Strong-Form Frequentist Testing In Communication Science: Principles, Opportunities, And Challenges

Lennert Coenen \textsuperscript{a,b} and Tim Smits \textsuperscript{c}

\textsuperscript{a}Tilburg Center for Cognition and Communication, Tilburg University, Tilburg, The Netherlands; \textsuperscript{b}School for Mass Communication Research, KU Leuven, Leuven, Belgium; \textsuperscript{c}Institute for Media Studies, KU Leuven, Leuven, Belgium

ABSTRACT
This paper discusses ‘strong-form’ frequentist testing as a useful complement to null hypothesis testing in communication science. In a ‘strong-form’ set-up a researcher defines a hypothetical effect size of (minimal) theoretical interest and assesses to what extent her findings falsify or corroborate that particular hypothesis. We argue that the idea of ‘strong-form’ testing aligns closely with the ideals of the movements for scientific reform, discuss its technical application within the context of the General Linear Model, and show how the relevant P-value-like quantities can be calculated and interpreted. We also provide examples and a simulation to illustrate how a strong-form set-up requires more nuanced reflections about research findings. In addition, we discuss some pitfalls that might still hold back strong-form tests from widespread adoption.

Concerns about the validity of research findings have been at the forefront of scientific debate over the past decade. This is also true in communication science, where movements for reform have recently been gaining more and more traction (Dienlin et al., 2020). Philosophically speaking, many arguments of such movements are falsificationist in spirit (Lakatos, 1978; Mayo, 2018; Popper, 1963; see also, Derksen, 2019): they stress that scientific claims should be considered viable only to the extent that they passed heavy scrutiny and stringent testing. To this end, policy surrounding the transparency of research methods, pre-registration of research hypotheses, and replication of results are being advocated and implemented. However, there is yet another way in which communication science, much like other behavioral sciences, could gear itself toward falsificationist ideals: by rethinking statistical practice. Traditionally, communication scientists have relied disproportionally on a statistical routine known as null hypothesis significance testing, also referred to as NHST (Levine, Weber, Hullet, Park, & Lindsay, 2008; Vermeulen et al., 2015). In NHST, a substantive research hypothesis (e.g., ‘violent video game play influences aggressive cognitions’) is evaluated on the basis of a P-value associated with a test of the statistical nil-null hypothesis (‘there is no relationship between video game play and aggressive cognitions’). Whenever the P-value lies below a conventional threshold denoted by $\alpha$ – usually, $\alpha = .05$ – the finding is said to be statistically significant, which is taken to mean that there is an ‘effect’ or ‘difference’ (e.g., ‘violent video game play significantly influences aggression’). In all other cases, the result is considered not statistically significant, which is interpreted as absence of evidence for, or evidence against, an effect or difference (e.g., “violent video game play did not significantly influence aggression”).
There are two well-known issues with this approach – both of which hamper its value as a means for stringent hypothesis testing. First, NHST leaves much of the actual statistical logic to conventional decision rules and the behavior of software. Researchers choose a statistical model for the data at hand and report the results as expected by APA-standards, but they rarely confront the inferential meaning of their results. They stick to a binary conclusion of a finding being ‘significant’ or not, which ignores more fundamental reflections about the strength of a finding or its theoretical relevance. As a consequence, the evidence gleaned from NHST – in particular, the _P_-value – is known to be often misrepresented and overblown (e.g., Greenland, Senn, Rothman, Carlin, Poole, Goodman, & Altman, 2016; Nickerson, 2000; Vermeulen et al., 2015). A second, and even more fundamental, problem with NHST is that it is generally unable to say anything useful about hypotheses of substantive interest. This is true even when the _P_-values derived from the test are, technically speaking, correctly interpreted: the nil hypothesis is set up as a straw man (i.e., the claim that not the tiniest effect or relation exists between the variables of interest), and its rejection (or lack thereof) tells us little about the types of substantive hypotheses researchers are typically interested in (e.g., De Groot, 1956/2014; Meehl, 1967, 1990). Indeed, when pressed, communication researchers would probably not claim to be interested in infinitesimally small effects. Much of communication science is conducted within an aura of social relevance. Time and money is spent on investigating the impact of communication styles and technology because it is believed that there might, in fact, be substantial and replicable effects that require regulation and/or intervention. _P_-values derived from NHST seem of little help in such a scenario.

In light of these issues, many voices have been calling for statistical reforms that move away from NHST, the use of _P_-values, and even frequentist statistics in general (for an overview, see the special issue in the American Statistician: Wasserstein et al., 2019). In this paper, we will not take such a fundamental stance. Rather, we will try to familiarize communication scientists with one line of thought within the reform movement that remains close to the familiar _P_-value concept but aligns it with the falsificationist ideal of stringent hypothesis testing. After Meehl (1967), Meehl (1990), we label this approach ‘strong-form/frequentist testing’ – although its core principles reside under different names as well (e.g., ‘severe testing’: Mayo, 2018; Mayo & Spanos, 2006; ‘minimal effect testing’: Murphy & Myors, 1999; ‘(non-)equiv-alency testing’: Weber & Popova, 2012; Wellek, 2010). The basic idea of strong-form testing is straightforward: instead of testing a nil-null-hypothesis we simply set up a test for a statistical hypothesis of (minimal) theoretical interest. This provides us with a different kind of _P_-value – which, in this paper, we choose to denote by _P_ – that may be more directly interpreted as a measure of falsifying or corroborating evidence (see also, Greenland, 2019; Haig, 2020): it provides information on how much evidence the data offer to refute (or corroborate) the hypothesized effect size of (minimal) interest.

In what follows, we first provide a refresher of key concepts within the frequentist hypothesis testing framework – in particular, sampling distributions, test statistics, _P_-values, and statistical power. Readers who are knowledgeable of frequentist inference will find the information in this section familiar and may decide to skip it entirely; the value in this part mainly lies in (1) gradually introducing the logic and terminology of frequentist statistics (which seems useful for non-expert audiences), (2) clarifying some terminological choices made in the paper, and (3) showing that the ‘strong-form’ construal flows naturally from a properly conceptualized frequentist logic. In the second part of the paper we formulate general guidelines for the application of strong-form testing within the context of the General Linear Model: we show how _P_ can be calculated, discuss some important considerations when interpreting _P_, and we illustrate its behavior through a simulation. In the last part of the paper, we consider some technical challenges for strong-form testing in more complicated use-cases, and we suggest pathways for future developments.
The meaning of \( P \)

While the \( P \)-value lies at the heart of statistical inference in communication science, the concept is notoriously misunderstood (Vermeulen et al., 2015). Some widespread misconceptions are, for instance, that \( P \)-values represent the probability that the null hypothesis is true, that they reflect the probability of results being replicable, or that small \( P \)-values imply practically important effects (e.g., Greenland et al., 2016; Nickerson, 2000). If we want to understand what \( P \)-values represent, we first need to be familiar with two key concepts from frequentist statistics: sampling distributions and test statistics. Sampling distributions form the building blocks of frequentism; they represent the theoretical distributions of sample statistics (e.g., means, regression coefficients, correlation coefficients, . . .) that would, hypothetically, arise from drawing infinitely many random samples of size \( n \) from a given population characterized by a population parameter \( \delta \) (e.g., a population mean or regression coefficient). To illustrate what this means, assume we are investigating video game play among teenagers and randomly sample, say, 1000 10- to 19-year-olds and find that mean daily game play is 42.545 minutes. Of course, we know that this estimate only represents the mean in our particular sample, and if we would have drawn a different sample we would have most likely gotten a different estimate – for instance, 43.123 minutes. Thus, if we would randomly draw 1000 10- to 19-year-olds from the population an infinite number of times, we would end up with a continuous distribution representing all possible values for the sample mean; this distribution is called the sampling distribution. It can be shown that the sampling distribution of a mean can be expressed mathematically by the curve – called the probability density function – of a normal distribution, centered around the true value of the population mean. This result is, known as the Central Limit Theorem, plays a crucial role in frequentist statistics. The reason for this is that it can be referenced to derive the sampling distributions of various other types of statistics; examples of this are differences between means (normal), regression coefficients (normal), variances (scaled chi-square), and all of their related test statistics. Test statistics are standardized summaries of data, constructed as a ratio of a statistic and its (estimated) variance or ‘error.’ In traditional frequentist applications, we will often be working with sampling distributions of these test statistics, and not those of the ‘raw’ statistics on which the test statistics were initially based.

Once we have a proper understanding of test statistics and sampling distributions, we naturally arrive at the set-up of a frequentist hypothesis test. That is, if we can derive that test statistics follow a mathematically defined sampling distribution parametrized by a given population parameter (denoted by \( \delta \)), then we may (1) hypothesize a value \( d \) for that population parameter (e.g., the mean standardized difference in aggression scores between the violent and nonviolent video game group is Cohen’s \( d \) of, say, 0 or 0.30), (2) derive the sampling distribution of test statistics if \( \delta = d \), and (3) evaluate where the observed test statistic \( t \) lies within that sampling distribution. The position of the observed test statistic \( t \) within the sampling distribution is typically expressed through the probability of observing an even more extreme test statistic than the one observed, assuming that the tested hypothesis were true. This quantity is known as the \( P \)-value. Symbolically, then, a typical (right-sided) \( P \)-value can be written as

\[
P = Pr(T > t; \delta = d),^2
\]

where \( Pr(T > t; \delta = d) \) is read as ‘the probability (\( Pr \))^2 of randomly drawing a test statistic \( T \) larger than observed test statistic \( t \) assuming that the population parameter \( \delta \) is equal to \( d \).’ If the observed test statistic is relatively extreme within the sampling distribution under \( \delta = d \), \( P \) will be small. This can be

---

1 For the sake of simplicity of these expressions we assume \( d > 0 \) throughout the paper, unless stated otherwise.
2 Most texts would prefer to \( Pr(T > t; \delta = d) \) instead of \( Pr(T > t; \delta = d) \). When talking about continuous distributions (as we will be doing throughout the paper) this is the same thing: the probability at any single point in a continuous distribution is \( Pr(T > t; \delta = d) = 0 \). Thus, \( Pr(T > t; \delta = d) = Pr(T > t; \delta = d) \).
interpreted as \( t \) being evidence \textit{against} the hypothesis that \( \delta = d \). If the observed test statistic is \textit{not} extreme within the sampling distribution under \( \delta, P \) will be large. In this case, we may \textit{not} consider \( t \) as evidence against the hypothesis that \( \delta = d \).

Notice how this description of \( P \)-values alludes to the intimate relationship between frequentist hypothesis testing and a (neo-)Popperian, falsificationist epistemology (Mayo, 2018; Meehl, 1967; Meehl, 1990; Popper, 1963). The key point of falsificationism is that science logically progresses through critical tests of theories; that is, what sets science apart from pseudo-science in the falsificationist sense is that the latter typically tries to gather observations that are able to \textit{confirm} theories, whereas the former actively tries to challenge theories and find \textit{counterevidence} for them. Only if observations are inconsistent with a theoretical hypothesis, a falsificationist concludes that a hypothesis – or, at least, a set of background assumptions guiding the test – has been refuted. If observations are not inconsistent with the hypothesis, the hypothesis is said to be temporarily \textit{corroborated} – that is, \textit{not refuted}, but \textit{not literally confirmed} either. Clearly, frequentist inference offers a probabilistic toolbox for this type of reasoning: if a \( P \)-value is low, there is a low probability of observing a more extreme test statistic if the tested hypothesis were true. In falsificationist terms, this means that test statistic \( t \) entails a high level of refutational information against the hypothesis \( \delta = d \) (or a low level of corroborating information). If a \( P \)-value is high, there is a high probability of observing a more extreme test statistic under the tested hypothesis \( \delta = d \). This implies low refutational information against \( \delta = d \) in \( t \) (or a high level of corroborating information), but not a literal ‘confirmation’ in any sense.

Unfortunately, the way in which frequentist hypothesis testing is typically used in communication science ignores its falsificationist roots, and it defeats much its promise for scientific reasoning: usually, communication researchers (implicitly) set their tested hypotheses to a \textit{nil-null} hypothesis in which \( \delta = 0 \). As a consequence, a \( P \)-value as it is typically reported will only encode (a lack of) refutational information against a statistical claim of ‘no difference between group means’ or ‘no linear relationship between variables \( x \) and \( y \).’ This is not completely without merit, of course: any study hypothesizing some effect or difference should aspire to refute, at the very least, the hypothesis that its results could have been caused by random noise alone (that is, without even the smallest systematic relationship or difference in the population). That said, it is also obvious that evaluating refutational information against a nil-null hypothesis is rarely the actual purpose of a study. In practice, researchers often use small \( P \)-values to infer some type of ‘meaningful’ effect (i.e., ‘significant’ in the non-statistical sense of ‘important,’ Nickerson, 2000), but such a conclusion ignores the logical asymmetry between \textit{falsifying} or \textit{corroborating} a \textit{nil-null} and \textit{corroborating} or \textit{falsifying} any particular hypothesis of interest: it is not because a finding is extreme in light of \( \delta = 0 \) that it is therefore able to corroborate a theoretically viable effect \( \delta = d \) (the ‘fallacy of rejection,’ Mayo & Spanos, 2006; Spanos, 2014). Likewise, the absence of counterevidence against \( \delta = 0 \) does not necessarily imply counterevidence against a meaningful alternative \( \delta = d \) (the ‘fallacy of acceptance,’ Mayo & Spanos, 2006; Spanos, 2014).

In short, there is very little of value in a test of the nil-null; at best, it will only ever provide very weak (counter) evidence against theoretically meaningful claims. This is also why Meehl (1967, Meehl, 1990) dubbed nil-null-testing the ‘weak’ form of frequentist testing. Meehl contrasted this with the ‘strong’ use of frequentist testing as it is typically applied in domains such as particle physics (see also, Cousins, 2020). In a ‘strong-form’ test, a researcher directly compares an observed test statistic \( t \) against a point prediction \( \delta = d \) that has been derived from theory (e.g., the standard model of physics) – literally, as per Expression (1). A \( P \)-value generated in this manner provides directly informative (i.e., ‘strong’) information with regard to the falsification or corroborating of an \textit{actual} theoretical claim, and it does not take an unnecessary detour by first setting up an easily refutable straw man nil-null hypothesis.

There appears to be no fundamental reason why communication scientists could not also turn to an application of frequentist testing in a ‘strong’ – or at least ‘stronger’ – sense. A main hurdle seems to be that it is not the standard in undergraduate textbooks or statistical software and, therefore, requires some technical involvement on part of the researcher: in order to conduct a strong-form test
a researcher will need to (1) specify the effect size that is theoretically hypothesized (instead of just assuming a standard nil-null across all studies), and (2) determine the sampling distribution of the statistics under this alternative hypothesis. Upon reflection, however, it will become clear that these steps do not actually require that much of a change to existing principles – at least not when researchers already abide by ‘best practices’ in their tests of the nil-null. As has been stressed in various papers (e.g., Dienlin et al., 2020) communication scientists relying on NHST should always calculate a quantity known as statistical power which, technically speaking, already requires them to go through most of the necessary steps for a strong-form test. Indeed, as we will see shortly, strong-form testing can be conceptualized as a relatively straightforward extension of power analyses – at least for the case of General Linear Models with fixed effects.

Before fleshing out the relationship between power analyses and strong-form testing in more detail, it seems useful to clarify four issues that are tangential to our further discussion. First, for the remainder of the paper we will simply refer to the ‘nil-null’ hypothesis as the ‘null hypothesis’. We do so despite knowing that the concept of null hypothesis was never meant to reflect only the situation where \( \delta = 0 \). In fact, as Fisher (1955) noted, the null hypothesis could refer any working hypothesis to be ‘nullified’ (see also, Gigerenzer, 2004), which, again, attests to the intimate relationship between frequentists hypothesis testing and a falsificationist epistemology. However, we believe that using the concept of ‘null hypothesis’ to reflect the situation where \( \delta = 0 \) aids the readability of the paper, and it also corresponds to its typical usage in the communication literature. When we refer to the situation where we are testing an alternative, theoretically interesting hypotheses \( \delta \neq 0 \), we will speak about the ‘substantive hypothesis’, ‘alternative hypothesis’, or ‘hypothesis of interest’.

Second, the paper will be discussing P-values as continuous measures of refutational evidence against the null hypothesis. It is useful to note that there is some controversy about the interpretation of P-values in such evidential terms: some scholars deny this ‘neo-Fisherian’ interpretation of P-values and stick only to a decision theoretical interpretation along the lines of Neyman and Pearson (e.g., Lakens, 2021). Others do, in fact, interpret P-values in similar ways to the current paper (e.g., Greenland, 2019). We believe that the preference for either interpretation is largely a philosophical matter and that both can be justified assuming that existing concerns about the preferred interpretation are adequately considered.\(^3\)

Third, the paper will only be discussing P-values as measures of refutational evidence against a hypothesis. We will not engage with the question how small a P-value should actually be to falsify a hypothesis, or how this should be weighed against the probability of a false refutation (a Type-I-error). Within the frequentist framework, this type of question is addressed by significance level \( \alpha \), not by P-values per se.\(^4\) In the debates on statistical reform many different opinions have been raised about the specification of \( \alpha \) – whether it should always be set to a very stringent level (e.g., \( \alpha = .005 \);

---

\(^1\)Important arguments against interpreting P-like quantities in terms of (counter)evidence are (1) that \( P \) does not directly say anything about the truth or falsity of a hypothesis; (2) that all P-values are equally likely under the null hypothesis; (3) that, under conditions of low power, relatively high P-values do not provide any evidence against a null hypothesis; and (4) that, under conditions of high power, relatively low P-values may actually provide more evidence in favor of the null rather than an alternative. The current paper takes note of these concerns: for (1), it will be stressed that P-values provide information about the extremity of an observation assuming the truth of a hypothesis. Thus, \( P \) is only used to quantify the observation; when extreme, the observation is considered to be a falsifier of the hypothesis (but this doesn’t mean the hypothesis has been shown to be false; see also Footnote 4). For (2), it will be noted that, under any alternative hypothesis \( \delta \neq 0 \), P-values nearing 0 are more likely (assuming power > 0.5). This is reflected in our discussion on the skewed distribution of P- and P-values; the skew, then, is taken to be suggestive of the direction of \( \delta \) compared to 0 or, more generally, \( d \) (i.e., whether the finding is a falsifier or corroborator). For (3) and (4), it will be stressed that any interpretation of a P- or P-value requires an explicit consideration of other metrics such as statistical power (i.e., the overlap between null and alternative distributions).

\(^2\)Note that we are actually cutting a terminological corner in this paper in order not to overcomplicate our discussion: it is very typical to say, as we have said here, that when \( P < \alpha \), the tested hypothesis can be rejected. However, there is a subtle problem in this wording, as it confuses a property of the observation (i.e., \( t \) is extreme enough to be called a falsifier of a hypothesis) with a property of the hypothesis (i.e., the hypothesis is falsified). Generally, in a frequentist framework – and in contrast with a Bayesian approach – we do not actually try to falsify or confirm any given hypothesis. That is, the focus lies on calculating the probability of observations assuming the truth of a hypothesis \( P(H|data) \) – not on the probability of a hypothesis conditioned on the observations, \( P(H|data) \). Strictly speaking, then, in frequentist terms \( P < \alpha \) does not allow us to say that “the tested hypothesis can
Benjamin et al., 2018), whether it should be justified on a case-by-case-basis (Lakens et al. 2018), or whether it should not be used at all (Amrhein & Greenland, 2017; McShane et al., 2019). Although we agree that pre-registering significance thresholds is important to guarantee an honest and critical interpretation after observing results (see Mayo, 2019), we remain largely agnostic about the existence of such thresholds here: whether or not a researcher specifies a significance level, interpreting the P-value in continuous terms will always provide more specific information about the refutational information entailed in an observed test statistic, thereby allowing for— even requiring— more nuanced conclusions.

Fourth, the paper will not go into detail on the deeply rooted philosophical debates between Bayesians and frequentists on what would be the ‘superior’ approach to statistical testing (for discussions, see Mayo, 2018). While the current paper is clearly embedded in frequentism, this is not to argue that it is necessarily to be preferred. We consider a choice for either side to be mostly a matter of epistemological belief (such as one’s vision on subjective and objective probability, inductive logic, and the role of falsification in science) and of dominant practice in a field. In communication science, the focus is mostly on frequentist hypothesis testing using P-values, which suggests the relevance of the approach taken in this paper. However, researchers accepting the philosophical foundations of Bayesianism may reasonably prefer other types of metrics, such as Highest Density Intervals to summarize the posterior probability distribution for an estimated parameter, or Bayes Factors to compare the relative strength of confirmatory evidence for two distinct hypotheses (e.g., a nil-null and a minimal effect size of interest).

**Statistical power, and the transition from a 'weak' to a 'strong' construal**

In any proper application of a frequentist null hypothesis test, researchers do not only specify a null hypothesis but also put forward a particular alternative [a priori](https://osf.io/sdu9m/) that defines a population effect size of theoretical interest (δ). This is done to calculate a sample size needed to attain a reasonable level of statistical power for their null hypothesis test: statistical power is the probability of correctly rejecting the null if, in fact, the alternative of interest were true (Cohen, 1988; Neyman, 1942). Assuming a one-sided test and δ > 0, this translates into the following symbolism:

\[
\text{Power} = \Pr(T > t_{\text{crit}}; \delta = d),
\]

where \( t_{\text{crit}} \) refers to the ‘critical’ value of the test statistic – that is, the value that is considered “extreme enough” (by whatever standard) to refute the null.

Notice the \( \delta = d \) in this expression: it shows that the calculation of statistical power does not assume test statistics to come from the null distribution; rather, it assumes that they are drawn from an alternative distribution defined by a population effect size \( \delta = d \). In statistical jargon, these alternative sampling distributions are called ‘non-central’ sampling distributions, whose probability density functions are parametrized by a non-centrality parameter \( \lambda \). We discuss the meaning of \( \lambda \) in more detail later in the paper and in the Technical Appendix (available from OSF, see [https://osf.io/sdu9m/](https://osf.io/sdu9m/)), but for now it will be enough to say that \( \lambda \) is some increasing function of the population effect size \( \delta \) and sample size \( n \) (i.e., \( \lambda = f(\delta)f(n) \)). When \( \delta = 0 \) (under the null hypothesis) non-central distributions reduce to so-called ‘central’ sampling distributions, with \( \lambda = 0 \). As such, central sampling distributions are the ones we typically use in null-hypothesis tests, whereas non-central distributions are the ones needed to conduct a power analysis. To visualize what all of this means, Figure 1 depicts 3...
Student $T$-distributions: one central $T$-distribution assuming that the null hypothesis is true, and 2 non-central distributions assuming that a given alternative hypothesis is true. Statistical power is represented by the area under the non-central distributions beyond the $t_{crit}$.

There are good reasons why statistical power is considered to be essential for a null hypothesis test. First, before conducting a study, we want to ensure a sample size that would give us a reasonable probability of rejecting the null if, in fact, the hypothesis of interest were true. If we know that there will be only a low probability of correctly rejecting the null hypothesis, there is little use to running the study at all! Second, high statistical power ensures that published null refutations will, on average, be more easily replicable and less inflated. This means that a literature relying on null hypothesis tests will be of more value if it is highly powered (Asendorpf et al., 2013). That being said, it should also be apparent that a consideration of statistical power still fails to address the most fundamental logical issue with tests of the null hypothesis: even though the sampling distribution under the hypothesis of interest $\delta = d$ is derived, the null hypothesis $\delta = 0$ is still the only one actually used for the statistical test. Fortunately, if we look at the statistical definition of power as in Expression (2), it is quite straightforward to arrive back at the definition of $P$-values as in Expression (1): we just need to plug in the observed test statistic $t$ for $t_{crit}$ in Expression (1) to arrive at $P$-values in a strong-form set-up.

Practically speaking, though, this ‘simple’ extension will still raise questions: how could a communication scientist ever be able to define a substantive hypothesis $\delta = d$ for which a literal falsification would be theoretically meaningful? In contrast to physics, we do not have theories that allows us to derive point predictions, so setting up a test to falsify exactly $\delta = d$—as Meehl advocated—might seem out of reach. Fortunately, this problem has a practical solution: within the context of a power analysis, we will also rarely use a hypothesis derived from theory (even though that would be optimal); often, we will define the smallest effect size of interest (SESOI; Lakens et al., 2018) for which a we want a high probability of refuting the null (denoted by $d_{min}$). Of course, testing against a minimal effect size $d_{min}$ does not exactly deliver the ‘strongest’ possible test in Meehl’s sense, which requires an
exact theoretical point prediction $\delta = d$. However, it does give way for what at least appears to be a 'stronger' test than a typical nil-null test. That is, if we are able to define a meaningful $d_{\text{min}}$ (ideally, based on a close reading of prior research findings or meta-analysis) we may also assess to what extent our findings serve as corroborating evidence that, in fact, $\delta \geq d_{\text{min}}$. To do so, we can evaluate the extremity of an observed test statistic $t$ within the sampling distribution under the population parameter of minimal interest, $\delta = d_{\text{min}}$. This requires the calculation of a one-sided $P$-value, which we will denote by $\hat{P}$ ('$P$-dot').

$$\hat{P} = \begin{cases} P(T > t; \delta = d_{\text{min}}) & \text{if } d_{\text{min}} > 0 \\ P(T < t; \delta = d_{\text{min}}) & \text{if } d_{\text{min}} < 0. \end{cases}$$

As Figure 2 visualizes, $\hat{P}$ represents the refutational information against the hypothesis of minimal interest that $\delta \geq d_{\text{min}}$ (assuming $d_{\text{min}} > 0$). When it is small, there is little refutational information against $\delta \geq d_{\text{min}}$ (i.e., it is a corroborator of $\delta \geq d_{\text{min}}$); a high value for $\hat{P}$ implies a test statistic with considerable refutational information against the hypothesis that $\delta \geq d_{\text{min}}$ (i.e., it is a falsifier of $\delta \geq d_{\text{min}}$). Synonymously, one could say that, in the former case, $t$ entails falsifying information against $\delta < d_{\text{min}}$; in the latter case, $t$ corroborates $\delta < d_{\text{min}}$.

![Sampling distribution of $T$-statistics given population effect sizes $\delta = d_{\text{min}}$. $\hat{P}$ is represented by the shaded area. The arrows show the direction and strength of corroborators and falsifiers.](image)

The way in which $\hat{P}$ is defined here would traditionally be said to correspond to a one-sided test of the form $\delta \leq d_{\text{min}}$, rather than a test of the form $\delta \geq d_{\text{min}}$. Therefore, this definition might not appear to be formally consistent with our requirement of putting the actual hypothesis $\delta d_{\text{min}}$, to the test. However, we prefer to stick with this definition for two reasons. (1) The tests of $\delta \leq d_{\text{min}}$ and $\delta \geq d_{\text{min}}$ are statistically identical for continuous distributions of $T$, in the sense that both tests evaluate $t$ at $\delta = d_{\text{min}}$. This means that $P$-values arising from both set-ups provide the exact same information (they are just each other's complement). (2) If we would have defined $\hat{P}$ as formally expected, a researcher would have needed to use inverse logic when interpreting $P$ and $\hat{P}$. That is, a low $P$-value typically means falsifying $\delta \leq 0$ (i.e., it is in line with a substantive claim; 'good'), but a low $\hat{P}$ would have meant 'falsifying $\delta \geq d$ (i.e., not in line with the substantive claim; 'bad'). In our definition, $P$ and $\hat{P}$ values maintain the same evidential rank-order. Yet another possibility would have been to also define traditional $P$-values as testing the hypothesis $\delta \geq 0$. This would also have maintained the rank-order compared to testing the claim $\delta \geq d_{\text{min}}$ but it would have inverted the interpretation of traditional $P$-values: instead of 'smaller is better' it would be 'larger is better'. This, we think, would have been unnecessarily confusing for researchers used to work with $P$-values.
Given this relatively straightforward extension from a typical $P$ (under the null) to $\hat{P}$ (under a hypothesis of minimal interest), it seems natural to ask: why has this approach not been adopted as a standard practice in the social-scientific literature? One reason could be that frequentist statistics has long been ritualized: the null routine has often been presented as “statistics per se” (Gigerenzer, 2004, p. 589), and even foundational principles such as statistical power took decades to be adequately considered (Cohen, 1992). A second reason could be that writings on strong-form testing have remained relatively scattered, operating under differing terminologies and, sometimes, lacking much technical and/or conceptual guidance. For instance, papers in the statistical reform movements have often been promoting the use of $P$-values under alternative, non-null-distributions, but they have not always attached conceptual or practical recommendations to it (Amrhein et al., 2019; Greenland, 2019; Nickerson, 2000). Others have provided more in-depth discussions on the relevant concepts and technicalities but discussed them within seemingly disparate frameworks. Some examples are minimal effects testing (Murphy & Myors, 1999), equivalency testing (Lakens et al., 2018; Weber & Popova, 2012; Wellek, 2010), magnitude-based inference (Aisbett et al., 2020), and severe testing (Mayo, 2018).

As the name suggests, minimal effects testing recommends an identical set-up to the one suggested in this paper (Murphy & Myors, 1999): define a minimal hypothesis of interest, derive the corresponding sampling distribution, and assess the extremity of an observed test statistic within this sampling distribution – that is, a quantity such as $\hat{P}$. Tests of equivalency are also technically similar, although the conceptual idea is somewhat different: within an equivalency test, researchers substantively predict the null-value, which requires them to (1) define the set of effect sizes that are ‘practically equivalent’ to the null and (2) refute the hypothesis that the population effect size is larger in absolute value than the upper and lower bounds of practical equivalence (Lakens et al., 2018; Weber & Popova, 2012; Wellek, 2010). This can be achieved on the basis of confidence intervals, or by calculating $P$-values from two one-sided tests against the largest effect size of practical equivalence (‘TOST’; Lakens et al., 2018). It should be clear that the latter is the same as setting up an estimate of $\delta = d_{\text{min}}$, deriving the non-central sampling distribution, and calculating $\hat{P}$. Similar principles apply to the framework of magnitude-based inference, although one should be cautioned about the flawed, pseudo-Bayesian interpretations that have been circulating in this corner of the literature (see, Aisbett et al., 2020, for a discussion).

The most advanced philosophical treatment of strong-form testing has been delivered by Mayo (2018). In Mayo’s terminology, ‘strong testing’ is referred to as ‘severe testing’- more generally, ‘error statistics’ – but her basic argument is similar to Meehl’s (1967). Mayo also emphasizes the falsificationist rationale of frequentist testing: as she puts it, the purpose of a statistical test should be to find out how severely the test probed a statistical hypothesis $H$ (e.g., $\delta > d$). To evaluate whether the hypothesis is severely probed we need to find out, first, how capable the test would have been to generate an observed test statistic as extreme as $t$ if $H$ had not been true. Technically, Mayo’s logic also boils down to (1) setting up the substantive hypothesis $\delta \leq d$, and (2) calculating the one-sided extremity of a test statistic $t$ within a non-central sampling distribution assuming $\delta = d$. Mayo uses the concept of severity (SEV) to refer to the $P$-value-like quantity that arises from this, and SEV corresponds to what we have defined as $\hat{P}$ above: when SEVs is low, there is a fairly low probability that $t$ would have been larger than it is, if the population effect size would be any smaller or equal to $d$. By the Frequentist Principle of Evidence (Mayo & Spanos, 2006), this implies that the data are evidence that $\delta > d$. In contrast, when SEV is high, there is a fairly high probability that $t$ would have been larger than it is if the population effect size had been at least $d$; by the Frequentist Principle of Evidence, this serves as counterevidence against $\delta > d$.

In this paper, we choose to stick to the symbol of $\hat{P}$ rather than SEV because severe testing, as discussed by Mayo (2018), entails much more than just the calculation of $P$-value-like quantities; for a claim to be severely tested researchers also need to probe the full chain of assumptions and auxiliaries underlying their tests (see also, Mayo, 2018; Scheel, Tiokhin, Isager, & Lakens, 2020; Spanos & Mayo, 2015). Here, we are only concerned with the calculation and interpretation of values such as $P$ and $\hat{P}$.
per se, so we prefer not to introduce the connotations attributed to the general concept of severe testing. Importantly, the fact that we are not concerned with probing model assumptions also entails an important caveat: throughout our discussion, we will already assume that all statistical assumptions for General Linear Models are met. If the model is incorrectly specified, $P$ and $\hat{P}$ naturally lose their interpretability. That said, in all situations where we consider a regular $P$-value to be reasonably meaningful, $\hat{P}$ should also apply.

**Strong-form frequentist testing: calculating $\hat{P}$ in the General Linear Model**

Let us now turn to considering what communication scientists should do, specifically, to apply the principles of strong-form testing (aka, minimal effects testing, severe testing, etc.). To begin with, it should be clear from our previous discussion that the technical application is relatively straightforward for General Linear Models with fixed effects: much of the set-up of a strong-form test will remain identical compared to the typical null-hypothesis case: the type (i.e., $Z$, $T$, $\chi^2$, $F$) and observed value of the test statistic will be the same, and the degrees of freedom will remain constant as well. This means that the information typically used and reported in a null-hypothesis test can be recycled into the calculation of $\hat{P}$. The only technical challenge arises in deriving the non-centrality parameter $\lambda$, which is needed to determine the sampling distribution of test statistics under the substantive hypothesis $\delta = d_{\min}$.

The concept of the non-centrality parameter is rarely discussed in much detail in the literature on applied statistics. This is very surprising, especially given its fundamental role in power analyses; in fact, even Cohen’s (1988) widely cited reference manual on statistical power does not provide a lot of detail on the meaning or calculation of $\lambda$ (see also, Liu & Raudenbush, 2004). Conceptually, $\lambda$ can be understood as a mathematical expression of “the degree to which the null hypothesis is false” (Kirk, 2013, p. 139). More technically, it is a parameter that arises in distributions of random variables that have been derived from other, normally distributed, random variables with non-zero means (Liu, 2014). It is these non-zero means of the initial variable that define the value of $\lambda$ for the newly defined random variable; $\lambda$, in turn, can be factored into the expected value (the mean) of that new variable. The exact formula of $\lambda$ will depend on the underlying transformation, but it generally holds that $\lambda$ can be expressed as an increasing function of sample size $n$ and population effect size $\delta$ – i.e., $\lambda = f(n)f(\delta)$, as noted above. We will not go into further detail on the derivation of $\lambda$ here; interested readers are referred to Liu (2014) and the Technical Appendix (available from OSF, see https://osf.io/sdu9m/), where the origins of $\lambda$ are discussed further.

Table 1 provides an overview of formulas of $\lambda$ for common applications of the General Linear Model. Alongside the non-centrality parameter, Table 1 also reports the formula for test statistics, degrees of freedom, and the functions that can be used to calculate $\hat{P}$ in R (R Core Team, 2020). Importantly, researchers do not necessarily need to calculate $\hat{P}$ manually through the formulas from Table 1; any software application conducting power analyses for the General Linear Model should be able to determine $\lambda$ (as it is used behind the scenes anyway). G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), for instance, routinely outputs whenever a researcher chooses a post-hoc power analysis procedure. To avoid confusion, this does not mean that a researcher using the post-hoc command in G*Power to calculate $\lambda$ will be conducting a post-hoc power analysis! In a post-hoc power analysis, one would use the observed effect size to calculate statistical power assuming that the observed effect size equals the population effect size. In contrast, when using G*Power to calculate $\lambda$ we only impute the smallest effect size of interest $d_{\min}$ into our calculation to obtain the value of $\lambda$ assuming that the population effect size were equal to the smallest effect size of interest. This set-up is clearly different, so there is no need to worry about the logical problems with post-hoc power analyses in the context of calculating $\hat{P}$.

In sum, the practical workflow for calculating $\hat{P}$ in General Linear Models will look something like this:
| Distribution Family | Type of hypothesis test | Test statistic | Degrees of freedom | Hypothesized (minimal) effect size of interest $d_{mn}$ | Noncentrality Parameter $\lambda$ | Function for $\hat{p}$ |
|---------------------|-------------------------|----------------|-------------------|------------------------------------------------|---------------------------------|-------------------------|
| Sample mean against constant, known variance. | $Z$ | $Z = \sqrt{n} \frac{(\bar{y} - \mu)}{\sigma}$ | / | Unstandardized: $\hat{\Delta} = (\bar{y} - \mu)$ | $\lambda = \sqrt{n} \frac{\hat{\Delta}}{\sigma}$ | $\lambda > 0: 1 - \text{pnorm}(x = z, \text{mean} = \lambda, \text{sd} = 1)$ |
| Difference of 2 independent sample means, known and equal variances. | $Z$ | $Z = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{(\bar{y}_1 - \bar{y}_2)}{\sigma}$ | / | Unstandardized: $\hat{\Delta} = (\bar{y}_1 - \bar{y}_2)$ | $\lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\hat{\Delta}}{\sigma}$ | $\lambda < 0: \text{pnorm}(x = z, \text{mean} = \lambda, \text{sd} = 1)$ |
| Sample mean against constant, unknown variance | $t$ | $t = \sqrt{n} \frac{(\bar{y} - \mu)}{s}$ | $v = n - 1$ | Unstandardized: $\hat{\Delta} = (\bar{y} - \mu)$ | $\lambda = \sqrt{n} \frac{\hat{\Delta}}{\sigma}$ | $\lambda > 0: 1 - \text{pt}(q = t, \text{df} = v, \text{ncp} = \lambda)$ |
| Difference of 2 dependent (paired) sample means. ("dependent t-test") | $t$ | $t = \sqrt{n} \frac{(\bar{y}_1 - \bar{y}_2)}{s(\bar{y}_1 - \bar{y}_2)}$ | $v = n - 1$ | Unstandardized: $\hat{\Delta} = (\bar{y}_1 - \bar{y}_2)$ | $\lambda = \sqrt{n} \frac{\hat{\Delta}}{s(\bar{y}_1 - \bar{y}_2)}$ | $\lambda < 0: \text{pt}(q = t, \text{df} = v, \text{ncp} = \lambda)$ |
| Difference of 2 independent sample means, unknown variances. Equal population variances assumed. ("t-test") | $t$ | $t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{(\bar{y}_1 - \bar{y}_2)}{s}$ | $v = n - 2$ | Unstandardized: $\hat{\Delta} = (\bar{y}_1 - \bar{y}_2)$ | $\lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\hat{\Delta}}{s}$ | $\lambda > 0: 1 - \text{pt}(q = t, \text{df} = v, \text{ncp} = \lambda)$ |

(Continued)
**Table 1. (Continued).**

| Distribution Family | Type of hypothesis test | Test statistic | Degrees of freedom | Hypothesized (minimal) effect size of interest $d_{min}$ | Noncentrality Parameter | Function for $\hat{\lambda}$ |
|---------------------|-------------------------|----------------|-------------------|---------------------------------|-------------------------|-----------------------------|
| **T**               | Difference of 2 independent sample means, unknown variances. Equal population variances not assumed. (*Welch’s T-test*) | $t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2/n_1 + s_2^2/n_2}{n_1 + n_2 - 2}}}$ | $v = n_1 + n_2 - 2$ | $\Delta = (\bar{y}_1 - \bar{y}_2)$ | $\lambda = \frac{\sqrt{n_1 n_2}}{\sqrt{s_1^2 n_1 + s_2^2 n_2}} \Delta$ | $\lambda > 0$: $1 - pt$ \hspace{1cm} \left( q = t, df = v, ncp = \lambda \right)$ |
|                     | Correlation coefficient  | $t = \frac{r_{xy} \sqrt{n - 2}}{\sqrt{1 - r_{xy}^2}}$ | $v = n - 2$ | $r_{xy}$ | $\lambda = \frac{r_{xy} \sqrt{n}}{\sqrt{1 - r_{xy}^2}}$ | $\lambda < 0$: $pt (q = t, df = v, ncp = \lambda)$ |
|                     | Regression Coefficient (1 independent variable) | $t = \frac{1}{\sqrt{n - 2}} \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$ | $v = n - 2$ | $b_1$ | $\lambda = \frac{\hat{b}_1}{\sigma^2_{xy}}$ | $\lambda = \frac{r_{xy} \sqrt{n}}{\sqrt{1 - r_{xy}^2}}$ |
|                     |                         |                |                   | $\hat{b}_1 = \frac{r_{xy} \sqrt{n}}{\sqrt{1 - r_{xy}^2}}$ |                           |                             |
|                     |                         |                |                   | $\hat{1} - \hat{2}$ | $\hat{R}_{change}^2 = \frac{\left( \hat{1} - \hat{2} \right)}{\left( 1 - \hat{R}_{total}^2 \right)}$ | $\lambda = \text{ } \hat{R}_{change}^2 \left( 1 - \hat{R}_{total}^2 \right)$ | $1 - pf (q = F, df = \nu_1, df_2 = \nu_2, ncp = \lambda)$ |
| **$\chi^2$**        | Independence of 2 categorical variables in a contingency table with k cells. All observations are independent. Decent sample size. | $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$ | $v = (r - 1)(c - 1)$ | $\hat{\nu} = \frac{m \hat{\nu}^2}{\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}}$ | $\lambda = m \hat{\nu}^2$ | $1 - p \text{chisq} (q = \chi^2, df = v, ncp = \lambda)$ |
|                     | Multiple linear regression (Full Model) | $F = \frac{SS_{b} \nu_2}{SS_{e} \nu_1}$ | $\nu_1 = p$ | $\nu_2 = n - 1 - p$ | $\hat{\nu} = \frac{\hat{R}^2}{\left( 1 - \hat{R}^2 \right)}$ | $\lambda = n \hat{\nu}^2$ |
|                     | Multiple linear regression (Specific predictors) | $\frac{SS_{b} \nu_2}{SS_{e} \nu_1}$ | $\nu_1 = p_{\text{test}}$ | $\nu_2 = n - 1 - p_{\text{total}}$ | $\hat{\nu} = \frac{\hat{R}_{change}^2}{\left( 1 - \hat{R}_{total}^2 \right)}$ | $\lambda = n \hat{\nu}^2$ |

(Continued)
Table 1. (Continued).

| Distribution Family | Type of hypothesis test | Test statistic | Degrees of freedom | Hypothesized (minimal) effect size of interest $d_{min}$ | Noncentrality Parameter | Function for $\lambda$ |
|---------------------|------------------------|----------------|-------------------|-----------------------------------------------|--------------------------|---------------------|
| Differences of more than 2 independent sample means ("One-Way ANOVA"). Equal variances assumed. | $F = \frac{SS_m v_2}{SS_e v_1}$ | $v_1 = k - 1$ $v_2 = n - k$ | $\hat{f}^2 = \frac{\eta^2}{(1 - \eta^2)}$ | $\lambda = \sum_{i=1}^{k} n_i \left( \frac{\mu_i - \mu_{grand}}{\sigma_{within}} \right)^2$ | $\lambda = \frac{nk}{(1 - \rho)^2}$, where $\rho$ is the average correlation in scores between the groups |
| Difference of more than 2 dependent sample means ("repeated measures ANOVA"). Sphericity assumed. | $F = \frac{SS_m v_2}{SS_e v_1}$ | $v_1 = k - 1$ $v_2 = k(n - 1) - n + 1$ | $\hat{f}^2 = \frac{\eta^2}{(1 - \eta^2)}$ | $\lambda = \frac{nk}{(1 - \rho)^2}$, where $\rho$ is the average correlation in scores between the groups |

Note. Except for the Welch $T$-test the formulas reported here have been restricted to cases where $\lambda$ can be straightforwardly expressed as a function of a common (standardized) effect size. See the Technical Appendix for more details. Formulas are also reported in Liu (2014), Cohen (1988), the G*Power manual, and the manual for (2020) software. Suggestive rules-of-thumb for the size of standardized effects $r$, $R^2$, $d$, $w$, and $\eta^2$ are mentioned by Cohen (1992, p. 157). $^d$ in the paired-samples-formula corresponds to $d$ as calculated by G*Power, with the standardizer in the denominator reflecting the standard deviation of the difference. It has been noted that the standard deviation of the difference is not very easy to interpret substantively, and it is sometimes recommended to use the standard deviation of the baseline to make $d$ more interpretable (Cumming, 2013). However, the formulation as given in this table most easily translates into $\lambda$, which serves the purpose of the current paper.
(1) Define the minimal effects size of interest prior to data collection. This is based on general insight in the research topic, previous research and literature review. Optionally, define (and pre-register) the cutoffs for the interpretation of statistics such as \( P \) (see step 5).

(2) Conduct the nil-null hypothesis test (as one would usually do). The analysis outputs an observed test statistic and, when applicable, degrees of freedom, correlations between (repeated) measures, etc.

(3) Calculate the non-centrality parameter for the sampling distribution of test statistics. This can be done either through the formulas in Table 1 or by choosing the ‘post-hoc’ procedure within G*Power and providing the necessary information (degrees of freedom, sample size, variances, minimal effect size of interest, correlations between repeated measures, . . .). Note that any value specified for \( \alpha \) in software such as G*Power will be irrelevant for the calculation of \( \lambda \).

(4) Use Table 1 to choose the appropriate formula for calculating the the non-central Cumulative Distribution Function of \( T \). Impute the observed test statistic \( t \), the non-centrality parameter \( \lambda \), and (when appropriate) degrees of freedom \( v \) as provided by steps 1 and 2. The outcome value, or its complement, generates \( P \) (see Table 1).

(5) Evaluate \( P \) in terms of how much corroborating or falsifying evidence it provides for or against the claim \( \delta \geq d_{min} \). When \( P \) is small, it offers (relatively speaking) more corroborating than falsifying information about \( \delta \geq d_{min} \). When \( P \) is large, it offers (relatively speaking) more falsifying than corroborating information about \( \delta \geq d_{min} \). How much falsifying or corroborating information one would need to consider it “convincing enough” will depend on the researcher’s testing philosophy and pre-registered cutoffs.

While this procedure is not overly complicated, steps 3 and 4 still require some manual operations. For this reason, a web application is being developed that allows researchers to automatically calculate \( P \) based on the relevant inputs. A link to the application will be available on the OSF page associated with this project (https://osf.io/sdu9m/).

**Illustrative examples**

We now turn to a discussion of three hypothetical examples to show how \( P \) may be calculated and interpreted. The code associated with the examples is also available from OSF (https://osf.io/sdu9m/).

Example 1. Difference between two independent group means, unknown variances (equal variances assumed).

Imagine a researcher planning to conduct a simple advertising experiment with an independent two-group design: in one group, she will expose participants to an advertisement with celebrity endorsement; in the second group she will expose them to an advertisement with an unknown endorser. After exposure, she measures attitudes toward the product using a validated measurement scale. Say that she will be using a Student \( T \)-test (assuming similar population variances in both groups) and predicts an effect size of minimal interest of at least Cohen’s \( d = .30 \) (a ‘small’ to ‘medium’ effect by Cohen’s standards). She wants to achieve reasonably high power — say, \( \text{power} = .80 \) — at a one-sided significance level of \( \alpha = .025 \) (corresponding to the critical value of a two-sided \( \alpha = .05 \)), which requires a total sample size \( n = 352 \). She conducts the study and observes the mean attitude in the celebrity endorsement group to be 2.70 (\( SD = 0.82; n = 180 \)); in the control group it is 2.33 (\( SD = 0.85; n = 172 \)). This corresponds to an observed Cohen’s \( d = 0.4430423 \), with \( t(350) = 4.153327 \), and \( P = .00002 \) (right-sided). This means (1) that the observation entails relatively strong falsifying evidence against the null hypothesis, and (2) that the observed effect size is larger than the effect size specified as being of minimal interest.

From the latter observation, it seems straightforward to conclude that, in fact, the findings corroborate \( \delta \geq d_{min} \). However, an important additional question is as follows: how strong is the corroborator? This can be evaluated on the basis of \( P \). Plugging in the formulas from Table 1, the researcher finds that \( d_{min} \) corresponds to a non-centrality parameter
Figure 3. Distributions of $\delta = 0$ and $\delta = d_{\text{exp}}$ belonging to examples 1, 2, and 3. The curves illustrate the position of the observed test statistic within the alternative and null distributions, the extent to which it provides corroborating/falsifying information, and the overlap between null and alternative distributions.
\[
\lambda = d_{min} \sqrt{\frac{n_1n_2}{n_1+n_2}} = 0.30 \sqrt{\frac{172 \times 180}{350}} = 2.813523.
\]
She then applies the corresponding function in \( R \) - technically, the complement of the cumulative distribution function for the \( T \)-distribution: 
\[
1 - pt(q = 4.153327, df = 350, ncp = 2.813523). \text{This results in } P = .09330167.
\]
This can be interpreted in three equivalent ways. First, in terms of probability, this result means that there would be a probability of at least 9.3\% to observe a larger test statistic than the one observed if the true population difference would be .30 or larger. Second, and more informally, the results suggest relatively strong counterevidence against \( \delta < d_{min} \). If \( \delta \) had been any smaller than \( d_{min} \), the probability of finding a test statistic as large as the one observed would have been even smaller than \( \hat{P} \approx .09 \). This is identical to a third way of interpreting the results: the finding \( t \) serves as a relatively strong corroboration that \( \delta \geq d_{min} \) (see also Figure 3).

**Example 2. A correlation coefficient.**

Now consider a media researcher testing a traditional 'cultivation effects' hypothesis, suggesting there being a bivariate correlation between total television exposure and the expressions of beliefs in a just world (Figure 3, second column). The researcher uses a large-scale survey in a random sample of 1900 adults, and observes \( r = -.08, t(1898) = -3.49649, P = .000241131 \). She concludes that the relationship is statistically significant at \( \alpha = .05 \) and therefore corroborates the cultivation hypothesis. However, a reviewer posits that this is not at all convincing: while the finding provides refutational information against the null hypothesis, a theoretically meaningful manifestation of a cultivation effect should have a minimal effect size of at least \( R^2 \geq .01 \), or \( |r| \geq .10 \). He notes that the observed effect size \( r = -.08 \) already falsifies the claim – but how strong of a falsifier is it? The minimal effect size \( R^2 \geq .01 \) implies \( r \leq -.10 \), which - according to Table 1 - corresponds to non-centrality parameter 
\[
\lambda = -.10 \sqrt{\frac{1900}{1-01}} = -4.380858.
\]
Calculating \( \hat{P} \) from this gives \( pt(q = -3.49649, df = 1898, ncp = -4.380858) = 0.811492 \). This means that the actual population effect size were no smaller (in absolute value) than a minimum effect size of interest \( |r| = .10 \) there would have been a probability of at least \( \hat{P} \approx .81 \) to observe a larger test statistic (in absolute value) than the one we have observed. In other words, the observation serves as a relatively substantive falsifier with regard to a theoretically meaningful manifestation of a cultivation effect – even though the relationship is statistically significant by common standards.

**Example 3. \( R^2 \)-change in multiple regression.**

Consider the scenario where a student conducts a relatively small-scale survey (\( n = 120 \)) to assess the relationship between social media use and depressive feelings, measured using a validated measurement scale. The model also contains 4 socio-demographic and psychological variables as controls. The student hypothesizes that social media use contributes substantively to the explanation of variance in scores on the depression scale, reflected in a minimal change of \( \Delta R^2 = .02 \). Using a hierarchical regression she finds that the four control variables (added as a first block) generate \( R^2 = .15 \). In a second block, she adds social media use, which increases \( R^2 \) by \( \Delta R^2 = .03, F(1, 114) = 4.170732, P = .04343538 \). This means that, by the typical standards of a null hypothesis test, we would reject the null hypothesis and find an effect size of \( R^2 = .03 \), larger than one proposed as a minimal effect size of interest. Now, do we consider this finding as a strong corroborator of the hypothesis \( R^2 \geq .02 \)? According to Table 1, the non-centrality parameter for \( R^2 \)-change is given by 
\[
\lambda = nf^2 = n \frac{R_{\text{change}}^2}{(1-R_{\text{finite}}^2)} = 120 \times \frac{0.02}{1-0.02-0.15} = 2.891566, \text{and } \hat{P} \text{ is obtained through the cumulative distribution function of the } F\text{-distribution:}
\]
\[
1 - pf(q = 4.170732, df1 = 1, df2 = 114, ncp = 2.891566), \text{which results in } \hat{P} = .3692041. \text{This does not seem fully convincing: there is still a } \hat{P} = .3692041 \text{ probability of observing a larger test}
\]
statistic than the one observed if the population effect size would be at the absolute minimum value of interest $R^2 = .02$; that is, if the population effect size were any smaller than $R^2 = .02$, the probability of observing a larger test statistic would still only be at least $\hat{P} \approx .37$. Hence, the result lie in the direction of corroboration, albeit not convincingly.

Note from this example that the formula for $\lambda$ used in multiple regression (i.e., based on $f^2$) is generalizable across tests within the General Linear Model: a relationship that can be expressed in terms of explained variance (i.e., $R$ or $\eta^2$) can be reformulated into the form of $\lambda = nf$. For ANOVA-type analyses, calculating $\lambda$ in this way will generally be easier than having to express all group-specific differences separately (see Table 1).

**The importance of interpreting overlap with the null distribution**

Looking at the specifics of our third example, and comparing the curves from the three examples in Figure 3, it should be clear that $\hat{P}$ should not just be interpreted on the basis of the curve of the non-central distribution alone. That is, when the null and alternative distributions overlap heavily, $\hat{P}$ values will often be pointing into the direction of corroboration, even when the test statistics were drawn from the null distribution. To put it in falsificationist terms, a test with large overlap between the null and alternative distributions is highly unfalsifiable: even under the absolute null scenario, it would be relatively likely to find a test statistic corroborating the hypothesis $\delta \geq d_{\text{min}}$. If the null and alternative distributions overlap completely, finding a corroboration for $\delta \geq d_{\text{min}}$ is equally probable under $\delta = 0$ and $\delta = d_{\text{min}}$, which means that finding a corroboration is completely uninformative. This is a reminder that the interpretability of test results always depends, at the very least, on the results the test is expected to generate under absolute null scenario! A general rule is this: for any given $\hat{P}$, a test with more overlap with the null distribution implies weaker corroborations of the claim $\delta \geq d_{\text{min}}$, as observations from the null distribution will be expected to generate relatively low $\hat{P}$ values with high probability (claim $\delta \geq d_{\text{min}}$ is less falsifiable; cf., Mayo, 2018). This shows why evaluating $\hat{P}$ also always requires a consideration of overlap, distance, or divergence between the null distribution and the alternative distribution. There are several pieces of information that could be used as part of this evaluation.

A first piece of relevant information can be found by simply comparing the observed $P$-value with $\hat{P}$. When the two sampling distributions overlap completely, $\hat{P}$ and $\hat{P}$ for any given test statistic $t$ will be identical. The less the two distributions overlap, the larger $\hat{P}$ should be compared to $\hat{P}$ for any given test statistic $t$. While this method is straightforward, it is also informal and subjective. More formal approaches will require the interpretation of statistics quantifying the overlap, distance, or divergence between the null and alternative distributions (see Figure 3 for a visualization). There are many options available, but we will focus our discussion on statistics with intuitive bounds between 0 and 1 (thereby excluding widely used statistics such as the Kolmogorov–Smirnov distance, the Kullback–Leibler divergence, and the Jensen–Shannon divergence).

One option is to evaluate statistical power of the null hypothesis test. Power has the advantage of being familiar to researchers, and it can be used as an indication of overlap in the sense that it is the probability of exceeding a critical (tail) value of the null distribution if the alternative is true. This may then lead one to conclude that, conceptually, higher power reflects higher distance and less overlap between the null and alternative distributions. The main problem with using power in this way is that it requires a significance level to be defined, which means that (1) it cannot be defined without requiring researchers to put a significance level in place, and (2) it is not comparable between different significance levels. These issues can be bypassed by extending the concept of power to what one could call “falsifiability.” “Falsifiability” can thus be defined as the probability of finding a test statistic that provides at least some falsifying information against $\delta \geq d_{\text{min}}$ if the null hypothesis were true. In other words, it corresponds to the probability of observing test statistics corresponding to $\delta < d_{\text{min}}$ if the null hypothesis were true. Symbolically,
\[ \varphi = \begin{cases} \Pr(T < t_{\delta=d_{\text{min}}}; \delta = 0) & \text{if } \delta > 0 \\ 1 - \Pr(T < t_{\delta=d_{\text{min}}}; \delta = 0) & \text{if } \delta < 0 \end{cases} \]

where \( \varphi \) (“phi”) stands for “falsifiability” (i.e., falsifiability of \( \delta > d_{\text{min}} \) when \( \delta = 0 \)), and \( t_{\delta=d_{\text{min}}} \) refers to the value of the test statistic corresponding to \( \delta = d_{\text{min}} \). Falsifiability is conceptually similar to power, but instead of using probabilities under the alternative based on a critical value of \( T \), it uses probabilities under the alternative based on the value of \( T \) corresponding to \( d_{\text{min}} \). Hence, the value for \( \varphi \) can easily be calculated through a minor extension of the test statistic formulas in Table 1: the expected effect size of interest \( (d_{\text{min}}) \) can be plugged into the formula for the test statistic, together with the observed sample size(s) \( n \) and (in case \( d_{\text{min}} \) is unstandardized) standard deviation(s) \( \sigma \). This will provide the test statistic corresponding to \( d_{\text{min}}, t_{\delta=d_{\text{min}}} \). From this, we can calculate the probability of observing test statistics smaller than \( t_{\delta=d_{\text{min}}} \) assuming that the null hypothesis is true. Ideally, this probability should be 1, as this means that \( \delta > d_{\text{min}} \) cannot be corroborated under the null hypothesis. Note that tests with reasonable power will already tend to spawn \( \varphi \) very close to 1; this is important to keep in mind, because it means that one may set very high standards for \( \varphi \) (with values nearing 1 being required).

When comparing null and alternative distributions one may also resort to measures that have been specifically designed to quantify distributional (dis)similarity. One notable example of this is the so-called Overlapping Coefficient (OVL), which is used as a similarity measure in various areas of statistics. OVL represents the percentual overlap of two distributions, meaning that it ranges from 0 (0% overlap) to 1 (100% overlap), with values close to 0 to be preferred. OVL can be implemented in R using the Overlapping package (Pastore, 2018). As yet another alternative, one could calculate the Total Variation Distance (TVD), which is implemented in R through the distrEx package (Rudkeshel et al., 2006). TVD also ranges from 0 to 1, and it represents the largest possible difference assigned to the same event by two probability distributions. If two distributions are identical, the largest possible difference in assigned probabilities is 0; if they are completely different (i.e., if they don’t overlap), the largest possible difference is 1. Hence, values close to 1 are also to be preferred here. Figure 4 illustrates the relationship of all statistics mentioned (power, \( \varphi \), OVL, and TVD) in case of a two-sample \( T \)-test.

Now let us briefly consider how the overlap or distance between the alternative and null distributions would affect our interpretation of our three examples. Eyeballing Figure 3, it is clear that the curves from the third example overlap heavily, whereas the curves of the other two examples do not: in the first and second example, there is relatively high power, little overlap and high falsifiability. In the third example, however, the distributions under the null and substantive hypothesis overlap heavily, which is also reflected in the values of power, \( \varphi \), OVL, and TVD. Considering \( \varphi \) specifically, the third example shows a \( \varphi = .87 \) probability of finding a test statistic in the direction of \( \delta < d_{\text{min}} \) if the null hypothesis is true. In other words, this still leaves a.13 probability of corroborating \( \delta > d_{\text{min}} \) if, in reality, \( \delta = 0 \). The values of \( \varphi \) are close to 1 for the other two examples, showing that, comparatively, they provide more informative and adequate tests. Note that the web application associated with this article will also include an option to output these measures of overlap and similarity alongside \( \hat{P} \) (see https://osf.io/sdu9m).

**What about the overlap with other, non-null-distributions?**

There is an infinite set of possible parameter values different from \( d_{\text{min}} \) all of which would spawn a different value for \( \lambda \), a different sampling distribution, and a different value for \( \hat{P} \). One could ask: should a researcher not evaluate all of these parameters? Our answer would be: no, not necessarily. We know that testing any effect below \( d_{\text{min}} \) would always suggest more corroborating information in test statistic \( t \). But given that we were calculating \( \hat{P} \) at the minimal effect size of interest, any \( \delta < d_{\text{min}} \) is not of relevance. Likewise, a given \( t \) will always generate less corroborating information for effect sizes.
larger than $d_{\min}$, but this is also not much of an issue for the initial hypothesis test: the only claim we are making is that $\delta > d_{\min}$, so finding less corroborating information for any alternative $\delta' > \delta > d_{\min}$ has no direct implications for our claim.

Importantly, this is not to say that it is irrelevant to evaluate the falsifying or corroborating information of an observed test statistic $t$ in light of parameter values other than $d_{\min}$. However, doing so seems of most use after the initial hypothesis test of interest – that is, as a follow-up, 'inductive' step. We see two ways in which a researcher could proceed at this stage. A first approach is to evaluate confidence intervals for an observed effect size. In recent years, the reporting of confidence intervals for effect sizes has been widely promoted, even to the extent that it has been dubbed “the new statistics,” ready to replace $P$-values (Cumming, 2014, p. 7). From the viewpoints expressed in this paper, confidence intervals should not be seen as a replacement of $P$-like quantities. Rather, we believe they are mutually compatible, of use at subsequent steps of the testing-and-inference-cycle: we first test a hypothesis of minimal interest, and then we may use confidence intervals to infer the range of parameter values that could not be rejected (at a fixed $\alpha$ level). That is, confidence intervals represent intervals of non-refutation (corroboration) given an observed test statistic $t$ and dichotomous decision rule (i.e., corroborate/falsify) based on some threshold $\alpha$ (“compatibility intervals”: Amrhein et al., 2019; Greenland, 2019; Hawkins & Samuels, 2021). A second approach – and one that does not require fixed $\alpha$ levels – is to eyeball so-called “severity curves” (Mayo, 2018). Severity curves represent the value of $SEV$, or $\hat{P}$, for a wide range of parameter values. Plotting severity curves can be done iteratively, through a simple application of the formulas in Table 1. The steps are as follows: (1) specify a range of values for the hypothesized effect size, (2) calculating $\hat{P}$ given all of these effect sizes, and (3) plot the values for $\hat{P}/SEV$ against their respective effect sizes. Figure 5 shows severity curves for the three examples described above. The accompanying web application will also include the option to plot severity curves automatically (see https://osf.io/sdu9m).

**Summary: the steps in a 'strong-form' test**

In sum, there will typically be four practical steps associated with a 'strong-form' test:
Figure 5. Severity curves for the three examples. Dots represent values at which $\hat{p} = .05$ and $\hat{\psi} = .95$. 
(1) Define minimal effect size of interest \( d_{\text{min}} \) and (possibly) significance thresholds.

(2) Calculate a traditional \( P \)-value. Critically evaluate the extent to which it corroborates or falsifies \( \delta \geq 0 \). Use thresholds (when those were defined) to label findings as “corroborators” or “falsifiers” of the null hypothesis.

(3) Calculate \( \hat{P} \). Critically evaluate to what extent it corroborates or falsifies \( \delta \geq d_{\text{min}} \). Use thresholds (when those were defined) to label findings as “corroborators” or “falsifiers” of the hypothesis of (minimal) interest.

(4) Evaluate the overlap, distance, or divergence between the alternative and null distributions. A corroboration should be considered as (relatively) less valuable when there is a large overlap, and (relatively) more valuable when there is little overlap.

(5) Assess the corroborating/falsifying information in \( t \) with regard to other possible values of the parameter \( \delta \). This can be achieved by interpreting a confidence interval for the effect size (i.e., an interval of compatibility/non-refutation) and/or by plotting severity curves.

**The behavior of \( \hat{P} \): a simulation**

We now turn to a simulation to illustrate our previous claims and to portray the long-term behavior of \( P \) and \( \hat{P} \) in more detail. The simulations revolve around the example of a correlation coefficient, but the definition of standardized quantities such as \( P \) and \( \hat{P} \) does not depend on the type of statistic being used. As a result, the general pattern of findings shown here will extend to other types of statistics as well – at least as long as the assumptions of the underlying statistical models are met.

**Set-up**

In our simulation, we begin by defining five population effect sizes for the correlation coefficient \( \rho = \{0, 0.05, 0.1, 0.2, 0.3\} \), five sample sizes \( n = \{30, 50, 100, 200, 300, 500\} \), five hypothesized (i.e., \( d_{\text{min}} \)) correlations that are supposedly put to the test \( \rho = 0, 0.05, 0.1, 0.2, 0.3 \), and one significance level of \( \alpha = .025 \) (one-tailed). For all combinations, we simulate 10,000 data sets and store observed test statistics, \( P \)-values, \( \hat{P} \)-values, non-centrality-parameters (\( \lambda \)), and measures of similarity or overlap. This allows us to simulate the behavior of \( P \) and \( \hat{P} \) (1) under various combinations of true effect sizes, tested effect sizes, and sample sizes, and (2) under various conditions of the overlap between the null and alternative distributions. The annotated code of the simulation is available from the paper’s OSF repository (https://osf.io/sdu9m/).

**Illustration of key points**

The non-centrality parameter \( \lambda \) is an increasing function of sample size and population effect size, so the distribution of simulated test statistics should shift further away from zero the higher (1) the population value of the correlation coefficient, and (2) the sample size. As visualized by Figure 6, this is reflected in the simulations: when the population correlation coefficient \( \rho = 0 \), the simulated distribution of test statistics follows a central distribution with \( \lambda = 0 \). When the population coefficient is non-zero, the distribution of observed test statistics shifts to a non-central \( T \) distribution with \( \lambda = \frac{r\sqrt{n}}{\sqrt{1-r^2}} \).

Of course, in reality, researchers will always be working with sampling distributions under a hypothesized effect size, and not the true effect size. This means that when we are calculating \( P \) or \( \hat{P} \) (i.e., under a hypothesized effect of 0 or \( d_{\text{min}} \)) we are assuming a distribution that may or may not differ from the true distribution. The distribution of \( P \) and \( \hat{P} \) values will therefore also depend on the difference between the hypothesized and actual sampling distributions of test statistics. To illustrate, consider the case of a regular \( P \)-value, where the hypothesized effect size is always equal to \( \delta = 0 \). It is a well-known fact that, if the null hypothesis is true (if \( \lambda = 0 \)), \( P \)-values will be uniformly distributed between 0 and 1 – regardless of sample size. This follows directly from the definition of a \( P \)-value,
Figure 6. Analytical and simulated pdf’s for $T$. This graph shows that the simulated probability density function of $T$ overlaps with the analytical solution.

Figure 7. The sampling distribution of $\theta$ under varying population effect sizes and sample sizes (upper figure), and the sampling distribution of $\theta$ under varying population effect sizes and sample sizes (lower figure).
which can be interpreted as stating that, if the null hypothesis were true, the observed test statistic $t$ would be at the $P$–th percentile of highest ranked test statistics. From this it follows that if the null hypothesis is actually true, exactly $P$ percent of all test statistics will result in a $P$-value smaller than $P$; that is, $\text{Pr}(P < P_{\text{observed}}; H_0) = P_{\text{observed}}$, which implies a uniform distribution. In contrast, if the true population effect size is any greater than 0, $P$ will have a right-skewed distribution. This also makes sense: if we are drawing test statistics from a distribution that, on average, spawns large tests statistics, then we will, on average, find small $P$-values. This phenomenon will be more and more pronounced the more the true non-centrality parameter $\lambda$ moves away from 0 (that is when the effect size and/or sample sizes increase). This behavior of $P$-values is visualized in Figure 7.

Unsurprisingly, the behavior of $P$ also translates into $\hat{P}$; $\hat{P}$ will also be uniformly distributed if the tested hypothesis $d_{\text{min}}$ is true in reality, and if the true effect size is larger (in absolute value) than the tested parameter $d_{\text{min}}$, the sampling distribution of $\hat{P}$ will be right-skewed as well. However, as Figure 7 shows, there is another interesting case to consider for $\hat{P}$: if the true effect size is smaller than the hypothesized effect $d_{\text{min}}$ the sampling distribution will be left-skewed. This means that for a population parameter greater than 0 but smaller than $d_{\text{min}}$, the distribution of $\hat{P}$ will be right-skewed, and the distribution of $\hat{P}$ will be left-skewed. The skew in the distribution of $\hat{P}$ will also more pronounced when the true non-centrality parameter diverges to a greater extent from the one assumed in the hypothesis test (e.g., when the population effect size differs greatly from the tested effect size).

The notion of skew in $\hat{P}$ distributions illustrates two fundamental points about the interpretation of strong-form tests. First, it illustrates the importance of sample size: for any fixed population effect size and tested effect size, it will be true that with increasing sample sizes, it will be easier to find falsifiers against the hypothesis $\delta \geq d_{\text{min}}$ if the population effect size were any smaller than the tested effect $d_{\text{min}}$. This means that higher sample sizes increase Popperian 'risk' (Meehl, 1968): there is a higher probability of refutation of a hypothesis $\delta \geq d_{\text{min}}$ if, in reality, $\delta$ is somewhat smaller than $d_{\text{min}}$. Likewise, higher sample sizes will also make it easier to detect strong corroborators if the actual population effect size were any greater than $d_{\text{min}}$. Both of these indicate why researchers would want to aim for large samples: (1) larger samples increase falsifiability and should therefore be lauded by reviewers; (2) when $\delta$ is any greater than $d_{\text{min}}$, higher sample sizes will, on average, spawn smaller values of $\hat{P}$ (i.e., the results will tend to be more convincing).

A second crucial point illustrated by the skew in the distribution of $\hat{P}$ is the relevance of (dis)similarity and overlap between the null and alternative distributions: when sample sizes and/or effect sizes are very small (i.e., when non-centrality parameters are very small) the null and alternative distributions overlap heavily. When this is the case $\hat{P}$ will, on average, be smaller under the null hypothesis, and, as a result, become less informative: even if the absolute null situation were true, it would be relatively likely to find strong corroborating information that $\delta > d_{\text{min}}$. This is also reflected in the distribution of $\hat{P}$-values: the more the distributions overlap, the less the distribution of $\hat{P}$ will be skewed if the null hypothesis is true – that is, the higher the probability becomes of finding small $\hat{P}$-values under the null. To illustrate, Figure 8 visualizes the distribution of $\hat{P}$-values at various levels of overlap with the null distribution, quantified by (1) power, (2) $\phi$, (3) OVL, and (4) TVD.

The added value of the strong-form construal

In our view, the construal of strong-form frequentist testing (severe testing, minimal effects testing, etc.) offers a promising addition to the statistical toolbox of communication scientists. It is a relatively straightforward extension of widespread practice, and it adds to the convergent evolution of the field toward critical theory assessment in a falsificationist sense: a strong-form test allows for a more direct evaluation of a statistical hypothesis, requires a more critical reading of findings that are “statistically significant,” and underlines the importance of making thoughtful methodological decisions before
conducting a study. That is, if we require strong-form tests (or some related conceptualization), researchers will need to pre-register a reasonable value for $d_{min}$ on the basis of their knowledge of the literature (see also, Dienlin et al., 2020). Fixing one general value of $d_{min}$ across studies will not be very useful, as the effect size of interest will be highly contingent on the type of hypothesis under investigation. For instance, when a researcher is conducting a large-scale survey to assess the overall relationship between media use and well-being it seems reasonable to set $d_{min}$ to a relatively low value. In contrast, when the purpose of a study is to test the behavioral effects of a costly intervention campaign, it will be necessary to specify a much larger effect size as being of minimal interest. For this reason, the specification of a given $d_{min}$ should be considered as a central topic of both a study’s pre-registration plan as well as its literature review. In general, and when feasible, larger values of $d_{min}$ are to be preferred as they represent more meaningful effects in a theoretical and practical sense; methodologically, they generate less overlap between the null and alternative distributions given a fixed sample size (i.e., higher falsifiability).

Another advantage of the strong-form construal is that it incentivizes researchers to gather large samples. This incentive is much more apparent compared to the traditional null hypothesis $P$-value, because it literally creeps into our interpretation of ‘evidence’: if we only gather small samples we know that, when evaluating the results, we will find larger overlap between the null and alternative distributions. As a result, a corroborator (if we already find one) will be considered of less value than it would have been if we had had a larger sample. Also, the larger the sample size, the smaller $P$ will be for relatively small deviations $\delta > d_{min}$. This means that larger sample sizes will make it easier to find corroborating information in favor of $\delta > d_{min}$ if, in reality, the population parameter is any greater than $d_{min}$. Hence, gathering larger samples will generally be more interesting for researchers and, at the same time, ensure a more risky test of hypotheses. In other words, in a strong-form set-up,
gathering large samples will be a win–win situation for both science and the scientist (which is not necessarily the case for null hypothesis tests, where scientifically meaningless deviations from 0 already spawn low \( P \)-values with large samples).

A third advantage of strong-form testing is that it underlines the value of oft-maligned distance measures such as \( P \), up and above confidence intervals (Cumming, 2014): when assessing the falsifying information of a finding in light of a fixed hypothesis, confidence intervals only provide binary information about a predicted parameter – that is, it lies within the interval or not. Distance measures are more specifically informative, as they quantify the extremity of an observation against a specific prediction. As we have mentioned earlier, measures such as \( P \) or \( \hat{P} \) and confidence intervals can be seen as complementary: we first test predictions by evaluating falsifying information as expressed through a distance measure; next, we may use an interval to (inductively) reason through the set of hypothetical parameter values for which the observation would not be among the \( 1 - \alpha \) most extreme observations. Alternatively, if we are unwilling to specify a given significance level, we may use severity curves to obtain an overview of \( \hat{P} \) values across hypothetical parameter values.

**Challenges, limitations and future directions**

Despite its potential for strengthening the statistical inferences of communication scientists, the strong-form framework carries along various challenges. Most crucially, it is in need of a generally established set of definitions and best-practices: as there are still many ways to define or conceptualize similar sets of ideas, researchers will need to add some technical detail to their statistical definitions and calculations when resorting to strong-form tests. While this might raise the bar for practically oriented scholars, we hope that the principles laid out in this paper are formulated clearly enough to be used as a reference guide.

A second challenge is finding a way to generalize the principles of strong-form testing to more complicated analyses. By this, we mainly refer to situations where the non-centrality parameter is not easily expressed in terms of standardized effect sizes, and/or the sampling distribution cannot be derived analytically. With regard to the first situation, it might happen that we are unable (or unwilling) to express effects sizes in standardized form: some methodologists have argued against the use of standardized metrics (e.g., Baguley, 2009) and, in some situations, there is no consensus on how a standardized effect should be defined at all (e.g., when variances are not assumed to be homogeneous, as in a Welch ANOVA). Under these circumstances, calculating the non-centrality parameter will often require us to separately define (1) raw effect sizes as well as (2) population variances. Of course, population variances are typically unknown, so the only way to proceed seems to be with using the sample variance to calculate the non-centrality parameter. This is not unreasonable – the sample variance is, after all, an unbiased estimator – but it does introduce an additional layer of variability: the sample variance has its own sampling distribution and may therefore generate both an under- or overestimation of the true population variance. The key question therefore is as follows: what would this use of a sample variance imply, exactly, for the conclusions drawn from a statistic such as \( \hat{P} \)? Detailing the answer to this question goes beyond the scope of the current paper, and we leave a full discussion to future studies. However, it seems that there are at least two competing dynamics to be considered: on the one hand, the sample variance will underestimate the population variance more than 50% of the time (its distribution is right-skewed). This means that, more often than not, the non-centrality parameter calculated through the sample variance will be greater than the non-centrality parameter obtained through the population variance. As a consequence, the overlap between the null and alternative distributions will be underestimated more than 50% of the time when using the sample variance, and quantities such as statistical power will be overestimated more than 50% of the time. On the other hand, as we know that \( \hat{P} \) will be calculated on the basis of a non-centrality parameter that is, on average, overestimated, \( \hat{P} \) should typically be overly conservative: that is, a given test statistic will, relatively speaking, generate a higher \( \hat{P} \) value (less corroborative information for \( \delta = d_{\min} \)) than it
would have had if we had used the population variance to calculate the noncentrality parameter. This would appear to provide yet-another incentive for researchers to obtain high sample sizes; with higher sample sizes, \( \hat{P} \) as calculated from sample variances would be underestimated less often.\(^7\)

Another complication for strong-form testing arises when non-central distributions are not easily expressed in analytical form. This occurs, for instance, for the widely used cases of logistic regression, multilevel models, and structural equation models. For logistic regression – which is a non-linear model – we need to consider that alternative distributions of the test statistic are contingent on the exact distribution of the independent variables (Demidenko, 2007); for multilevel models we need to specify separate variance components (Snijders, 2005); and for SEM we naturally need to take into account measurement reliabilities (Liu, 2014). While there are workarounds to approximate these non-central distributions without changing much to the principles for the General Linear Model (see Faul et al., 2009, for logistic regression, or Liu, 2014, for multilevel models and SEM) it seems useful to think about developing a framework for strong-form testing that does not require analytical solutions. For instance, one could consider extending the strong-testing framework to simulation-based methods which is, in fact, the way in which frequentist testing proceeds in physics or how power analyses are set up for complicated use-cases. A detailed treatment or validation of simulation-based applications is far beyond the scope of the present paper, so future studies should flesh out those principles in more detail. We believe, however, that this exercise would be particularly useful: a generalizable framework for strong-form testing would not just add to the statistical practices of communication scientists, but it could aid all of the social sciences evolve toward a common standard for critical hypothesis testing.

With all of this said, there is still one important caveat to be stressed: conducting what we called strong-form tests should not be conflated with ‘having strong theories.’ Strong-form testing makes a more stringent assessment of theoretical claims possible – at least compared to nil-null testing – but it provides no safeguard against testing theoretically meaningless (or methodologically contrived) claims. Hence, communication scientists should be wary not to overinterpret the results of a strong-form test – or any test, for that matter. A qualitative appraisal of theoretical arguments and methodological choices remains essential, regardless of the value obtained for any summary statistic such as \( P \) or \( \hat{P} \).

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Notes on contributors**

*Lennert Coenen* (PhD, KU Leuven) is an assistant professor at the Tilburg Center for Cognition & Communication (Tilburg University, The Netherlands) and a visiting professor at the Leuven School for Mass Communication Research (KU Leuven, Belgium). His research mainly aims to develop theory, metatheory, and methodology in the area of media effects on public opinion.

*Tim Smits* (PhD, KU Leuven) is a full professor in Persuasion and Marketing Communication at the institute for Media Studies, KU Leuven, Belgium. Tim published on various topics within these fields, but his main research focus pertains to persuasion and marketing communication dynamics that involve health and/or consumer empowerment and how these are affected by situational differences or manipulations. He also has a more methodological line of research on science replicability.

**Funding**

Lennert Coenen’s contribution to the paper was made possible (in part) by his postdoctoral fellowship at the Research Foundation Flanders [12J7619N].

\(^7\)All of this being said, initial simulations suggest that over- and underestimation of the population variance will occur with almost equal probability when errors are normal and sample sizes are reasonably large.
References

Aisbett, J., Lakens, D., & Sainani, K. (2020). Magnitude based inference in 23Relation to one-sided hypotheses testing procedures. SportRyiv. https://doi.org/10.31236/osf.io/pn9s3

Amrhein, V., & Greenland, S. (2017). Remove, rather than redefine, statistical significance. Nature Human Behaviour, 2 (1), 4–4. https://doi.org/10.1038/s41562-017-0224-0

Amrhein, V., & Greenland, S. (2018). Remove, rather than redefine, statistical significance. Nature Human Behaviour, 2 (1). 4. https://doi.org/10.1038/s41562-017-0224-0

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don’t expect replication. The American Statistician, 73(sup1), 262–270. https://doi.org/10.1080/00031305.2018.1543137

Asendorpf, J. B., Conner, M., de Fruyt, F., de Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. European Journal of Personality, 27(2), 108–119. https://doi.org/10.1002/per.1919

Baguley, T. (2009). Standardized or simple effect size: What should be reported? British Journal of Psychology, 100(3), 603–617. https://doi.org/10.1348/000712608X377117

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., & Cesarini, D. (2018). Redefine statistical significance. Nature Human Behaviour, 2(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155. https://doi.org/10.1037/0033-2909.112.1.155

Cousins, R. D. (2020). Connections between statistical practice in elementary particle physics and the severity concept as discussed in Mayo’s Statistical Inference as Severe Testing. ArXiv:2002.09713 [Hep-Ex, Physics:Physics, Stat]. http://arxiv.org/abs/2002.09713

Cumming, G. (2013). Cohen’s d needs to be readily interpretable: Comment on Shieh (2013). Behavior Research Methods, 45(4), 968–971. https://doi.org/10.3758/s13428-013-0392-4

Cumming, G. (2014). The new statistics: Why and How. Psychological Science, 25(1), 7–29. https://doi.org/10.1177/0956797613504966

De Groot, A. (1956/2014). The meaning of “significance” for different types of research (translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kiviet, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas). Acta Psychologica, 148, 188–194. https://doi.org/10.1016/j.actpsy.2014.02.001

Demidenko, E. (2007). Sample size determination for logistic regression revisited. Statistics In Medicine, 26(18), 3385–3397. https://doi.org/10.1002/sim.2771

Derksen, M. (2019). Putting Popper to work. Theory & Psychology, 29(4), 449–465. https://doi.org/10.1177/095935419838343

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., Johnson, B. K., Huskey, R., Schneider, F. M., Breuer, J., Parry, D. A., Vermeulen, I., Fisher, J. T., Banks, J., Weber, R., Ellis, D. A., & de Vreeze, C. (2020). An Agenda for open science in communication. Journal of Communication, 71(1), 1–26. https://doi.org/10.1093/joc/jqz052

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2), 175–191. https://doi.org/10.3758/bf03193146

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fisher, R. (1955). Statistical methods and scientific induction. Journal of the Royal Statistical Society: Series B (Methodological), 17(1), 69–78. https://doi.org/10.1111/j.2517-6161.1955.tb00180.x

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis (3rd ed.). Chapman & Hall/CRC.

Gigerenzer, G. (2004). Mindless statistics. The Journal of Socio-Economics, 33(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. European Journal of Epidemiology, 31(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3
Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, 73(sup1), 106–114. https://doi.org/10.1080/00031305.2018.1529625

Haig, B. D. (2020). What can psychology's statistics reformers learn from the error-statistical perspective? *Methods in Psychology*, 2, 100020. https://doi.org/10.1016/j.metip.2020.100020

Hawkins, A. T., & Samuels, L. R. (2021). Use of confidence intervals in interpreting nonstatistically significant results. *JAMA*, 326(20), 2068–2069. https://doi.org/10.1001/jama.2021.16172

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Sage Publications, Inc.

Lakatos, I. (1978). Falsification and the methodology of scientific research programmes. In J. Worrall & G. Currie (Eds.), *The methodology of scientific research programmes: Philosophical papers* (Vol. 1, pp. 8–101). Cambridge University Press.

Lakens, D., Adolff, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. F., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., & Zwaan, R. A. (2018). Justify your α. *Nature Human Behavior*, 2(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D. (2021). The practical alternative to the P value is the correctly used P value. *Perspectives on Psychological Science*, 16(3), 639–648. https://doi.org/10.1177/1745691620958012

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. L. M. (2008). A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. *Human Communication Research*, 34(2), 171–187. https://doi.org/10.1111/j.1468-2958.2008.00317.x

Liu, X., & Raudenbush, S. (2004). A note on the noncentrality parameter and effect size estimates for the F test in ANOVA. *Journal of Educational and Behavioral Statistics*, 29(2), 251–255. https://doi.org/10.3102/1076986029002251

Liu, X. S. (2014). Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques. Routledge.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science*, 57(2), 323–357. https://doi.org/10.1093/bjps/axl003

Mayo, D. (2018). Statistical inference as severe testing. *How to get beyond the statistics wars*. Cambridge University Press.

Mayo, D. (2019). P-value thresholds: Forfeit at your Peril. *European Journal of Clinical Investigation*, 49(10), e13170. https://doi.org/10.1111/eci.13170

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Meehl, P. E. (1967). Theory-Testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2), 103–115. https://doi.org/10.1086/288135

Meehl, P. E. (1990). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1(2), 108–141. https://doi.org/10.1207/s15327965pi0102_1

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *The Journal of Applied Psychology*, 84(2), 234–248. https://doi.org/10.1037/0021-9010.84.2.234

Neyman, J. (1942). Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society*, 105(4), 292–327. https://doi.org/10.2307/2980436

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241–301. https://doi.org/10.1037/1082-989x.5.2.241

PASS. (2020). *Power analysis and sample size software [Computer Software]*. NCSS, LLC. ncss.com/software/pass

Pastore, M. (2018). Overlapping: A R package for estimating overlapping in empirical distributions. *Journal of Open Source Software*, 3(32), 1023. https://doi.org/10.21105/joss.01023

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. Routledge and Kegan Paul.

R Core Team (2020). *R: A language and environment for statistical computing*. https://www.R-project.org/

Ruckdeschel, P., Kohl, M., Stabla, T., & Camphausen, F. (2006). S4 classes for distributions. *R News*, 6(2), 2–6. https://cran.r-project.org/web/packages/distrDoc/vignettes/distr.pdf

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, 16(4), 744–755. https://doi.org/10.1177/1745691620966795

Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1570–1573). Wiley.

Spanos, A. (2014). Recurring controversies about P values and confidence intervals revisited. *Ecology*, 95(3), 645–651. https://doi.org/10.1890/13-1291.1

Spanos, A., & Mayo, D. G. (2015). Error statistical modeling and inference: Where methodology meets ontology. *Synthese*, 192(11), 3533–3555. https://doi.org/10.1007/s11229-015-0744-y
Vermeulen, I., Beukeboom, C. J., Batenburg, A., Avramiea, A., Stoyanov, D., van de Velde, B., & Oegema, D. (2015). Blinded by the light: How a focus on statistical “Significance” may cause p-value misreporting and an excess of p-values just below .05 in communication science. *Communication Methods and Measures, 9*(4), 253–279. https://doi.org/10.1080/19312458.2015.1096333

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond “ p < .05. *The American Statistician, 73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Weber, R., & Popova, L. (2012). Testing equivalence in communication research: Theory and application. *Communication Methods and Measures, 6*(3), 190–213. https://doi.org/10.1080/19312458.2012.703834

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). CRC Press.