate the appropriateness of one another's choices. This type of evaluative process may be practiced just as effectively for modeling choices. Analysis should be practiced and evaluated like all other aspects of research, as a subjective expertise-added endeavor upon which reasonable people may show variation.

Element #6: Readers Evaluate Results

One aspect of modern practice of statistics in psychology is that it entails almost no formally-defined responsibility for readers. Because models are not often discussed, and because readers are forced by conventional rules to honor researchers' decisions assuming that they are made in accordance with the rules, readers have little to do other than note rejection and fail-to-reject decisions. Readers of course exercise judgment, but they do so outside the rules of statistical analysis.

In our framework, readers have a formal role of active evaluation of the statistical analyses. Researchers add expertise by instantiating theories as models, comparing target models against carefully chosen alternatives, and by reinterpreting the evidence for one model relative to another as statements about the plausibility of theories. Readers should actively evaluate all of these steps, say, whether the models well instantiate the theory and whether the alternative models offer a suitable contrast. Moreover readers should form their own opinion about the impact and interpretation of the evidence.

As readers, we would find the choices of Team A, B, and C broadly defensible and appropriate, and, consequently, find the reported evidence values interpretable and relevant to the evaluation of the theory. By the same token, we find the alternative used by Team D to be inappropriate and unjustifiable. The problem is that there is no theoretical defense for specifying an effect to be at a single point with a seemingly arbitrary value. The model provides overly rigid constraint on the data, and in this regard, it is not judicious. As a consequence, we would find the evidence to be less interpretable for the evaluation of theory. As readers, we would also form our own opinion of the Bayes factors. In the case of the results above which were obtained in a within-subjects design, we would not be content with the relatively small values. We would hope that follow-up work focuses on providing clearer results, especially considering that data of this type are relatively easy to obtain.

Critiques

We consider two critiques of the Bayesian/subjectivist approach:

Critique 1. Researchers rarely have enough background knowledge to specify models that predict data. In practice, their models may be too arbitrary, and the resulting evidence values are too subjective for the assessment of theory.

We think such a critique is overstated for the following reasons:

I. Researchers, reviewers, and readers have much background knowledge they may bring to bear [53]. In the examples in this paper, for instance, models are specified on an effect-size parameterization. The community of researchers has much insight into the size of effects (on the effect-size metric) should they occur. For example, effect sizes frequently fall between .1 and 1, and we can call this a common range. An effect model that did not put much predictive mass on this range and instead favored a different range, say one with effect sizes between 10 and 20, would be difficult to justify. This type of background knowledge exists for several dependent measures (RT, ERP voltages, fMRI signals) and across numerous domains.

II. To help researchers with the task of specifying alternative models, we recommend a set of default models that are broadly applicable for many situations, especially ones where the researcher may have little guidance to make stronger commitments. Team A in **Figure 1A** uses the default model originally recommended by Jeffreys [49] and advocated by Rouder et al. [34]. We do not recommend a single default model, but a collection of models that may be tuned by a single parameter, the scale of the distribution on effect size. **Figure 2**, left side, shows the default model with four different settings of scale, from 10 (upper left) to .02 (lower right). Settings of scale over 1 seem unnecessarily wide as then encompass too much mass on impossibly large effect sizes, say those of 10. Likewise, settings smaller than .2 are unecessarily narrow as they do not give enough credence to effect sizes normally observed in well-executed behavioral-psychology experiments. Values of scale between .2 and 1.0 seem appropriate for many common experiments. The value used by Team A in **Figure 1A**, .707, is perfectly reasonable in most contexts. The right panel of **Figure 2** shows the effect of the choice of scale on the resulting Bayes factor. There certainly is an effect of scale. The largest effects occur for $t = 0$, and changing the scale from .2 to 1.0 increases Bayes factors from 5-to-1 for the null to 10-to-1 for the null. This factor characterizes the top of the range of variation we may expect from different reasonable specifications at moderate sample sizes. Perhaps more importantly, Bayes factors depends far less on reasonable variation in scale than it does on the observed t-value. Differences in observed t-value affects the Bayes factor by orders of magnitude, and this range dwarfs that for variation in scale (see also [52]).

In summary, variation in inference from model specification is expected and, in our opinion, relatively modest for reasonable choices. In our opinion, not only is this variation acceptable and appropriate, it is less than variation from other subjective aspects of research including the choice of procedures, materials, and subject pools.

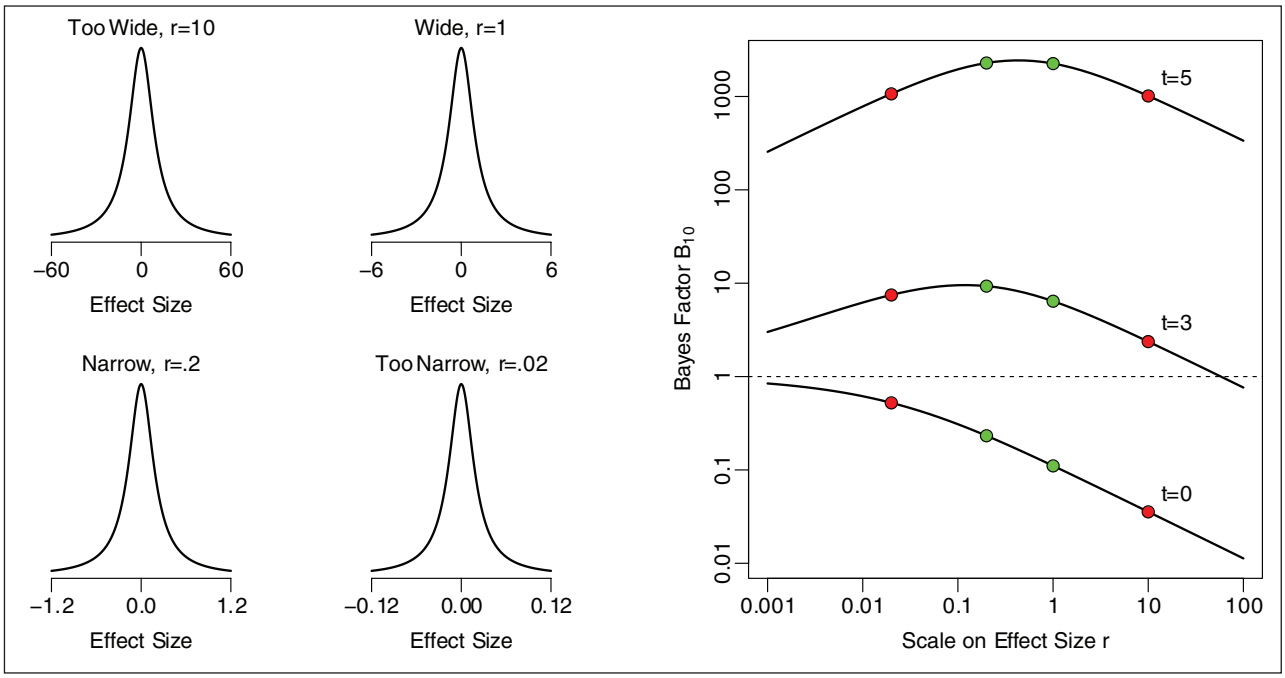

Figure 2: Alternatives may be specified on effect-size measures. **Left.** Four possible specifications. These specifications share the common property of being symmetric around zero and specifying that smaller effect sizes are more common than larger ones. The difference between them is the scale, denoted r_δ. Scales of 10 and .02, top left and bottom right panels, respectively, are too extreme to be useful. Scales of 1 and .2, top right and bottom left panels respectively, reflect limits of reasonable specifications in most contexts. **Right.** Bayes factor as a function of prior specification (r) shows that while the specification matters as it must, the changes across reasonable specifications are not great, especially when compared to the sensitivity of the Bayes factor on the data (*t*-values). The plot is drawn for a sample size of $N = 50$.

III. Researchers may explore a number of models of the same theory without penalty. For example, we could have constructed a "big effect" alternative, say with a scale of 1.0, and a "small effect" alternative, say with a scale of .2. The evidence for these models can serve as brackets on the range of evidence from the data. Models are simply models, and many may be used to understand the evidence for a phenomena (see [74], for a related perspective).

IV. Although analyses would be considerably simpler if one could somehow avoid specification of alternative models, we believe the only path to principled inference is to consider multiple models that predict data. Methods that consider only a null model, such as significance testing, overstate the evidence for effects because they cannot appropriately account for the flexibility of having an alternative that makes no predictions against a null that makes firm predictions [20, 22, 75, 76]. It is these overstatements which have contributed in part to the current methodological crisis.

V. We believe that psychological scientists can rise to the challenge of constructing models that predict data, and can cogently defend these constructions with recourse to theory, previous results, and common sense. Not only can we do it, we have peer review to check that individual authors have done a good job constructing and defending their models.

Critique 2. It seems difficult to compute the predictions of models. Computation of predictions involves integrating out all the parameters, and such a task becomes difficult in high-dimensional models. Indeed, the computation of Bayes factors is a difficult and topical area of Bayesian statistics [77]. Many researchers might not have the skills to integrate out any parameters much less those in high-dimensional space. In contrast to, say, sums-of-squares computations, computation of model predictions may be formidable.

It is with these computational difficulties in mind that we have proposed default models. As all default choices, they may not be appropriate for all possible cases, and we look forward to new choices in future development [53, 78]. But the default priors work well in the usual types of experiments, especially if the scale parameters are tuned. At the time of this writing, we have proposed default models for *t*-tests [54, 34], ANOVA [79, 80], regression [81], and correlations [82]. For these default models, convenient algorithms for the Bayes factor evidence ratios exist and are implemented in easy-to-use software. These are Rouder's website for *t*-tests and regression pcl.missouri.edu/bayesfactor; Morey and Rouder's Bayes factor package for R (BayesFactor Package, Morey & Rouder, 2014); and the new SPSS clone, JASP (JASP Team, 2016, jasp-stats.org. Development of these tools is ongoing, and they are now sufficiently rich to support inference in wide variety of common experimental settings. Hence, with a modest

amount of time and effort, it is possible to compute Bayes factors across a wide variety of settings.

Conclusion

An integrated, comparative modeling viewpoint provides a principled and transparent approach to statistical practice. This approach is intellectually far more satisfying and honest than current significance-test approaches based on the myopic consideration of one hypothesis in isolation [83]. Emphasis is placed on models as theoretically motivated statements of constraint in data. The main analytic activity is model comparison in which evidence for one model over another is stated as a relative quantity, and this evidence is defined as how well one model predicts observed data relative to another model. If a model and its alternatives are appropriate instantiations of theories, then statements of relative evidence about models may be interpreted as statements of relative evidence about theories. The critical and innovative part of the endeavor is in specifying appropriate, compelling models that are sufficiently constrained to predict structure in data. These specifications will necessarily be subjective, and the subjective element is informed by expertise and common sense rather than being formulaic and bureaucratic. Scrutinizing how well models instantiate theories is part of the reader's responsibility, and they are invited to do so through the lens of their expertise and common sense as well.

Competing Interests

The authors declare that they have no competing interests.

Author Note

Jeff Rouder, Department of Psychological Sciences, 210 McAlester Hall, University of Missouri, Columbia, MO 65211, rouderj@missouri.edu. This research was supported by National Science Foundation grants BCS-1240359 and SES-102408.

Notes

[1] The Wikipedia entry, en.wikipedia.org/wiki/Sally_Clark provides an overview Sally Clark's case.

[2] The point null of $\delta = 0$ is not the only instantiation of a no-effects model. One could take δ on a small interval around zero as the null. The general approach works similarly, and the details are provided in Morey & Rouder [54].

References

1. Carpenter, S. 2012. Psychology's bold initiative. Science 335: 1558–1561. DOI: http://dx.doi.org/10.1126/science.335.6076.1558
2. Kahneman, D. J. 2012 (September). Open letter: A proposal to deal with questions about priming effects.
3. Nosek, B. A., and Lakens, D. 2014. Registered reports: A method to increase the credibility of published results. Social Psychology 45: 137–141. DOI: http://dx.doi.org/10.1027/1864-9335/a000192
4. Pashler, H., and Wagenmakers, E.-J. 2012. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? Perspectives on Psychological Science 7: 528–530. DOI: http://dx.doi.org/10.1177/1745691612465253
5. Yong, E. 2012. Replication studies: Bad copy. Nature 485: 298–300. DOI: http://dx.doi.org/10.1038/485298a
6. Bem, D. J. 2011. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. Journal of Personality and Social Psychology 100: 407–425. DOI: http://dx.doi.org/10.1037/a0021524
7. Storm, L., Tressoldi, P. E., and Di Risio, L. 2010. Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. Psychological Bulletin 136: 471–485. DOI: http://dx.doi.org/10.1037/a0019457
8. Stroebe, W., Postmes, T., and Spears, R. 2012. Scientific misconduct and the myth of self-correction in science. Perspectives on Psychological Science 7: 670–688. DOI: http://dx.doi.org/10.1177/1745691612460687
9. Cohen, J. 1994. The earth is round ($p < .05$). American Psychologist 49: 997–1003. DOI: http://dx.doi.org/10.1037/0003-066X.49.12.997
10. Meehl, P. E. 1978. Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology 46: 806–834. Retrieved from http://www.psych.umn.edu/faculty/meehlp/113TheoreticalRisks.pdf, DOI: http://dx.doi.org/10.1037/0022-006X.46.4.806
11. Rozenboom, W. W. 1960. The fallacy of the null-hypothesis significance test. Psychological Bulletin 57: 416–428. DOI: http://dx.doi.org/10.1037/h0042040
12. Wilkinson, L., and the Task Force on Statistical Inference. 1999. Statistical methods in psychology journals: Guidelines and explanations. American Psychologist 54: 594–604. DOI: http://dx.doi.org/10.1037/0003-066X.54.8.594
13. Simmons, J. P., Nelson, L. D., and Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science 22: 1359–1366. DOI: http://dx.doi.org/10.1177/0956797611417632
14. John, L. K., Loewenstein, G., and Prelec, D. 2012. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science 23(5): 524–532. Retrieved from http://pss.sagepub.com/content/23/5/524.abstract, DOI: http://dx.doi.org/10.1177/0956797611430953
15. Chambers, C. D. 2013. Registered Reports: A new publishing initiative at Cortex. Cortex 49: 609–610. DOI: http://dx.doi.org/10.1016/j.cortex.2012.12.016

16. Nosek, B. A., and Bar–Anan, Y. 2012. Scientific utopia: I. Opening scientific communication. Psychological Inquiry 23: 217–243. DOI: http://dx.doi.org/10.1080/1047840X.2012.692215
17. Nosek, B. A., Spies, J. R., and Motyl, M. 2012. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science 7: 615–631. DOI: http://dx.doi.org/10.1177/1745691612459058
18. Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., and Kievit, R. A. 2012. An agenda for purely confirmatory research. Perspectives on Psychological Science 7: 627–633. DOI: http://dx.doi.org/10.1177/1745691612463078
19. Rouder, J. N. in press. The what, why, and how of born-open data. Behavioral Research Methods. DOI: http://dx.doi.org/10.3758/s13428-015-0630-z
20. Edwards, W., Lindman, H., and Savage, L. J. 1963. Bayesian statistical inference for psychological research. Psychological Review 70: 193–242. DOI: http://dx.doi.org/10.1037/h0044139
21. Mulaik, S., and Steiger, J. 1997. What if there were no significance tests. Mahwah, New Jersey: Erlbaum.
22. Wagenmakers, E.-J. 2007. A practical solution to the pervasive problem of p values. Psychonomic Bulletin and Review, 14: 779–804. DOI: http://dx.doi.org/10.3758/BF03194105
23. Eich, E. 2014. Business not as usual. Psychological Science 25: 3–6. DOI: http://dx.doi.org/10.1177/0956797613512465
24. Erdfelder, E. 2010. A note on statistical analysis. Experimental Psychology 57(1–4). DOI: http://dx.doi.org/10.1027/1618-3169/a000001
25. Loftus, G. R. 1993. Editoral comment. Memory & Cognition 21: 1–3. DOI: http://dx.doi.org/10.3758/BF03211158
26. Psychonomics Society. 2012. Psychonomic Society guidelines on statistical issues. Retrieved from http://www.springer.com/psychology?SGWID=0-10126-6-1390050-0
27. Trafimow, D., and Marks, M. 2015. Editorial. Basic and Applied Social Psychology 37: 1–2. DOI: http://dx.doi.org/10.1080/01973533.2015.1012991
28. Gigerenzer, G. 1998. We need statistical thinking, not statistical rituals. Behavioral and Brain Sciences 21: 199–200. DOI: http://dx.doi.org/10.1017/S0140525X98281167
29. Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. 2013. Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14: 1–12. DOIs: http://dx.doi.org/10.1038/nrn3502; http://dx.doi.org/10.1038/nrn3475
30. Cumming, G. 2014. The new statistics: Why and how. Psychological Science 25: 7–29. DOI: http://dx.doi.org/10.1177/0956797613504966
31. Lee, M., and Wagenmakers, E.-J. 2005. Bayesian statistical inference in psychology: Comment on Trafimow (2003). Psychological Review 112: 662–668. DOIs: http://dx.doi.org/10.1037/0033-295X.112.3.668; http://dx.doi.org/10.1037/0033-295X.112.3.662
32. Gallistel, C. R. 2009. The importance of proving the null. Psychological Review 116: 439–453. Retrieved from http://psycnet.apa.org/doi/10.1037/a0015251, DOI: http://dx.doi.org/10.1037/a0015251
33. Myung, I.-J., and Pitt, M. A. 1997. Applying Occam's razor in modeling cognition: A Bayesian approach. Psychonomic Bulletin and Review 4: 79–95. DOI: http://dx.doi.org/10.3758/BF03210778
34. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. 2009. Bayesian *t*-tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin and Review 16: 225–237. DOI: http://dx.doi.org/10.3758/PBR.16.2.225
35. Devine, P. 1989. Stereotypes and prejudice: Their automatic and controlled components. Journal of Personality and Social Psychology 56: 680–690. DOIs: http://dx.doi.org/10.1037/0022-3514.56.1.5; http://dx.doi.org/10.1037/0022-3514.56.5.680
36. Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., and Covert, A. E. 2004. Neural signals for the detection of unintentional race bias. Psychological Science 15: 88–93. DOI: http://dx.doi.org/10.1111/j.0963-7214.2004.01502003.x
37. Payne, B. K. 2001. Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. Journal of Personality and Social Psychology 81: 181–192. DOI: http://dx.doi.org/10.1037/0022-3514.81.2.181
38. Todd, A. R., Thiem, K. C., and Neel, R. 2016. Does seeing faces of young black boys facilitate the identification of threatening stimuli? Psychological Science 27: 384–393. DOI: http://dx.doi.org/10.1177/0956797615624492
39. Masin, S. C., Zudini, V., and Antonelli, M. 2009. Early alternative derivations of Fechner's law. Journal of the History of the Behavioral Sciences 45(1): 56–65. DOI: http://dx.doi.org/10.1002/jhbs.20349
40. Vanpaemel, W. 2010. Prior sensitivity in theory testing: An apologia for the Bayes factor. Journal of Mathematical Psychology 54: 491–498. DOI: http://dx.doi.org/10.1016/j.jmp.2010.07.003
41. Vanpaemel, W., and Lee, M. D. 2012. Using priors to formalize theory: Optimal attention and the generalized context model. Psychonomic Bulletin & Review 19: 1047–1056. DOI: http://dx.doi.org/10.3758/s13423-012-0300-4
42. Wasserman, L. 2004. All of statistics: A concise course in statistical inference. New York: Springer. DOI: http://dx.doi.org/10.1007/978-0-387-21736-9
43. Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., and Morey, R. D. in press. The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. Waldman (Eds.), Psychological science under scrutiny: Recent challenges and proposed solutions. John Wiley and Sons.

44. Dawid, A. P. 2005. Statistics on trial. Significance 2: 6–8. DOI: http://dx.doi.org/10.1111/j.1740-9713.2005.00075.x
45. Hill, R. 2005. Reflections on the cot death cases. Significance 2: 13–15. DOI: http://dx.doi.org/10.1111/j.1740-9713.2005.00077.x
46. Nobles, R., and Schiff, D. 2005. Misleading statistics within criminal trials: The Sally Clark case. Significance 2: 17–19. DOI: http://dx.doi.org/10.1111/j.1740-9713.2005.00078.x
47. Morey, R. D., Romeijn, J.-W., and Rouder, J. N. 2013. The humble Bayesian: model checking from a fully Bayesian perspective. British Journal of Mathematical and Statistical Psychology 66: 68–75. DOI: http://dx.doi.org/10.1111/j.2044-8317.2012.02067.x
48. de Finetti, B. 1974. Theory of probability (Vol. 1). New York: John Wiley and Sons.
49. Jeffreys, H. 1961. Theory of probability (3rd edition). New York: Oxford University Press.
50. Morey, R. D., and Rouder, J. N. 2015. *BayesFactor* 0.9.12–2. Comprehensive R Archive Network. Retrieved from http://cran.r-project.org/web/packages/BayesFactor/index.html
51. Kass, R. E., and Raftery, A. E. 1995. Bayes factors. Journal of the American Statistical Association 90: 773–795. DOI: http://dx.doi.org/10.1080/01621459.1995.10476572
52. Raftery, A. E. 1995. Bayesian model selection in social research. Sociological Methodology 25: 111–163. DOIs: http://dx.doi.org/10.2307/271066; http://dx.doi.org/10.2307/271063
53. Dienes, Z. 2014. Using Bayes to get the most out of non-significant results. Frontiers in Quantitative Psychology and Assessment. DOI: http://dx.doi.org/10.3389/fpsyg.2014.00781
54. Morey, R. D., and Rouder, J. N. 2011. Bayes factor approaches for testing interval null hypotheses. Psychological Methods 16: 406–419. DOI: http://dx.doi.org/10.1037/a0024377
55. Rouder, J. N., and Morey, R. D. 2011. A Bayes factor meta-analysis of Bem's ESP claim. Psychonomic Bulletin & Review 18: 682–689. DOI: http://dx.doi.org/10.3758/s13423-011-0088-7
56. Wagenmakers, E.-J., Grünwald, P., and Steyvers, M. 2006. Accumulative prediction error and the selection of time series models. Journal of Mathematical Psychology 50: 149–166. DOI: http://dx.doi.org/10.1016/j.jmp.2006.01.004
57. Neyman, J., and Pearson, E. S. 1928a. On the use and interpretation of certain test criteria for purposes of statistical inference: Part i. Biometrika 20A(1/2): 175–240. Retrieved from http://www.jstor.org/stable/2331945, DOIs: http://dx.doi.org/10.1093/biomet/20A.1-2.175; http://dx.doi.org/10.2307/2331945
58. Neyman, J., and Pearson, E. S. 1928b. On the use and interpretation of certain test criteria for purposes of statistical inference: Part ii. Biometrika 20A(3/4): 263–294. Retrieved from http://www.jstor.org/stable/2332112, DOIs: http://dx.doi.org/10.1093/biomet/20A.3-4.263; http://dx.doi.org/10.2307/2332112
59. Fisher, R. A. 1935. The logic of inductive inference. Journal of the Royal Statistical Society 98: 39–82. DOI: http://dx.doi.org/10.2307/2342435
60. Hacking, I. 1965. Logic of statistical inference. Cambridge, England: Cambridge University Press.
61. Edwards, A. 1972. Likelihood: An account of the statistical concept of likelihood and its application to scientific inference. London: Cambridge University Press. Retrieved from http://www.ams.org/mathscinet-getitem?mr=348869
62. Royall, R. 1997. Statistical evidence: A likelihood paradigm. New York: CRC Press.
63. Gillispie, C. C., Gratton-Guinness, I., and Fox, R. 1999. Pierre Simon Laplace: A life in exact science. Princeton, N. J.: Princeton University Press.
64. Laplace, P. S. 1986. Memoir on the probability of the causes of events. Statistical Science 1(3): 364–378. Retrieved from http://www.jstor.org/stable/2245476, DOI: http://dx.doi.org/10.1214/ss/1177013621
65. Lindley, D. V. 1985. Making decisions (2nd ed.). London: Wiley.
66. Yu, E. C., Sprenger, A. M., Thomas, R. P., and Dougherty, M. R. 2014. When decision heuristics and science collide. Psychonomic Bulletin & Review. DOI: http://dx.doi.org/10.3758/s13423-013-0495-z
67. Lindley, D. V. 1957. A statistical paradox. *Biometrika* 44: 187–192. DOIs: http://dx.doi.org/10.1093/biomet/44.1-2.187; http://dx.doi.org/10.2307/2333251
68. Rouder, J. N. 2014. Optional stopping: No problem for Bayesians. Psychonomic Bulletin & Review 21: 301–308. DOI: http://dx.doi.org/10.3758/s13423-014-0595-4
69. Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., and Wagenmakers, E.-J. 2015. The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. Journal of Psychology: General 144(1): e1-e15. Retrieved from http://psycnet.apa.org/doi/10.1037/xge0000038, DOI: http://dx.doi.org/10.1037/xge0000038
70. Gigerenzer, G. 2004. Mindless statistics. The Journal of Socio-Economics 33: 587–606. DOI: http://dx.doi.org/10.1016/j.socec.2004.09.033
71. Cave, C., and Squire, L. 1992. Intact and long-lasting repetition priming in amnesia. Journal of Experimental Psychology: Learning, Memory, and Cognition 18: 509–520. DOI: http://dx.doi.org/10.1037/0278-7393.18.3.509
72. Rouder, J. N., Ratcliff, R., and McKoon, G. 2000. A neural network model of priming in object recognition. Psychological Science 11, 13–19. DOI: http://dx.doi.org/10.1111/1467-9280.00208

73. Banaji, M., and Hardin, C. 1996. Automatic stereotyping. Psychological Science 7: 136–141. DOI: http://dx.doi.org/10.1111/j.1467-9280.1996.tb00346.x
74. Gelman, A., and Shalizi, C. R. 2013. Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology 66: 57–64. DOIs: http://dx.doi.org/10.1111/j.2044-8317.2011.02037.x; http://dx.doi.org/10.1111/j.2044-8317.2012.02066.x
75. Berger, J. O., and Berry, D. A. 1988. Statistical analysis and the illusion of objectivity. American Scientist 76: 159–165.
76. Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., and Wagenmakers, E.-J. (submitted). Is there a free lunch in inference?
77. Sarbanés Bové, D., and Held, L. 2011. Hyper-*g* priors for generalized linear models. Bayesian Analysis 6: 1–24. DOI: http://dx.doi.org/10.1214/ba/1339616469
78. Johnson, V. E., and Rossell, D. 2010. On the use of non-local prior desities in Bayesian hypothesis tests. Journal of the Royal Statistical Society, Series B 72: 143–170. DOI: http://dx.doi.org/10.1111/j.1467-9868.2009.00730.x
79. Rouder, J. N., Morey, R. D., Speckman, P. L., and Province, J. M. 2012. Default Bayes factors for ANOVA designs. Journal of Mathematical Psychology 56: 356–374. DOI: http://dx.doi.org/10.1016/j.jmp.2012.08.001
80. Wetzels, R., Grasman, R. P., and Wagenmakers, E.-J. 2012. A default Bayesian hypothesis test for ANOVA designs. American Statistician 66: 104–111. DOI: http://dx.doi.org/10.1080/00031305.2012.695956
81. Rouder, J. N., and Morey, R. D. 2012. Default Bayes factors for model selection in regression. Multivariate Behavioral Research 47: 877–903. DOI: http://dx.doi.org/10.1080/00273171.2012.734737
82. Wetzels, R., and Wagenmakers, E.-J. 2012. A default Bayesian hypothesis test for correlations and partial correlations. Psychonomic Bulletin & Review 19: 1057–1064. DOI: http://dx.doi.org/10.3758/s13423-012-0295-x
83. Nuzzo, R. 2015. Fooling ourselves. Nature 526: 182–185. DOI: http://dx.doi.org/10.1038/526182a

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://dx.doi.org/10.1525/collabra.28.opr

How to cite this article: Rouder, J N, Morey, R D and Wagenmakers, E-J 2016 The Interplay between Subjectivity, Statistical Practice, and Psychological Science. *Collabra,* 2(1): 6, pp. 1–12, DOI: http://dx.doi.org/10.1525/collabra.28

Submitted: 09 July 2015 **Accepted:** 28 March 2016 **Published:** 11 May 2016

Copyright: © 2016 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

UNIVERSITY of CALIFORNIA PRESS | Collabra

Collabra is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS