Why Figures with Error Bars Should Replace *p* Values

Some Conceptual Arguments and Empirical Demonstrations

Fiona Fidler[1] and Geoffrey R. Loftus[2]

[1]La Trobe University, Melbourne, Australia, [2]University of Washington, Seattle, WA, USA

Abstract. Null-hypothesis significance testing (NHST) is the primary means by which data are analyzed and conclusions made, particularly in the social sciences, but in other sciences as well (notably ecology and economics). Despite this supremacy however, numerous problems exist with NHST as a means of interpreting and understanding data. These problems have been articulated by various observers over the years, but are being taken seriously by researchers only slowly, if at all, as evidenced by the continuing emphasis on NHST in statistics classes, statistics textbooks, editorial policies and, of course, the day-to-day practices reported in empirical articles themselves (Cumming et al., 2007). Over the past several decades, observers have suggested a simpler approach – plotting the data with appropriate confidence intervals (CIs) around relevant sample statistics – to supplement or take the place of hypothesis testing. This article addresses these issues.

Keywords: confidence intervals, *p*-values, error bars, figures, statistical reform, statistical thinking

This article is divided into two sections. In the first section, we review a number of what we consider to be serious problems with NHST, focusing particularly on ways by which NHST could plausibly distort one's conclusions and, ultimately, one's understanding of the topic under investigation. In this first section, we also describe confidence intervals (CIs) and the degree to which they address the problems of NHST. In the second section, we present empirical data that extend prior findings indicating that what we consider to be the most serious of these problems – an unwarranted equation of "failure to reject the null hypothesis" with "the null hypothesis is true" – does indeed influence and bias interpretations of typical experimental results. In two experiments we compare the degree to which such false conclusions issue from results described principally by way of NHST versus from results illustrated principally by way of visually represented confidence intervals.

Part 1: Problems with NHST

Lurking close to the heart of scientific practice is *uncertainty*: Any sophisticated scientist would readily agree that the interpretation of a result coming out of a scientific study must be tempered with some probabilistic statement that underscores and quantifies such uncertainty. Recognition of the underlying uncertainty attendant to any scientific conclusion is essential to both scientific efficiency and scientific integrity; conversely, *failure* to recognize such uncertainty is bound to engender misunderstanding and bias. One theme that runs through this article is that NHST has the effect of sweeping such uncertainty under the rug, whereas use of CIs, especially when presented graphically, leaves the uncertainty in the middle of the floor for all to behold.

In many scientific studies – indeed, in the majority of those undertaken in the social sciences – uncertainty is couched in terms of the relation between a set of sample statistics measured in an experiment and the underlying population parameters of which the sample statistics are estimates. In what follows, we will, for the sake of expositional simplicity, constrain our discussion to the relations between sample and population *means*, although our arguments apply to any statistic-parameter relation.

In a typical experiment, there are J conditions and, accordingly, J sample means (the M_js) are measured. These M_js are presumed to be estimates of the J corresponding population means (the μ_js). Of fundamental interest in the experiment (although rarely stated explicitly as such) is the *pattern* of the μ_js over the J conditions. The experimenter, of course, does not know what the μ_js are; all that he or she has available are the measured M_js. Thus, the uncertainty lies in lack of knowledge about the expected discrepancy between each M_j and the corresponding μ_j of which the M_j is an estimate. If, in general, the uncertainty is small then

DOI 10.1027/0044-3409.217.1.27

pattern of the M_js may be construed as a relatively precise estimate of the underlying pattern of μ_js. Conversely, if the uncertainty is large, the pattern of the M_js may be construed as a relatively imprecise estimate of the underlying pattern of μ_js. Note that "precise" and "imprecise" estimates maps onto what, within the domain of NHST is normally referred to as "high" or "low" statistical power.

NHST: A Dichotomization of Conclusions

The process of NHST pits two hypotheses against one another: A specific *null hypothesis* that is almost always a *nil null*, stating that the μ_js are all equal to one another and an *alternative hypothesis* which, usually, is "anything else." If the differences among the M_js (embodied in "mean squares between") are sufficiently large compared to the error variability observed within groups (embodied in "mean squares within") then the null hypothesis is rejected. If mean squares between is not sufficiently large, then the null hypothesis is not rejected; technically, that is, one is left in a nonconclusive "limbo" state. Within the context of NHST, the uncertainty of which we have spoken is compressed into, and expressed by, a single number, the "*p*-value" whose simple, but often misinterpreted meaning is elucidated below.

The NHST process has associated with it some serious problems. Most have been discussed at length in past commentaries on NHST and we will not rediscuss them in detail here[1]. Briefly, they include the following.

The Null Hypothesis Cannot Be Literally True

According to the (typical) null hypothesis, every μ_j is identically equal to every other μ_j. In fact, however, in most branches of science such a state of affairs cannot be true, i.e., the μ_js will not equal one another to an infinite number of decimal places. As Meehl (1967), for example, has pointed out, "Considering . . . that everything in the brain is connected with everything else, and that there exist several "general state-variables" (such as arousal, attention, anxiety, and the like) which are known to be at least *slightly* influenceable by practically any kind of stimulus input, it is highly unlikely that *any* psychologically discriminable situation which we apply to an experimental subject would exert literally *zero* effect on any aspect of performance." (p. 109). Thus, to reject the null hypothesis as false does not tell an investigator anything that was not known already; rather rejecting the null hypothesis allows only the relatively uninteresting conclusion that the experiment had sufficient power to detect whatever differences among the μ_js that must have been there to begin with. As an analogy, no sane scientist would ever make a claim such as "based on spectrometer results, we can reject the hypothesis that the moon is made of green cheese." In this cartoonish context, it is abundantly clear that such a claim would be silly because the falsity of the hypothesis is so apparent a priori. Yet a logically isomorphic claim is made whenever one rejects a null hypothesis.

Misinterpretation of "Rejecting the Null Hypothesis"

Normally, either implicitly or explicitly, a typical results-section assertion goes something like, "Based on these data, we reject the null hypothesis, $p < .05$." In normal everyday discourse, such an assertion would be tantamount to – and is close to literally – saying, "Given these data, the probability that the null hypothesis is true is less than .05." Of course, as every introductory-statistics student is taught, this is wrong: The *p*-value refers to the opposite conditional probability, and instead implies, "Given that the null hypothesis is true, the probability of getting these (or more extreme) data is less than .05." However, in much scientific discourse, both formal and informal, this critical distinction is often forgotten, as people are seduced, by endless repetitions of "Based on these data we reject the null hypothesis," into believing, and acting on the validity of, the normal-discourse interpretation, rather than the nonobvious, mildly convoluted, statista-speak interpretation of the phrase.

What Does "$p < .05$" Mean, Anyway?

A corollary of this problem is in that "$p < .05$" does not actually refer to anything very interesting. Typically, the fundamental goal of NHST is to determine that some null hypothesis is false (let's face it; that's the kind of result that gets you a full professorship). So knowing the probability that a null hypothesis is false would be important and, notwithstanding the serious problem sketched in Point 2 above, might arguably warrant the rigid reliance placed on *p*-values for making conclusions. However, because, in fact, the *p*-value refers to something more obscure and considerably less relevant – the probability of the data given a null hypothesis – the importance of the *p*-value is, within our scientific culture, highly overemphasized. It is far more interesting to know (a) the magnitude of the difference or effect size, (b) the uncertainty associated with our effect estimate (e.g., standard error or CI), and (c) whether the estimate is within a clinically important or theoretically meaningful range. CIs give us (a) and (b) and should at least lead to thinking about (c).

[1] Examples of detailed discussions of these problems include: Bakan (1966), Carver (1978), Greenwald, Gonzalez, Harris, & Guthrie (1996), Loftus (1991; 1996), Lykken (1968), Meehl (1990), Rosenthal & Rubin (1985), Rozeboom (1960), Schmidt (1994); see Kline (2004) for an overview.

Accepting the Null Hypothesis

Above we observed that, "If mean squares between is not sufficiently large, then the null hypothesis is not rejected; technically, that is, one is left in a nonconclusive 'limbo' state." In point of fact – and we shall return to this issue in the experiments that we report below – humans do not like to be in nonconclusive limbo states, and "fail to reject the null hypothesis" often, implicitly, morphs into "the null hypothesis is true." This kind of almost irresistible logical slippage can, and often does, lead to all manner of interpretational mischief later on. If the statistical power to detect even a small effect is very high, then one might reasonably argue that accepting the null is unproblematic, particularly when a yes/no decision must be made in an applied setting. In practice, however, statistical power is rarely so high as to warrant this interpretation (e.g., Sedlmeier & Gigerenzer, 1989).

Failure to See the Forest for the Trees

If an experiment includes only two conditions, then the range of possible (qualitative) states of reality is limited: Either the two conditions differ or (putting aside, for the moment, Point 1 above) they do not. Frequently, however, experiments contain more than two conditions, i.e., in our notation, $J > 2$. When this is true, rejecting the null hypothesis reveals virtually nothing about what is of principal interest, namely the underlying pattern of population means. The typical practice at this point is to carry out a series of posthoc *t* tests comparing individual pairs of means, the result of which is usually a dichotomization of which of the $(Jx(J–1))/2$ pairs differ significantly and which do not. We assert that such a practice (a) encourages the misdeed of accepting null hypotheses (see Point 4 above) and (b) does little to provide any intuitive understanding of the overall nature of the underlying pattern of μ_js.

Emphasis on Qualitative Rather than Quantitative Results

It is a truism that a stronger quantitative result is more informative than a weaker, qualitative result that subsumes it. For instance, saying that "ingestion of two ounces of alcohol increased subjects' mean reaction time by 50 ms compared to ingestion of one ounce of alcohol" is more informative than saying, "subjects were slower with two compared to one ounce of alcohol." NHST, however, emphasizes qualitative conclusions, e.g., "The null hypothesis of no alcohol effect isn't true" or "two ounces of alcohol reduces reaction time by a significantly greater amount than one ounce."

This qualitative-at-the-expense-of-quantitative emphasis is most readily seen in too many results sections in which hypothesis-testing results are provided (typically in the form of *p*-values) but condition means are not. Failure to report this crucial information is perhaps more common than one might think: As reported by Fidler et al., (2005), in *Journal of Consulting and Clinical Psychology* – a leading clinical journal – only 60% of articles using ANOVA between 1993 and 2001 reported condition means or mean differences. (We do note that, to *JCCP's* credit, this figure had risen to 82% by 2003, which is certainly an improvement, but one that took serious editorial intervention to achieve, and is unfortunately not standard practice in most journals.)

CIs: A Direct Depiction of Uncertainty

CIs are common in many branches of science. A CI constructed around a sample statistic is designed to provide an assessment of the corresponding population parameter's whereabouts. They also provide a direct indication of the uncertainty attendant to interpretation of results that we sketched earlier: The larger the CI, the greater is the uncertainty. Ideally, this will be depicted *visually*, as part of a graphical representation of the experimental results. Unfortunately, graphical presentation of CIs is not currently common practice in psychology.

CIs address, to varying degrees, the problems with NHST that we enumerated above[2]. Here we sketch specifically how they do so.

The Null Hypothesis Cannot Be Literally True

A CI does not presume any single null hypothesis. Instead, we can investigate multiple hypotheses on a relevant scale, where values in the interval are more likely than those outside, and in turn, values at the center of the interval are more likely than those at the ends.

Misinterpretation of "Rejecting the Null Hypothesis"

Again, with a CI, there is no null hypothesis whose rejection can be misinterpreted. However, as Abelson (1997) warned, "Under the law of the diffusion of idiocy, every foolish application of significance testing is sooner or later going to be translated into a corresponding foolish practice for confidence limits" (p. 130). If one simply looks for whether zero, or some other null value, is in or outside of the interval, then substituting CIs for *p*-values achieves little. Because CIs make precision immediately salient we

[2] Examples of detailed discussion of CI advantages include: Cumming & Fidler (this issue), Cumming & Finch (2005, 2001), Loftus & Masson (1994), Schmidt (1996).

expect them to help alleviate this dichotomous thinking. This is one of the questions we address experimentally in Section 2.

What Does "$p < .05$" Mean, Anyway?

The analog to $\alpha = 05$, to which a *p*-value is compared, would be the arbitrarily chosen confidence level, typically 95%. A rough analog to the *p*-interpretation problem with NHST exists in the construction of CIs. Technically, a CI – say a 95% CI – is interpreted thusly: In the long run, 95% of all CIs generated by this specific process will include the relevant population mean. However it is usually interpreted as: A CI around a single sample mean has a 95% probability of including the relevant population mean.

This latter, somewhat Bayesian, interpretation is certainly more satisfying. However, strict frequentists will quickly identify it as a misconception and, in fact, a manifestation of the *inverse probability fallacy*. This fallacy is also widespread in the interpretation of *p*-values (Oakes, 1986; Haller and Krauss, 2002), where it results in the mistaken belief that $p(D|H) = p(H|D)$. In the context of *p*-values, the inverse probability fallacy has had drastic consequences. For example, Oakes holds it responsible for the neglect of statistical power (i.e., If I already know the probability of my hypothesis why would I care about the probability of detecting an effect of some specified size if it was really there? I wouldn't!). Our argument here is that the equivalent mistaken belief in CIs, i.e., that there is a 95% probability a single interval includes the population parameter, is far less likely to bring damaging consequences. Hoenig & Heisey made the same argument in 2001:

> Although we cannot demonstrate it formally, we suspect that imperfectly understood confidence intervals are more useful and less dangerous than imperfectly understood *p* values and hypothesis tests. For example, it is surely prevalent that researchers interpret confidence intervals as if they were Bayesian credibility regions; to what extent does this lead to serious practical problems? (p. 23).

Having said that, it is important to be aware of the limits of this imperfect definition. 'A CI around a single sample mean has a 95% probability of including the relevant population mean' should only be treated as correct if the existing data are the *only* source of information upon which judgments about population mean locations is based. Additional information can render the interpretation suspect. Imagine, for example, an experiment for which a one-tailed hypothesis test would be appropriate, e.g., an experiment in which the effect of amount of alcohol consumption on reaction time is measured. Longstanding existing knowledge would allow one to disallow the possibility that additional alcohol *decreases* reaction time and accordingly a one-tailed hypothesis test would be carried out. A CI around a difference score (more alcohol reaction time minus less alcohol reaction time) might reasonably, however, extend into negative values. One would not, in this situation, infer that the CI included the true population value with the intended 95%.

Accepting the Null Hypothesis

Use of a CI does not entail the dichotomous conclusion of rejecting or failing to reject any hypothesis. As discussed above, such interpretations fail to fully exploit the information a CI provides. However, construction of CIs provides one critical piece of relevant information, particularly if shown visually: It permits an immediate, visual assessment of how uncertain the observed pattern of sample means is as an estimate of the underlying pattern of population means. To the degree that the CIs are small, one infers that the sample mean pattern is a more precise reflection of the underlying population mean pattern; conversely, to the degree that CIs are large, one infers that the sample mean pattern is a less precise reflection of the underlying population mean pattern.

As a side product of this CI property, one is in a position to make judgments about the viability of the null hypothesis or of any other specific hypothesis that one wishes to consider. Suppose, for example, that the sample mean difference(s) were sufficiently small that if the corresponding population mean difference(s) were similarly small, then the null hypothesis could be "accepted for all intents and purposes." This issue is usually relevant in applied settings where some decision must be made (as in our experiments as described below). For example, suppose that two clinical treatments, A and B, were being compared and suppose further that Treatment A is less expensive than Treatment B. Obviously, a finding of essentially no difference between the treatment outcomes (as in Panel 1 of Figure 1), would suggest supremacy of Treatment A and would, thus, be interesting and useful. Now suppose that, indeed, the sample means were relatively close to one another but not identical, i.e., that the null hypothesis of no difference was not rejected. Should this null result still be used as a basis for declaring Treatment A's superiority, given that it is cheaper and not statistically different from Treatment B? Construction of CIs would complete the picture in terms of the degree to which such a conclusion would be warranted. Large CIs would demonstrate immediately and intuitively that such a conclusion might be premature (as in Panel 2 of Figure 1), while small CIs would imply that such a conclusion would be safer to make (as in Panel 3 of Figure 1). We address this issue directly in the experiments that we report below.

Failure to See the Forest for the Trees

Much of what is relevant to this issue was discussed in our last point. A presentation of sample means with CIs – particularly a *visual* presentation – allows an immediate and

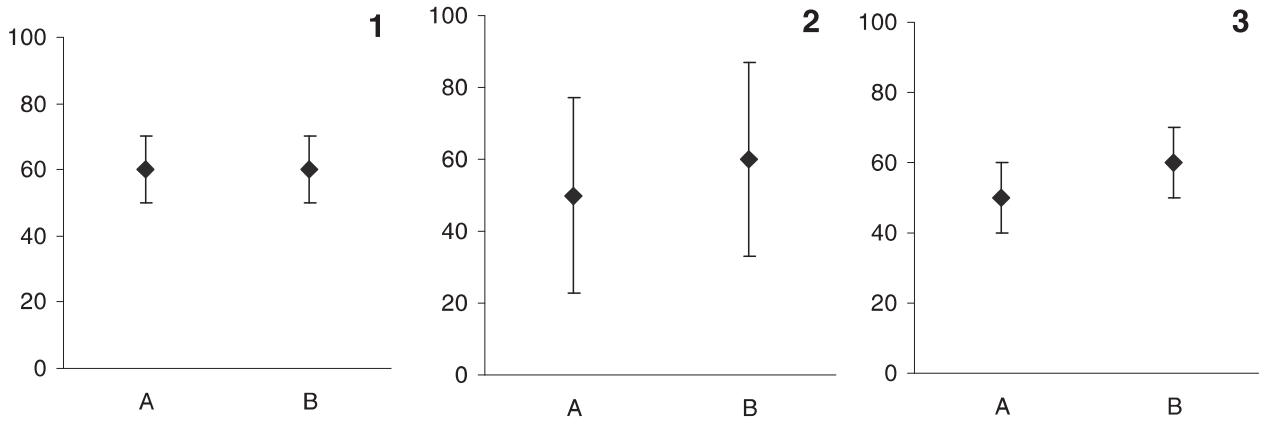


Figure 1. Three hypothetical outcomes (panels 1, 2, 3) of two treatments (A, B). The Y axis in each panel shows percentage improvement after treatment. The error bars are 95% confidence intervals. Panel 1 shows a scenario in which treatments A and B are exactly the same; so if A is less expensive it would obviously be preferred. Panels 2 and 3 show a scenario in which A and B are different to some statistically nonsignificant extent. The precision in panel 3 is much greater than in panel 2, as can be seen from the relative confidence interval width. We can therefore be more certain in proceeding with the cheaper treatment A in a panel 3 world than we can in a panel 2 world; in panel 2, study precision (roughly equivalent to statistical power) is low, and uncertainty about the true mean difference is high.

intuitive assessment of: (a) via the pattern of sample means itself, the best experimental estimate of the corresponding pattern of population means and (b) via the CI sizes, the degree to which the pattern of sample means should be taken seriously as a reflection of the underlying pattern of population means.

Emphasis on Qualitative Rather than Quantitative Results

Much of what is relevant to this issue was discussed above. A visual assessment of sample means plus CIs provides not only a sense of the qualitative pattern of means, but also a sense of the *pattern*, which is to say the magnitudes of all mean differences.

Part 2: Empirically Investigating the Cognitive Benefits of CIs

A variety of experiments have sought to determine whether the alleged problems with NHST and the concomitant virtues of using CIs are borne out in actual practice. The former (i.e., studies of NHST misconceptions) are far more common and have a much longer history (e.g., Falk & Greenbaum, 1995; Lecoutre, Poitevineau, & Lecoutre, 2003; Oakes, 1986; Rosenthal & Gatio, 1963) than parallel studies of CI understanding and interpretation. This imbalance in research is problematic for statistical reform: If the most common arguments for abandoning *p*-values are the severe and robust misconceptions associated with them and the dichotomous thinking that NHST promotes, then we should expect evidence that the alternatives offered to replace *p*-values are relatively free of such misconceptions and lead to a more sophisticated, less dichotomous approach to interpretation. Yet, recommendations for *p*-value replacements or supplements have not been investigated empirically. The following studies are an attempt to build such an evidence base.

One of the most serious problems associated with NHST is interpretation of statistical nonsignificance as evidence of no effect (see above). This misinterpretation is especially problematic in disciplines where statistical power is, on average, low and routinely unreported. We know from Cohen (1962) and subsequent studies in that tradition (e.g., Sedlmeier & Gigerenzer, 1989; Rossi, 1990) that the statistical power in psychological studies (for typical, medium effects) is roughly 50%. We also know from journal studies that statistical power is reported in less than 5% of psychology journal articles (Finch, Cumming, & Thomason, 2001; Fidler et al., 2005). Combined, these two practices leave considerable scope for this misconception to play out.

In all disciplines, misinterpreting a statistically nonsignificant result as evidence of "no effect" (without accounting for statistical power) is a mistake, but in some disciplines the consequences may be catastrophic. For example, in ecology false-negative results may lead directly to a lack of necessary conservation or precautionary action in endangered populations that have little scope to recover from the error. Similarly, in medicine, such errors can mean missing potentially life-saving treatments. As Altman and Bland (1995) reminded us "Absence of evidence is not evidence of absence." (p. 311). In psychology, too, there can be drastic consequences, including delays in application of potentially useful interventions.

It might seem as if simply reporting statistical power should rectify this misinterpretation. However, one conclusion we can draw from the experimental results below is

that reporting power is not sufficient to resist the lure of the "significant-nonsignificant" dichotomy. The studies below provide empirical evidence that presenting (at least these simple results) as visual CIs leaves students much less prone to this misconception, even when compared to "best practice" NHST presentations (i.e., those that include complete information including statistical power).

The two experiments that follow were conducted with undergraduate students in environmental science and ecology. Criticisms of NHST and calls for changes to practice in ecology have existed independently but in parallel with those in psychology. Misconceptions of *p*-values typical in the psychology literature have also been identified in ecology journals (e.g., Fidler, Burgman, Cumming, Buttrose, & Thomason, 2006). In this sense, the disciplines are comparable. Furthermore, the ecology students in our experiments receive virtually identical training in probability and statistics as psychology students at the same university. In fact, several may have taken the exact same service courses offered by the statistics department. The admission procedure and requirements of entry into an ecology major are also very similar to those for a psychology major. Finally, the scenarios and questions we ask in these surveys are parallel to many typical scenarios in psychology, and could easily be translated into psychology problems. For example, the uncertainty around the air quality measurement in the Experiment 2 could easily be uncertainty around an IQ estimate (or some other test score) in a situation where a decision about classification or diagnosis needs to be made. We, therefore, argue that the effects here can reasonably be generalized beyond the domain investigated in the experiments.

Experiment 1: Visually Presented CIs Result in Fewer Misconceptions than NHST

Experiment 1 investigated whether CIs reduce the temptation to interpret statistical nonsignificance as "no effect" in experimental scenarios with low statistical power. We focus on low power scenarios because, as indicated earlier, they are typical in many disciplines. The context of our study – environmental science and ecology – is no exception. For example, in studying threatened or endangered populations, sample sizes are constrained by small population sizes. In almost all cases, there are ethical or legal restrictions on areas being investigated and/or to simulating predicted environmental change experimentally. Adding to this problem is that statistical power is rarely reported in these research fields. In our recent survey of conservation biology journals (Fidler et al., 2006) only 8% of articles reported a statistical power calculation, yet *almost half* of the articles surveyed interpreted statistical nonsignificance as evidence for no effect. Because CIs make information about precision salient, particularly when they are presented graphically, we expected fewer misinterpretations of this kind in scenarios where this format was used.

Method

Participants, Materials, and General Design

Participants were 79 final-year Bachelor or Masters students from three separate environmental science classes – "Environmental Risk Assessment," "Environmental Problem Solving" and "Environmental Risk Assessment (Intensive)" – at the University of Melbourne, in Australia[3]. All participants had at least one prior semester of statistics, and were more than half-way through a second quantitative course.

Students viewed scenarios that required them to make conclusions based on certain statistical outcomes. All were randomly assigned to two of four possible scenarios (see below) with fictional data and asked to answer some multiple-choice questions. Students saw one scenario with NHST results, and a second scenario with CI results. Every effort was made to match scenarios for perceived environmental importance, interest, and accessibility. Scenarios were developed in consultation with two PhD ecologists to improve plausibility and symmetry. Even so, we used four scenarios in total, rather than two, to assess scenario effects.

The key characteristic of the scenarios is that they entailed statistically nonsignificant results, along with low statistical power (0.38–0.60). In addition, they entailed ecologically nontrivial observed effect sizes (the equivalent to "clinically important effect sizes"), by which we mean observed effect sizes that closely approached a biologically dangerous or important threshold or slightly exceeded the threshold. All scenarios were simple research designs, entailing either a single sample or two independent groups. The content varied from soil and water contaminants with potential human health effects, to the population decline of popular and endangered flora and fauna (see below).

In one version of each scenario, the results were presented as a *t* test. In these cases the *t* statistic was accompanied by a corresponding mean difference, standard deviation, degrees of freedom, *p* value, and a priori statistical power calculation for a predetermined biologically important effect. In the other version of each scenario, the results were presented graphically with CIs. In each individual survey, the scenario introductions were followed by either an NHST version of the results or a CI version.

[3] The University of Melbourne is one of the top three universities in Australia, and entry is extremely competitive. These were very bright students who had been taught by well-respected academics. At the time of the survey, students were enrolled in one of the listed courses that were taught by Professor Mark Burgman.

Scenarios. Briefly, the scenarios themselves were these:
- Scenario 1. "Toe-clipping is commonly used to mark frogs in population ecology studies because other methods of marking don't work on their skin. It is a valuable technique but there is some controversy over whether it affects recapture rates and, therefore, frog survival. This study examined the decline in recapture rate of frogs that had toes clipped . . ."
- Scenario 2. "Cadmium is a toxic heavy metal used, amongst other things, to make batteries. The cadmium level in stream near a battery factory has just been monitored . . ."
- Scenario 3. "A new park land is being developed near an old gas works. A lot of soil has already been cleaned and replaced. Since the clean up, the concentration of petroleum has been surveyed . . ."
- Scenario 4. "The monkey puzzle tree is a vulnerable species, endemic to South America. A study recently investigated whether two populations of monkey puzzle trees could be mixed for reforestation. If there are sufficient genetic differences between the two populations they should be kept separate; if not, they can be mixed. One important and common measure of genetic difference is the 'root to shoot' ratio, which measures how drought tolerant the trees are . . ."

In each scenario, subjects were given an explicit definition of the null hypothesis, which, for the four scenarios were:
- Scenario 1. Zero decline in frog recapture rate.
- Scenario 2. Normal background level of cadmium = 1ppb.
- Scenario 3. Average nonharmful level of petroleum = 2000 mg/kg.
- Scenario 4. Root-to-shoot ratio = 1.

In addition, subjects were provided with information about the size of a biologically important effect. In keeping with typical practice, the biologically important effects did not correspond to the null hypotheses above. The null hypotheses above are equivalent to nil nulls, which, as noted earlier, are hypotheses of no effect. In contrast, the biologically important effects given were:
- Scenario 1. Frog recapture rate decline of 10%.
- Scenario 2. Environmental Protection Agency maximum acceptable level = 5ppb cadmium.
- Scenario 3. 5000 mk/kg of petroleum is dangerous to human health.
- Scenario 4. Root-to-shoot ratio = 5 is of substantive genetic importance.

Figure 2 shows examples of one scenario (toe-clipping of frogs) in both formats. Subjects were asked to answer by circling one of the five statements below.

"In response to this information, the researcher who conducted this study should conclude that:"
- There is strong evidence in support of an important effect.
- There is moderate evidence in support of an important effect.
- The evidence is equivocal.
- There is moderate evidence of no effect.
- There is strong evidence of no effect.

The exact wording of these statements changed with each scenario depending on what the particular "effect" was. For example, in Scenario 1 the final option read: "There is strong evidence that toe clipping does *not* cause unacceptable decline." In Scenario 2, it read: "There is strong evidence that the factory has *not* breached EPA [Environmental Protection Agency] standards." (Underlining in original.)

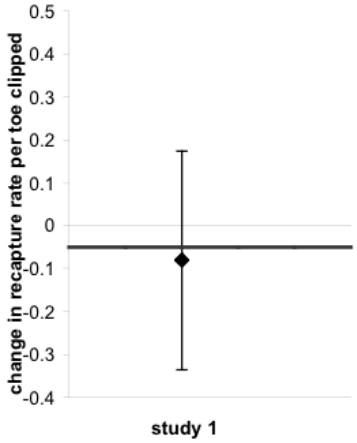
Responses indicating moderate or strong evidence for the null hypothesis were considered misconceptions, because, as mentioned, in all scenarios: (a) the statistical power of the scenario was low (power between 0.38 and 0.60) and (b) observed effect sizes were nontrivial in comparison to biologically important effects. Therefore, accepting or "failing to reject" the null was an uncontroversial error, and entailed interpreting statistical nonsignificance as "no effect." More appropriate responses noted the lack of power or precision and deemed the evidence equivocal, or attended to the large effect sizes and suggested that the evidence favored the alternative hypothesis.

Results and Discussion

Sixty-one percent (48 of 79, 95% CI: 50 to 71%) of students did *not* demonstrate the misconception that statistical nonsignificance means "no effect" when given results presented in the NHST format. This in itself is notable. Previous research has found this misconception to be far more widespread (Haller & Krauss, 2002; Oakes, 1986). However, it is important to note that these students had, throughout the semester, received several warnings of the misconception, along with formal instruction regarding statistical power analysis. Also, statistical power was clearly stated in all scenarios, and the biologically important effect size was stated. This reporting of what might be considered "best practice" NHST is far from typical, as discussed in the introduction to this experiment. Given that the presentation of results in these scenarios was much more complete than a typical journal article, it should, perhaps, be alarming that still 39% (31 of 79; 95% CI: 29 to 50%) *did* demonstrate the misconception.

Of those 31 students who demonstrated the misconception in the NHST scenarios, an overwhelming majority (87%; 95% CI: 71%, 95%) answered correctly when they were instead shown CI scenarios (presentation order was counterbalanced). Almost a third (32% of 31; 95% CI: 19%, 50%) moved 1 point on the 5-point Likert scale *away* from statements of accepting the null when given a CI; 52% (95% CI: 35%, 68%) moved 2 points and 3% (95% CI: 0 to 16%) moved fully 3 points. This amounts to an average shift of 1.67 points on a 5-point scale (see Figure 3).

Confidence Interval Picture Format



Toe-clipping is commonly used to mark frogs in population ecology studies because other methods of marking don't work on their skin. It is a valuable technique but there is some controversy over whether it affects recapture rates and, therefore, frog survival.

This study examined the decline in recapture rate of 60 frogs that had toes clipped.

In the figure above, the Y axis shows proportion change in recapture rate (negative values show proportion decline, positive values show increase). The horizontal line crossing at 0 indicates no effect on recapture rate. The thicker, horizontal line crossing at -.05 indicates the minimum decline we understand to be ecologically unacceptable. If the true proportion decline exceeds .05, toe clipping is an unacceptable practice. The black diamond is the change in recapture rate for this sample; the error bar is a 95% confidence interval.

NHST Format

Toe-clipping is commonly used to mark frogs in population ecology studies because other methods of marking don't work on their skin. It is a valuable technique but there is some controversy over whether it affects recapture rates and, therefore, frog survival.

This study examined the decline in recapture rate of 60 frogs that had toes clipped.

The minimum ecologically unacceptable decline in recapture rate is known to be .05. If the true proportion decline exceeds .05, toe clipping is an unacceptable practice.

The proportion decline in this sample was .08. This proportion (.08) is statistically *not* significantly different from zero (one sided t test = 1.1, p =.27; df = 59). The *a priori* statistical power of this test, to detect a decline of .05, was 40%.

Figure 2. The toe-clipping scenario in two formats – confidence-interval picture and NHST. In both the effect is biologically important, and the statistical power (or precision) low.

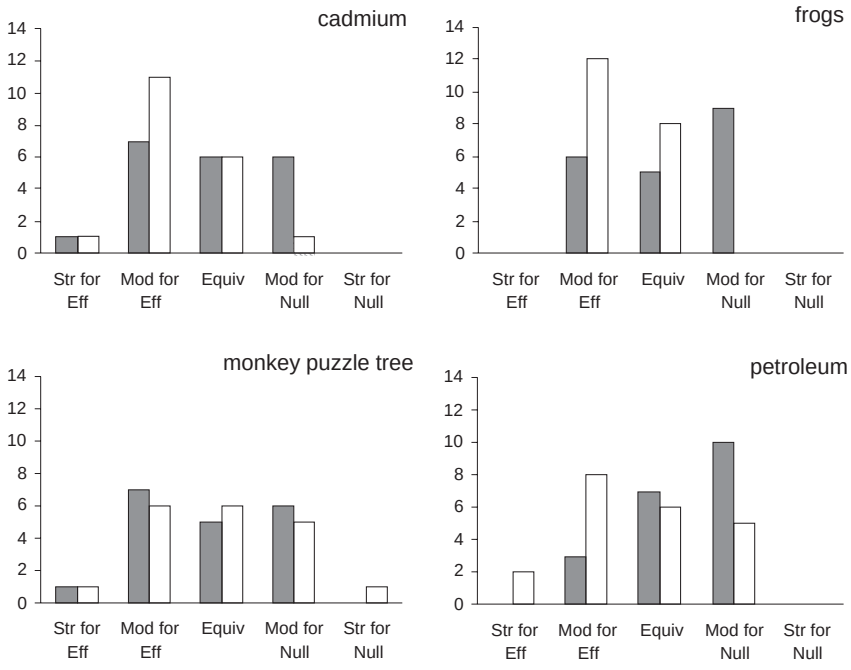


Figure 3. Frequency of five response types (left to right: strong support for effect, moderate support for effect, equivocal, moderate support for null, strong support for null) for the four scenarios when results were presented in NHST format (shaded bars) and CI format (white bars). Number of respondents in each scenario: cadmium n = 20 (21 in CI format); frogs n = 20; monkey puzzle tree n = 19; petroleum n = 20 (21 in CI format).

Table 1. Percentage of students who agreed that results moderately or strongly supported the null hypothesis (i.e., demonstrating the misconception that statistical nonsignificance equals "no effect") for each scenario and format

	Results as NHST % (n)	Results as CI picture % (n)	Improvement[a] with CI picture (NHST-CI)	95%CI[b] for improvement
Cadmium	30.0 (20)	4.8 (1 of 21)	25.2	3% – 50%
Frog	45.0 (20)	0.0 (0 of 21)	45.0	23% – 67%
Monkey puzzle	31.6 (19)	36.8 (7 of 19)	–5.2	–35% – 25%
Petroleum	50.0 (20)	23.8 (5 of 21)	26.2	–2% – 55%
Average	39.1	16.4	22.8	

[a]Negative values reflect more misconceptions in confidence-interval format of the scenario than in the NHST format. [b]Confidence intervals calculated according to method recommended for proportions by Newcombe and Altman (2000).

So far, this is an undeniably impressive result in favor of using CIs rather than NHST as a means of conveying the meaning of a data set. However, what of students who did not demonstrate the misconception to start with? Recall that 61% of students did not demonstrate the misconception in the NHST scenario. Were they perhaps led astray by the CI format? Table 1 presents these results. Any "reverse" effect is obviously undesirable, although in this case it seems isolated to one particular scenario – the monkey puzzle tree – where we suspect the direction of a "positive" result may have been ambiguous.

Conclusion

The results show promise for visually presented CIs. Given, in particular, (a) variability across scenarios and (b) that NHST used was best practice (e.g., a priori power, biologically important effect-size clearly stated) these results should be interpreted as encouraging. This misconception that statistical nonsignificance means no effect is not entirely absent when results are presented in CI picture format (as we might first have optimistically expected), but it is considerably less frequent.

Experiment 2: A Fully Within-Groups Replication

In Experiment 2, Experiment 1 was partially replicated as a fully within-subjects design, with a new sample of 55 second-year ecology subjects at the University of Melbourne. These participants were, on average, a year behind the previous sample and so arguable less statistically sophisticated than the previous sample. All had at least been exposed to both CIs and NHST during the previous semester.

Method

Each student was given a single research scenario and presented with both CI and NHST presentations of the results. The presentation order of these formats was counterbalanced, such that roughly half of students saw the NHST format followed by the CI format and the remainder saw the CI format followed the NHST format. This replication was designed to eliminate any confounding effects of scenario content, including varying effect sizes, levels of statistical power, and subjects of study. The scenario was introduced as follows:

> There are concerns about the air quality in a freeway tunnel. This study monitored the concentration of carbon monoxide (CO) during peak-hour traffic over 2 weeks, taking a total of 35 samples. Normal background levels of carbon monoxide are between 10–200 parts per million (ppm). A 1 h exposure time to CO levels of 250 ppm can lead to 5% carboxylated hemoglobin in the blood. Any level above this is abnormal and unsafe. If the true level of CO concentration in the tunnel exceeds 250 ppm, the tunnel will be closed and a surface road built. However, the surface road proposal has problems of its own, including the fact that threatened species inhabit an area near the surface site. First consider Presentation A. Please answer the question following Presentation A and then move on to Presentation B.

As in Experiment 1, the scenario was statistically nonsignificant with low statistical power and an effect size close to the biologically important cut off (i.e., observed effect 230 ppb). In addition, this scenario also included an incentive against overly precautionary answers that subjects might be naturally biased toward – the economic and biological costs of the tunnel alternative.

Question 1 was exactly the same as for Experiment 1, i.e., it asked whether the results provided strong or moderate evidence for an unsafe circumstance (the alternative hypothesis), strong or moderate evidence for a safe circumstance (the null) or whether the evidence was equivocal. As in Experiment 1, responses suggesting that results provided moderate or strong evidence for the null hypothesis were considered misconceptions, for the same reason: Statistical power was low and the effect size was nontrivial in comparison to biologically important effects, and so accepting the null was an uncontroversial error.

Results and Discussion

Given the NHST format, 44% (24 of 55; 95% CI: 31%, 57%) of students inappropriately claimed the results as ev-

idence *for* the null hypothesis. Less than half as many (18%, 10 of 55; 95% CI: 10%, 30%) made this mistake in the CI condition.

In Experiment 1 there appeared to be a reversal effect, i.e., some (17%) students demonstrated the misconception only with CIs and not at all with NHST. In this replication, which eliminated the confounding effect of certain scenarios, this reversal effect was much less pronounced (6%, 2 of 31 who did not have misconception in the NHST format).

The average shift away from misconceptions in the CI conditions was 1 point on the 5-point scale – a substantial effect, but perhaps not quite as large as we might have anticipated. However, the effect may be diluted by a learning effect. Indeed, students who saw the CI first appeared to do better with the NHST format than students who did not. (Numbers here are too small to analyse in any formal way.)

Conclusion

This within-groups replication provided even stronger evidence that CIs have a cognitive advantage over NHST. Visually represented confidence-intervals substantially alleviated the misinterpretation of statistically nonsignificant results as evidence for the null hypothesis, over and above a best practice NHST report including statistical power.

General Conclusions

The process of NHST has an enormous amount of inertia: In the social sciences, and other sciences as well, it has been the overwhelming data-analysis procedure for almost as long as these sciences have lain claim to be sciences. However, it is seriously flawed for, at the very least, the reasons that we articulated earlier. At its heart, it fails to underscore the magnitude and nature of uncertainty associated with any scientific result – uncertainty that is critical for understanding the data set, and, hence, uncertainty that the researcher should place front and center.

CIs have many theoretical advantages over NHST: They always include an estimate of the magnitude of the effect and information about precision, and they lend themselves readily (in most cases) to graphical representation. They do not necessarily entail a dichotomous decision on the basis of single studies, and should, therefore, help build a more cumulative approach to scientific knowledge. However, whether in practice CIs afford these benefits remains an open question. We have, for some time, had evidence of the widespread misconceptions about *p*-values, and yet there has been relatively little empirical study of the most commonly proposed alternatives. Here we have presented two studies that help to build an evidence base for the shift away from statistical significance and toward a science of estimation.

Acknowledgments

Supported by an Australian Research Council Postdoctoral Fellowship to Fiona Fidler and NIMH Grant MH41637 to Geoffrey Loftus.

References

Abelson, R.P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12–15.

Altman, D.G., & Bland, M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal, 311,* 485.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399.

Cohen, J. (1962). The statistical power of abnormal – social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A. et al. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18,* 230–232.

Cumming G., & Finch S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–574.

Cumming G., & Finch S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist, 60,* 170–180.

Falk, R., & Greenbaum, C.W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5,* 75–98.

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. *Educational and Psychological Measurement, 61,* 181–210.

Fidler, F., Burgman, M., Cumming, G. Buttrose, R. & Thomason, N. (2006). Impact of criticism of null hypothesis significance testing on statistical reporting practices in conservation biology. *Conservation Biology, 20,* 1539–1544.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P. et al. (2005). Toward improved statistical reporting in the Journal of Consulting and Clinical Psychology. *Journal of Consulting and Clinical Psychology, 73,* 136–143.

Greenwald, A.G., Gonzalez, R., Harris, R.J., & Guthrie, D. (1996). Effect Sizes and *p*-values: What should be reported and what should be replicated? *Psychophysiology, 33,* 175–183.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7,* 1–20.

Hoenig J.M., & Heisey D.M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55,* 19–24.

Kline, R.B. (2004). *Beyond significance testing: Reforming data-analysis methods in behavioral research.* Washington DC: American Psychological Association.

Lecoutre M.-P., Poitevineau J., & Lecoutre B. (2003). Even statisti-

cians are not immune to misinterpretations of null hypothesis significance tests. *International Journal of Psychology, 38*, 37–45.

Loftus, G.R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology. 36*, 102–105.

Loftus, G.R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 161–171.

Loftus, G.R., & Masson, M.E.J. (1994) Using confidence intervals in within-subjects designs. *Psychonomic Bulletin & Review, 1*, 476–490.

Lykken, D.T. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70*, 131–139.

Meehl, P.E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103–115.

Meehl, P.E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, Monograph Supplement* 1–V66.

Newcombe, R.G., & Altman, D.G. (2000) Proportions and their differences. In D.G. Altman, D. Machin, T.N. Bryant, & M.J. Gardner (Eds.) *Statistics with confidence* (pp. 39–50). London: BMJ Books.

Oakes, M.W. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* Chichester: Wiley.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55*, 33–38.

Rosenthal, R., & Rubin, D.B. (1985). Statistical analysis: Summarizing evidence versus establishing facts. *Psychological Bulletin, 97*, 527–529.

Rossi, J. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*, 646–656.

Rozeboom, W.W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*, 416–428.

Schmidt, F.L. (1994). *Data analysis methods and cumulative knowledge in Psychology: Implications for the training of researchers.* APA (Division 5) Presidential Address.

Schmidt, F.L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115–129.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309–315.

Geoffrey R. Loftus

Department of Psychology
Box 351525
University of Washington
Seattle, WA 98195-1525
USA
Tel. +1 206 543-8874
Fax +1 206 685-3157
E-mail gloftus@u.washington.edu