CANADIAN PSYCHOLOGICAL ASSOCIATION
SOCIÉTÉ CANADIENNE DE PSYCHOLOGIE

Canadian Psychology / Psychologie canadienne

ISSN: 0708-5591

2020, Vol. 61, No. 4, 331–341
http://dx.doi.org/10.1037/cap0000200

Post *p* Value Education in Graduate Statistics: Preparing Tomorrow's Psychology Researchers for a Postcrisis Future

Rex B. Kline
Concordia University

The first course in statistics plays an important role in many graduate programs, but instructors face challenges. One is that the standard curriculum in many undergraduate statistics courses overemphasizes traditional significance testing, which fills students' head with myths that work against the development of statistical literacy. There is also great variety in the quality of undergraduate courses, so new graduate students can vary markedly in their knowledge about statistics at the beginning of a graduate course. The aim of this article is to offer constructive advice for those who teach first courses in graduate statistics, including suggestions about how to assess background statistical knowledge and take early corrective action to address knowledge deficits. Essential topics that should be covered in a reform-oriented, post *p* value graduate statistics course are also recommended. The aim is help instructors promote a more modern—and truly graduate-level—understanding of statistics among their students.

Public Significance Statement
How to better educate student researchers in psychology and related disciplines in modern methods of statistics and data analysis that also promote better practices, such as replication, is the focus of this work. Suggestions about how to modernize course content and deal with potential obstacles that can arise in teaching graduate-level statistics courses are also considered.

Keywords: graduate statistics, teaching strategies, statistics reform, new statistics, post *p* value statistics

Supplemental materials: http://dx.doi.org/10.1037/cap0000200.supp

> The windows must be flung open again, we must see the world, the sky and the earth once more.
>
> —Emeritus Pope Benedict XVI (Waters, 2017, ¶13)

The *replication crisis* in psychology is characterized by recent failures to reproduce findings from classical laboratory studies (e.g., Open Science Collaboration, 2015; but see also Gilbert, King, Pettigrew, & Wilson, 2016) and a longer-term dearth of published studies explicitly devoted to replication (e.g., Makel & Plucker, 2014). It is part of an even bigger *credibility crisis* about psychology research and also the wasting of perhaps most of the research funds in some areas (Szucs & Ioannidis, 2017). Other aspects of the credibility crisis about psychology research include the following:

1. *Measurement crisis*, or the widespread failure of researchers to estimate and report the reliabilities of scores analyzed, even when using statistical techniques, such as multiple regression, that allow for no measurement error in predictors (e.g., Aguinis, Edwards, & Bradley, 2017; Meier, 1994). An aggravating factor was the disappearance in the 1980s–1990s of psychometrics courses from many undergraduate and graduate psychology programs in North America (Aiken et al., 1990). Although a slight improvement in measurement training was reported in a follow-up survey by Aiken, West, and Millsap (2008), the median length of measurement-related course work was only 4.5 weeks (i.e., still inadequate).

2. *Reporting crisis,* or the realization that critical information supporting replication, results synthesis (i.e., meta-analysis), and scientific transparency is omitted in too many journal articles for empirical studies, such as randomized clinical trials (e.g., Grant, Mayo-Wilson, Melendez-Torres, & Montgomery, 2013).

3. *Significance testing crisis*, or the ongoing controversy, now occurring in many disciplines, about the proper role of statistical significance testing, *if any*, in data analysis (e.g., Amrhein, Greenland, & McShane, 2019; Gelman, 2018; Wasserstein, Schirm, & Lazar, 2019).

This article was published Online First August 13, 2020.
Correspondence concerning this article should be addressed to Rex B. Kline, Department of Psychology, Concordia University, 7141 Sherbrooke Street West, Montréal, QC H4B 1R6, Canada. E-mail: rex.kline@concordia.ca

Journal article reporting standards for both quantitative and qualitative research in psychology (Appelbaum et al., 2018; Levitt et al., 2018) and other disciplines, such as those available on the Enhancing the Quality and Transparency of Health Research (EQUATOR) Network (n.d.), should address the reporting crisis. One hopes that improved reporting about psychometrics would help to ease the measurement crisis and also that more complete and transparent reporting in general would facilitate replication—for example, see Appelbaum et al. (2018, pp. 7, 17) for reporting standards on, respectively, psychometrics and replication.

The significance testing controversy actually consists of three overlapping subcrises, all of which contribute to the replication, reporting, and measurement crises. They are briefly stated and elaborated afterward:

1. Many, if not most, p values reported in empirical studies could be wrong.

2. Misinterpretation of p values is widespread, and altogether they can create false confidence in analysis results.

3. Through p-hacking, significance testing can engender scientific fraud, intentional or unintentional, which can also seriously erode the credibility of psychology research.

Outcomes of significance testing in the form of p values have been reported in most published empirical studies in psychology since about the mid-1950s (e.g., Hubbard & Ryan, 2000). Many—and perhaps most—of those p values, though, could be wrong due to violations of core assumptions in significance testing—such as random sampling from population distributions with known characteristics (e.g., normality, homoscedasticity) and absence of all sources of error other than sampling error—that are generally implausible. For example, too many researchers fail to verify distributional or other requirements of significance tests and, if they actually do report such results, incorrect methods may have been used (e.g., significance tests for normality; Hoekstra, Kiers, & Johnson, 2012; Kline, 2013).

Most researchers do not correctly understand p values, even among researchers with the highest levels of statistical training (McShane & Gal, 2016). For example, about half of psychology professors—including those who teach statistics or methods courses—endorsed the false belief that $1 - p$ is the probability that a result in one sample will replicate in a new sample (Haller & Krauss, 2002; Oakes, 1986). This misconception can lead to the false hope that "stylized facts" gleaned from applying significance tests to noisy data in small samples will replicate but, alas, they generally do not (Gelman, 2018). Authors in 402 of 791 (51%) empirical studies reviewed by Amrhein et al. (2019) committed the zero fallacy, which means that the absence of statistical significance is wrongly interpreted as meaning that the population effect size is zero. There are many other cognitive distortions associated with significance testing (Cumming & Calin-Jageman, 2017; Kline, 2013; Lambdin, 2012), and most involve exaggerating what can be inferred from either rejecting or failing to reject a null hypothesis.[1] This myriad of false beliefs can lead to confirmation bias where evidence that supports the researcher's hypotheses is accepted uncritically but evidence to the contrary is overlooked—see Calin-Jageman and Cumming (2019) for examples.

The phenomenon of p-hacking is now more widely known, and it concerns the fact that basically *any* result can be presented as "significant" through decisions, such as score transformations or covariate selection, that are not always disclosed (Simmons, Nelson, & Simonsohn, 2011). Such flexibility may encourage harking (hypothesizing after the results are known), or the presentation in a research report of a post hoc hypothesis informed by data analysis as if it were an a priori hypothesis. There is evidence that both p-hacking and harking are relatively common (John, Loewenstein, & Prelec, 2012; Kerr, 1998; Masicampo & Lalande, 2012), and these insidious practices are expected to generate false positive results where population effect sizes are routinely overestimated (Gelman, 2018; Ioannidis, 2005). Study preregistration, or the practice of publicly stating hypotheses and the analysis plan before data are analyzed, may counter both p-hacking and harking (Gehlbach & Robinson, 2018), but preregistration of analysis plans is yet relatively uncommon. Some research journals, such as *Psychological Science*, award open science badges to authors who preregister the design and analysis plan, share the data, or make study materials or stimuli publically available. Each badge has a specific graphical design by the Centre for Open Science (https://cos.io/our-services/open-science-badges/), and there is evidence that awarding badges actually works at least for data or materials sharing (Kidwell et al., 2016).

Of course, p values can be hacked in the other direction, too, that is, increased so that a significant finding is rendered as a nonsignificant result in a reanalysis. Of special concern is when a significant result that is unfavorable for some hypothesis or purpose is not reported or, even worse, intentionally made nonsignificant without disclosing the manipulation. In health research, such "reverse" p-hacking can have disastrous consequences, if adverse results are dropped or omitted (e.g., the Vioxx tragedy; Ziliak & McCloskey, 2008). Another example where results that are not statistically significant actually favor researchers' hypotheses is the model chi-square test in structural equation modelling (SEM). In this case, a significant chi-square statistic provides covariance evidence against the model because the null hypothesis of exact fit is rejected. But it is often possible to render a significant model chi-square statistic as nonsignificant by adding effects to the model (specifying free parameters), even if those additions make little theoretical sense. This is because increased complexity itself will typically improve model fit (Kline, 2016). Thus, the temptation for *model-hacking* in SEM: Through haphazard respecification, basically any model can be made to fit the data. Reporting standards for SEM call for researchers to explain their rationales for model respecification (Appelbaum et al., 2018), which may help to prevent model-hacking.

The ease with which results can be skewed in one way or another in significance testing can also facilitate intentional scientific fraud. This is why Ziliak (2016, p. 83) remarked that a "science, business, or law that is basing its validity on the level of p values, t statistics and other tests of statistical significance is looking less and less relevant and more and more unethical." Ziliak

[1] Widespread misunderstanding of confidence intervals is a related problem (e.g., Hoekstra, Morey, Rouder, & Wagenmakers, 2014).

also called on university research centers and government agencies to add the misuse of statistical significance to definitions of scientific misconduct. Kmetz (2019) described steps to counter the corrupting effects of *p*-hacking and other poor practices in data analysis.

Remediation and Prevention

Remediation and prevention are the two basic ways to ameliorate effects from the misuse of significance tests. Remedial efforts are for practicing researchers, usually postdegree, and include the aforementioned journal article reporting standards. Journal editorial policies, such as author guidelines, can serve a similar purpose. For example, many research journals in the health sciences now require the reporting of effect sizes. Quoted in the author guidelines for many of these journals is the recommendation by International Committee of Medical Journal Editors (2016, p. 15) to "avoid relying solely on statistical hypothesis testing, such as *p* values, which fail to convey important information about effect size and precision of estimates." Another example is the call by editors of health economics journals for increased attention to nonsignificant (negative) findings about interventions or programs that may nevertheless benefit real people ("Editorial statement on negative findings," 2015). The reporting and interpretation of effect sizes is also called for in reporting standards for quantitative research in psychology (Appelbaum et al., 2018).

Preventive efforts are usually educational, that is, for young scholars in training, such as students enrolled in graduate programs. There is a relatively large literature about how to prepare graduate students to teach introductory statistics to undergraduate students without strong quantitative backgrounds (e.g., Garfield & Everson, 2009). Unfortunately, there are fewer works about how to teach a first course in statistics at the graduate level (e.g., Moran, 2005; Steel, Liermann, & Guttorp, 2019), from this point on called GS1. A 3-credit, one-semester course that lasts 12–14 weeks is assumed.

This situation is unfortunate because GS1 classes play an important role in graduate training, especially when the terminal degree is the PhD or its counterpart in professional training programs (e.g., PsyD, EdD). Graduates of PhD programs should know something—a lot, actually—about research, so strong knowledge about statistics is paramount. Although research demands in graduate programs with a more applied focus may be lower compared with those in PhD programs, good statistics comprehension is also needed by professionals devoted to evidence-based practice (Shernoff, Bearman, & Kratochwill, 2017). Graduate training in statistics that begins with a strong GS1 course supports these goals.

There are some hurdles in teaching GS1 classes and dealing with these challenges must be addressed by instructors assigned to teach them. These obstacles are listed next and then discussed afterward with suggestions about how to ameliorate them:

1. Statistical significance testing is so overemphasized in undergraduate statistics (UGS) courses that many, and perhaps most, students fail to learn fundamental principles. They also may not realize that classical significance testing is not the sole inference framework for evaluating hypotheses.
2. Just as among established researchers, most students hold many false beliefs about outcomes in significance testing even after taking multiple UGS courses (Haller & Krauss, 2002). These myths can lead to overconfidence in analysis results and also impede new learning, if not corrected early.
3. Most new graduate students earned high grades in their prior UGS classes, but some may actually score at a failure level when tested on basic concepts at the beginning of GS1. To be fair, class size in UGS courses is often much larger than in GS1 classes and, thus, in-class examinations in UGS courses may be comprised mainly of multiple-choice items or calculational problems, which are easier to grade than essay questions. Essay questions are more demanding than multiple-choice questions, which require only recognition, and also harder than computational questions, which basically get at whether students can correctly apply a formula, not whether they understand the results. But comprehension deficits are more readily apparent in written responses. Anyhow, high grades in previous UGS courses can give some new graduate students the false impression that they have strong knowledge about statistics. A crisis of confidence can occur when such students learn the truth.

Great *p* Value Blank-Out

Summarized next are my impressions of student experiences in both UGS and GS1 courses. I do not claim that these experiences are universal among students, but I think that elements of these experiences can pose challenges when teaching GS1 courses, so my intention is to signal potential obstacles. With these limitations in mind, I believe that things in traditional introductory UGS courses usually go fairly well over the first month or so when basic descriptive statistics and types of distribution plots are introduced. These topics are generally straightforward with relatively few theoretical aspects.

Next, a crisis point in introductory UGS courses may be encountered around the middle of the semester with the introduction of the more abstract concepts about standard errors and sampling distributions. Many students struggle with these relatively abstract ideas; specifically, they have trouble distinguishing between the empirical (sample) distribution, the population distribution, and the sampling distribution (Green & Blankenship, 2014). Related distinctions that may be confused include the facts that (a) two of these distributions (population, sampling) are *theoretical* and may also be defined under particular assumptions (e.g., normality) while the third (empirical) is *observed*. Also, (b) a subset of these distributions (population, empirical) generally concerns *scores* at the case level while the third distribution (sampling) involves *statistics* (i.e., aggregated scores). The failure to correctly distinguish scores from statistics makes it difficult to comprehend the difference between standard deviations when used to describe observed variation at the score level in a particular sample and standard errors, which refer to expected variation among statistics from random samples (Altman & Bland, 2005).

Significance testing is often introduced just as students are still struggling with the basics of sampling theory. Many UGS students

at this time also start to internalize myths about *p* values mentioned earlier. Unless explicitly instructed otherwise, undergraduate students can easily come to overvalue significance testing as a kind of gold-standard method that somehow detects "real" effects above the static of sampling error (i.e., $p < \alpha$) or dismisses results due to sampling error ($p \geq \alpha$). For example, Kühberger, Fritz, Lermer, and Scherndl (2015) found that even undergraduate students with advanced statistical training associate statistical significance with medium to large effect sizes and assume that statistical significance is due to real effects. In contrast to the subtleties of sampling theory, this illusion of certainty afforded by significance testing may be welcomed by students, but there are costs, as elaborated next.[2]

Strong significance testing mentality can hinder continued learning about statistics (e.g., Huck, 2016; Rodgers, 2010; Ziliak & McCloskey, 2008), including at the graduate level. This phenomenon is the *great p value blank-out,* and it happens when students or researchers put little effort into learning what is for them a new statistical technique until they get to the part about significance testing (Kline, 2013). In computer output, they skip right to the *p* values but show little comprehension of or interest in the rest of the results. Ziliak and McCloskey (2008) used the term *trained incapacity*, and here it means that cognitive distortions associated with significance testing can make it difficult, if not impossible, for researchers to fully understand their own results and also the status of a research literature, if *p* values are their sole frame of scientific reference. Greenland (2017) described similar concerns about the capability of epidemiology researchers trained in traditional significance testing to correctly interpret and report study results.

Early Detection and Remediation of Poor Comprehension

Presented in Table 1 is a 16-item background knowledge test administered at the beginning of our GS1 course. Test-item content comes from UGS courses. New graduate students are asked to explain in written form the meaning(s) of the specific numerical value(s) for each statistic (Items 1–15). Readers are free to use or adapt the items listed in the table.

The average percentage of correct items on the test in Table 1 among our new graduate students is typically <50%, or what may be deemed a failure level of performance in some numerical grade systems or equivalent to a grade of "F" in some letter grade systems. That is, most of our new graduate students do not really understand even some quite basic statistical concepts despite achieving letter grades typically in the "A" range (i.e., "A-" to "A+") in prior UGS courses. That many new graduate students have poor mastery of statistical fundaments has been reported by others (Jannarone, 1986), so I do not believe that our students are any worse (or better) than those enrolled in psychology graduate programs at other universities. Indeed, most of our new graduate students at a Canadian university have taken *two* prior 3-credit UGS courses, but many new psychology graduate students in the United States have taken only a single 3-credit course.

It is worthwhile to consider some of the difficulties apparent in the answers of our new graduate students to the items in Table 1. Most students can define an arithmetic mean in terms of how it is calculated (i.e., as $\Sigma X/N$), but not as the balance point in a distribution from which the sum of absolute deviation scores is zero nor as the point from which the sum of squared deviations is the smallest (Item 1). The standard deviation is typically defined as the average absolute distance between the mean and the scores (Item 2), a common misconception (Huck, 2016).

Almost none of the students can correctly define a standard error, neither that of the mean nor that of the difference between two independent means (Table 1, Items 3–4). A frequent error involves describing the standard error as the standard deviation in the population distribution instead of as the standard deviation in a distribution of random statistics. Another frequent error is to define the standard error as the absolute distance between the sample statistic and the corresponding parameter. Confidence intervals are usually misinterpreted as ranges for which the probability is either .95 or .99 that the interval includes the correspond-

Table 1
Statistics Background Knowledge Test for New Graduate Students

For each of Items 1–15, interpret the specific numerical value(s); that is, indicate what each numerical value measures or estimates. For items with multiple numerical values, interpret each and every one of these values.

1. $M = 15.50$
2. $SD = 3.75$
3. $s_M = 7.50$
4. $s_{M1-M2} = 5.60$
5. 95% CI for μ, [102.00, 108.00]
6. 99% CI for $\mu_1 - \mu_2$, [−.50, 3.50]
7. $\alpha = .05$
8. $\beta = .40$
9. Independent-samples $t(48) = 2.25$, $p = .029$
10. Dependent-samples $t(29) = 1.90$, $p = .067$
11. Single-factor between-subjects $F(3, 36) = 40.00/12.50 = 3.20$
12. Factorial design, $MS_{AB} = 39.75$
13. $SS_T = 1{,}115.30$
14. $r = .65$
15. Unstandardized bivariate regression line $\hat{Y} = 5.60\ X + 10.20$
16. Statistical significance at the .05 level (i.e., $p < .05$) means . . . (check all that apply)
    _____a. The probability that the results are due to chance is < .05
    _____b. The probability that the null hypothesis (H_0) is true is < .05
    _____c. The probability that a Type I error was just made in rejecting H_0 is < .05
    _____d. The probability that the alternative hypothesis (H_1) is true is > .95
    _____e. The probability that the result will be replicated is > .95
    _____f. The probability of the data given the null hypothesis is < .05

Note. M = arithmetic mean; z = normal deviate; s_M = standard error of the mean; s_{M1-M2} = standard error of the difference between two independent means; CI = confidence interval; μ = population mean; α = criterion for statistical significance; β = probability of Type II error; MS_{AB} = mean square for the $A \times B$ interaction; SS_T = total sum of squares; r = Pearson correlation; $\hat{Y}$ predicted raw score on Y.

[2] It also does not help that significance testing is still presented in too many introductory statistics text books as a series of "black boxes"—without authorship (citations), context, or mention of controversies over time (Moran, 2005; Williams, 2005). This type of presentation gives the false impression that significance testing is a set of undisputed facts that is universally accepted in psychology as the only way to test hypotheses—see Gurnsey (2018) for an exception.

ing parameter (Items 5–6). In contrast, most students can correctly define both Type I and Type II errors in significance testing (Items 7–8).

Students have much difficulty with Items 9–10 in Table 1 about results of the *t* test for mean differences. Some can correctly describe the degrees of freedom (*df*) values as determined by sample or group sizes, but few manage to correctly define *t* as the magnitude of the mean contrast in standard error units nor can most students offer a correct definition for its *p* value. About all they can say here without error it that one result is "significant" (Item 9) while the other result is "not significant" (Item 10) at the .05 level, but with little real comprehension about either outcome. Although the basic *t* test is dealt with in most UGS courses, it may be true that many students do not really understand what its results actually mean.

For Item 11 in Table 1, some students can correctly define the *df* values for the *F* test and also remember that *F* equals the ratio of the two mean squares, but few can state the meaning of the numerical values of those mean squares. For example, they do not know that the denominator in Item 11, or 12.50, is the average within-cells variance that is derived at the level of *scores*, nor do they realize that cell size per se does *not* affect the value of the error term for the *F* test. They also may not realize that the numerator, or 40.00, is a variance at the level of cell *means* (i.e., their variation around the grand mean) or that the value of this variance *is* affected by cell size. Likewise, few students can correctly explain the numerical value for the mean square for a two-way interaction effect in factorial analysis of variance (ANOVA; Item 12), and few can correctly explain the meaning of the total sum of squares (Item 13). The results just described suggest that the typical new graduate student may fail to correctly define any of the individual entries in an ANOVA source table except possibly the *df* values.

Most of our new graduate students know that a Pearson correlation estimates the strength of the linear association between two continuous variables (Item 14 in Table 1), but they usually do not know how to further interpret its specific numerical value; that is, as a proportion of shared variance—$.65^2$, or .423—or that a difference in one variable of one standard deviation predicts a difference in the other variable of .65 standard deviations in a bivariate regression analysis. Most students correctly label the slope and intercept in the unstandardized regression equation (Item 15), but not all can state that a difference in *X* of one point in its raw score metric predicts a difference in *Y* of 5.60 points in its original metric. Items 16a–16e correspond to common myths about *p* values. Only Item 16f is actually true, but most students, about 90% or so, incorrectly endorse at least one item among 16a–16e as true.

Students' results on the test in Table 1 have no bearing on their grades in our GS1 course. Feedback about test performance includes review of correct definitions, which begins the process of corrective intervention. Students spend the next month in the tutorial (lab) section for our GS1 course reading Huck (2016) about statistical misconceptions and completing a series of related exercises, which are available in the online supplementary materials for this article with pedagogical comments for instructors. The activities just mentioned are led by a graduate student teaching assistant. In the lecture section, the topics of sampling theory, interval estimation, and correct versus incorrect interpretations of outcomes in significance testing are also reviewed.

At the end of a month of review, the students are administered a mastery test about undergraduate-level statistical concepts. They must pass the mastery test in order to continue in the GS1 class; otherwise, they are required to rewrite an alternate version of the mastery test. If a student should fail a second administration of the mastery exam, then that student would be asked to withdraw from the class. In my experience, though, this outcome would be very rare; instead, most graduate students go on to successfully finish the course. Indeed, the amount of progress made by the typical graduate student over the course of a single semester is impressive, as it should be. After all, newly admitted graduate students were among the best and brightest of their undergraduate cohorts, so it is realistic to hold them to a higher standard (i.e., they must understand what they are doing when they analyze their data).

Crisis of Confidence

This discussion reflects my impression of student experiences in GS1 classes but not anything I would claim is always true. Once again the goal is to highlight possible challenges for instructors of GS1 courses. The graphic in Figure 1 represents what I call the *Dunning-Kruger-Amara effect* for the expected relation between student confidence and competence in statistics through different educational phases. The *Dunning-Kruger effect* is a cognitive bias where people with relatively poor skills tend to overestimate their competence (Kruger & Dunning, 1999). The same effect may lead poor performers to push back against attempts to intervene, that is, to both reject the suggestion of inadequacy and offers for remediation (Dunning, 2011). *Amara's law* refers to a saying attributed to the scientist and futurist Ray Amara, who suggested that the impact of a new technology is overestimated in the short term but underestimated in the long run (Isenberg, 2001).

Students in UGS courses may experience a sense of increasing competence until the introduction of sampling theory and standard errors (Figure 1, Point 1), as mentioned. But any stall in rising confidence may soon be alleviated by the introduction of tradi-

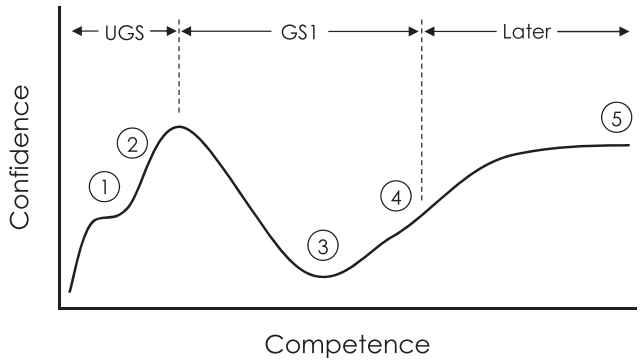

Figure 1. Dunning-Kruger-Amara effect for the expected relation between student confidence and competence through undergraduate statistics (UGS), the first graduate statistics course (GS1), and later. Phases: 1, struggle with concept of standard error; 2, significance testing introduced and rise of false confidence; 3, midterm in graduate statistics and confidence crisis; 4, beginning of realism; 5, continued learning and eventual attainment of wisdom.

tional significance testing (Point 2), during which students might conclude that p values are all that researchers really need to know. Next, confidence increases to its peak at the end of the UGS phase, when both high grades among future graduate students and myths about the meaning of statistical significance combine to boost confidence to an unrealistically high level, given still relatively low real levels of competence in statistics.

A second crisis of confidence may occur at the beginning of a GS1 course, if new graduate students find out that their actual knowledge of backgrounds concepts is actually deficient. A confidence low point might occur about midway through a GS1 course (Figure 1, Point 3), if the midterm exam grade is not high. A mature student response to such an outcome is to seek advice about how to correct misconceptions and improve subsequent performance. Although rare, a less civil, more antagonistic response can occur, one that involves blaming the instructor for unfair evaluation while simultaneously failing to take responsibility for shortcomings in the student's work. Most university instructors know how to constructively deal with student incivility about grading; see Holdcroft (2014) or Jannarone (1986) for examples.

My experience is that despite a confidence crisis in the first half of a GS1 course, most students recover as they gain mastery over both background concepts and more advanced topics. Indeed, the amount of progress that bright graduate students make between the middle and conclusion of a GS1 course can be extraordinary, and their sense of competence may begin to increase once more (Figure 1, Point 4). After a taking a GS1 course, both student competence and confidence should continue to gradually increase (Point 5) as they gain further experience.

Modernize Course Content

Sometimes GS1 courses offer little more than additional ways to test effects for statistical significance. An example is what I call *ANOVA voodoo*, or ritualistic ways to apply significance testing in complex designs, such as conducting the F test for a main effect with $\geq$2 degrees of freedom and then, if significant, testing all pairwise contrasts with a post hoc method, such as the Student-Newman–Keuls method (Wilkinson, 1999, nicely explained the illogic of this particular ritual). Such courses are so old-fashioned that little is accomplished except for reinforcing the great p value blank-out.

New Statistics

A contemporary GS1 course should cover what Cumming and Calin-Jageman (2017) called the *new statistics*, or effect sizes, confidence intervals, and meta-analysis, or how to synthesize results over replication studies. As mentioned, effect size estimation is required by many journals for health-related research. Most meta-analyses deal with the estimation of average effect sizes over sets of related studies. Another focus in many meta-analyses is whether effect size varies as a function of study characteristics, such as the composition of samples or the use of particular methods or procedures. To understand meta-analytic articles, graduate students need a good understanding of effect sizes including how to estimate the precision of results in primary studies (i.e., confidence intervals for effect sizes).

Understanding effect size also means *interpreting* such results, that is, addressing the question, what does it *mean* that an effect of a particular magnitude has been found? For example, is the observed effect size smaller or larger compared with those reported in related studies, whether those prior results were summarized in a previous meta-analysis (if available) or calculated by the researcher from summary statistics reported in primary studies (if the original authors did not report effect sizes but sufficient summary statistics were available to calculate effect sizes)? Another example concerns studies with interventions, such as clinical trials: Is the effect of the intervention large enough to be meaningful, such as exceeding a minimal clinically important difference, or the smallest change in status that patients consider as important (e.g., Angst, Aeschlimann, & Angst, 2017; Norman, Sloan, & Wyrwich, 2003)?

In too many journal articles, effect sizes are reported, but then not interpreted. For example, the hypothetical result for a mean contrast

$$t\,(50) = 2.25,\ p = .029,\ d = .30$$

is reported in the Results section with particular emphasis on statistical significance, but nothing is said about the standardized mean difference (d) effect size that accompanies the t test outcome. Well, just what does an average difference that corresponds to 30% of a standard deviation imply? Reporting but not interpreting effect sizes while elaborating on whether results are significant or not significant is "business as usual" and poor practice, too.

Standards for Replication

The role of effect size in replication studies is another important topic, especially if replication is eventually both valued and expected in psychology research. Students should know that statistical significance, or the requirement for significant results (e.g., $p < .05$) in both original and replication analyses, is a flawed standard for many reasons, including the distorting effect of low power and expected overestimation of population effect sizes (Anderson, Kelley, & Maxwell, 2017; Maxwell, Lau, & Howard, 2015). Alternatives based on effect size are more promising, such as whether the effect size from an original study falls within the 95% confidence interval for the same effect size in a replication (e.g., Open Science Collaboration, 2015). Another example is the technique of continuously cumulating meta-analysis where effects sizes from replications are synthesized along with all prior results and then updated as new findings subsequently appear (Braver, Thoemmes, & Rosenthal, 2014).

Bayesian methods are another alternative that can synthesize effect sizes from studies as they are reported and update the status of a research area. Bayesian statistics generally require strong knowledge of probability distributions, but relatively few psychology students have such prerequisite knowledge, so Bayesian methods are not a silver bullet for psychology GS1 courses. Gannon, Pereira, and Polpo (2019) described ways to blend Bayesian concepts with more traditional statistical methods.

Reforms of Significance Testing

Wasserstein et al. (2019) reminded us that there is no method that will magically replace significance testing in all research contexts and that there may never be such an alternative. Thus, it is still necessary to teach about significance testing in GS1 courses,

but students should be made aware of various ways to reform its application while at the same time emphasizing correct interpretations of results (i.e., warn them away from myth). Some possibilities about reformed significance testing are briefly listed next, but see the 2019 special issue "Statistical Inference in the 21st Century: A World Beyond $p < 0.05$" in *The American Statistician* for more information:

1. *Smart* α. Values of *dumb* α are either .05 or .01 that are blindly specified by the researcher with no rationale about whether either conventional level actually makes sense in a particular study. If at the least the *relative costs* of making a Type I versus Type II error can be estimated, then it is possible to specify a smart level of α that properly balances the risks of these decision errors (Aguinis et al., 2010; Mudge, Baker, Edge, & Houlahan, 2012). If the optimal level is, say, $\alpha = .30$, then it replaces a value of dumb α (.05 or .01) in the analysis.

2. *No* α. Hurlbert and Lombardi (2009) reminded us there is actually no requirement to specify any value for α at all. In this variation, *p* values are not compared against any criterion value for α; thus, there is no dichotomization of *p* values against an arbitrary standard such as .05, *nor against any standard whatsoever*. Also, the clichéd term *significant* is never used, which is consistent with recent calls to drop this expression from the language of statistical analysis along with the equally insipid "$p < .05$" (Wasserstein et al., 2019).

3. *Combine effect sizes and p values.* There are various methods to explicitly incorporate hypotheses about effect size, but just two are mentioned next. Generation of *second generation* p *values* involves the specification of an interval null hypothesis that includes not just a null (zero) effect size, but also the limits of trivial effect sizes, or nonzero effects considered as too small to be of interest (Blume, Greevy, Welty, Smith, & Dupont, 2019). In the *minimum effect size plus* p *value* approach, rejection of the null hypothesis also requires that observed effect sizes are meaningfully large and practically significant (Goodman, Spruill, & Komaroff, 2019).

No Significance Testing

The variations just described reform various aspects of traditional significance testing, or what has been called the Intro Stats method, but students in modern GS1 classes should not be given the impression that generating *p* values is the ultimate analysis goal. This is because the design–analysis gap between what significance testing requires versus what happens in actual studies remains vast even after reforming the Intro Stats method. For example, random sampling from populations with particular distributional characteristics and no other sources of error other than sampling error are assumed whenever *p* values are generated. This is true even for methods of significance testing, such as those based on nonparametric bootstrapping, that supposedly require no distributional assumptions except that the shape of the empirical distribution matches that of the population, which is itself a distributional assumption albeit a less stringent one. Study of true probability samples is rare in psychology research about humans, strong distributional assumptions such as normality or homoscedasticity are flawed in many actual samples, and even fairly modest levels of measurement error can severely distort *p* values (e.g., Gelman, 2018; Keselman et al., 1998; Westfall & Yarkoni, 2016), among other ways that significance testing can go off the rails (cognitive distortions, testing null hypotheses known to be false before any data are collected, etc.).

One way to prepare GS1 students for the future is to teach them how to describe their analysis results *with no use of significance testing at all*, that is, without reference to *p* values. Doing so may seem odd to some students at first, but they should know that significance tests are not generally necessary to detect substantive effects, which, as Cohen (1994) noted, should be obvious to visual inspection of relatively simple kinds of graphical displays. Show students instead how to refer to effect sizes with confidence intervals with emphasis on whether those effects sizes are *large enough* and *precise enough* to be meaningful in a particular research context (Calin-Jageman & Cumming, 2019). The description of results at the level closer to the data may also help students to develop more effective communication skills (Kline, 2020).

It might also help to compare for students the two situations described next: Researcher 1 has data from a single sample, that is, there no available data for replication. Instead, the researcher reports *p* values from various analyses, some of which correspond to results designated as "significant" results whereas other *p* values are associated with "nonsignificant" findings. The proper, scientific response should be, "So what?" or, more to the point, "Who cares?" Those significant results may or may replicate in other samples; likewise, their poor, embarrassing, unwashed cousins, the nonsignificant results, may or may not replicate. But expounding on the implications of the significant results while ignoring the nonsignificant results with no evidence for replication *is* the crisis to which this special issue is devoted. As Ziliak (2016) put it, "the deep, dark night of mindless significance testing has to go" (p. 89). In this case, replicated results warrant attention, not asterisks (a typographical symbolism for significant results), as elaborated next.

Now suppose that Researcher 2 conducted no significance testing at all but reported meaningful effects sizes that fall within the confidence intervals from two or more independent samples that are also acceptably precise in each sample (Calin-Jageman & Cumming, 2019). Things get even better if those other samples were collected by investigators other than Researcher 2 because the potential obstacles to replication were greater in this case. But the main point is that in this second scenario, there is now something of scientific interest. This second case corresponds in part to what Spellman (2015) referred to as "Revolution 2.0" and what Vazire (2018) called the "credibility revolution," where *revolution* means how we should be doing science: open, transparent, and *repeatable*, that is, a change in researcher mentality and habits, if not a change in the actual content of psychology research. In this way, replication should trump significance testing in single-sample analyses just about any time.

Computer Tool Use

Use of computer tools is part of many statistics courses, and undergraduate students who conducted thesis research projects and analysed raw data in SPSS or other statistical software have some practical computer skills, too. There are two additional kinds of computer exercises that I think that can be especially useful for GS1 students. One is to give students a raw data file with many problems, including missing data, duplicated cases, distributions of quantitative variables that are severely non-normal, variables that are extremely collinear, entries that are miscoded, and so on. The task is for students to diagnose the problems and recommend corrections (i.e., screen and clean the data). Data screening is a critical skill even for students who will analyze archival data sets, which are often riddled with problems even when such data sets were analyzed for published studies. The current state of reporting about data screening is abysmal (e.g., Osborne, 2013), but contemporary reporting standards call for more explicit detail in this area (e.g., Appelbaum et al., 2018).

Another type of computer exercise useful to students involves the analysis of summary statistics instead of raw data. In my experience, students are generally unaware that many kinds of statistical analyses do not require raw data files, if the appropriate summary statistics—such as group means, standard deviations, and sizes (n) for between-subjects effects and also correlations for within-subjects effects—are available. The capability to analyze summary statistics can also enable students to verify results from published studies or test hypotheses not evaluated by the original authors, even though students may have no access to the original raw data file. Cheng and Phillips (2014) discussed challenges and opportunities of secondary data analyses.

The online supplementary materials for this article include examples of computer exercises for data screening and analyses with summary data. The data screening exercise includes a synthesized raw data file with $N = 500$ cases—also available for download in the online supplementary materials—that has many issues about missing data, coding mistakes, outliers, and distributional characteristics. Readers are free to adapt these exercises for use in their own statistics courses.

Summary of Recommendations

Spence, Stanley, and Newby-Clark (2018) noted that part of educating graduate students in research-based programs is ensuring that they are aware of change, in this case teaching and encouraging them to adopt better analysis and research practices that could help to guide psychology research out of its credibility crisis. Core recommendations for instructors of GS1 courses aimed at facilitating these goals are summarized next:

1. Most new graduate students in GS1 courses earned high grades in prior UGS courses, but not all those students have a good grasp of basic statistical concepts (e.g., standard deviations vs. standard errors). Assess students' background knowledge and provide them with feedback about concepts they need to review.

2. Expect that students hold multiple false beliefs about p values (just like the rest of us). Help students to identify, challenge, and remediate cognitive distortions that can lead to false confidence about analysis results and impede new learning.

3. Acquaint students with principles of open science that foster transparency and accountability, such as preregistration of analysis plans and principles of open data and research materials access, that may help to prevent the worst abuses of traditional significance testing, including p-hacking. Also direct students to relevant reporting standards for their respective research areas.

4. Tell students about legitimate variations on traditional significance testing that include the elimination of arbitrary criterion levels of statistical significance, such as $p < .05$, especially if relative costs of Type I versus Type II error can be estimated. Otherwise, encourage students to stop dichotomizing continuous p values and also stop using the hackneyed, meaningless term "significant," which is both a symptom and cause of cognitive distortions among researchers.

5. Teach students how to both estimate and interpret effect size and also quantify uncertainty about those estimates. That is, emphasize interval estimation with effect sizes, but also help students to avoid widespread cognitive errors about confidence intervals, which generally parallel interpretive errors in significance testing. Also inform students about how to evaluate replicated results from the perspective of interval estimation and meta-analysis.

These ideas about post p value education in graduate statistics are represented next in the variation on the first verse of the song *Imagine* by Lennon and Ono (1971) listed next:

> Imagine there's no significant
>
> It's easy if you try
>
> No nonsignificant below us
>
> Around us only effects of different precision
>
> Imagine all the psychology researchers
>
> Living for replication

Résumé

Le premier cours de statistique joue un rôle important dans bon nombre de programmes de deuxième cycle, mais les enseignants font face à certains défis. D'un côté, le programme standard de bon nombre de cours de statistique du premier cycle accorde trop d'importance au test d'hypothèse traditionnel, ce qui remplit le cerveau des étudiants de mythes qui vont à l'encontre du perfectionnement de la littératie statistique. De l'autre, la qualité des cours de premier cycle varie grandement, si bien que les nouveaux étudiants de deuxième cycle peuvent présenter des niveaux fort variables de connaissances en statistique au début d'un cours de deuxième cycle. Le présent article vise à fournir des conseils constructifs aux enseignants des premiers cours de statistique de deuxième cycle, notamment des suggestions sur la façon d'évaluer les connaissances statistiques et prendre des mesures correctives

précoces pour pallier les lacunes dans les connaissances de leurs étudiants. Les sujets essentiels qu'il conviendrait de couvrir dans un cours de statistique réformiste post valeur p sont aussi recommandés. Le but est d'aider les enseignants à promouvoir une compréhension plus moderne et véritablement digne du deuxième cycle des statistiques parmi les étudiants.

Mots-clés : statistique de deuxième cycle, stratégies d'enseignement, réforme des statistiques, nouvelles statistiques, statistiques post valeur *p*.

References

Aguinis, H., Edwards, J. R., & Bradley, K. J. (2017). Improving our understanding of moderation and mediation in strategic management research. *Organizational Research Methods, 20,* 665–685. http://dx.doi.org/10.1177/1094428115627498

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods, 13,* 515–539. http://dx.doi.org/10.1177/1094428109333339

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63,* 32–50. http://dx.doi.org/10.1037/0003-066X.63.1.32

Aiken, L. S., West, S. G., Sechrest, L., Reno, R. R., Roediger, H. L., Scarr, S., . . . Sherman, S. J. (1990). Measurement in psychology: A survey of PhD programs in North America. *American Psychologist, 45,* 721–734. http://dx.doi.org/10.1037/0003-066X.45.6.721

Altman, D. G., & Bland, J. M. (2005). Standard deviations and standard errors. *British Medical Journal, 331,* 903. http://dx.doi.org/10.1136/bmj.331.7521.903

Amrhein, A., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature, 567,* 305–307. http://dx.doi.org/10.1038/d41586-019-00857-9

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28,* 1547–1562. http://dx.doi.org/10.1177/0956797617723724

Angst, F., Aeschlimann, A., & Angst, J. (2017). The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. *Journal of Clinical Epidemiology, 82,* 128–136. http://dx.doi.org/10.1016/j.jclinepi.2016.11.016

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force report. *American Psychologist, 73,* 3–25. http://dx.doi.org/10.1037/amp0000191

Blume, J. D., Greevy, R. A., Welty, V. F., Smith, J. R., & Dupont, W. D. (2019). An introduction to second-generation *p*-values. *American Statistician, 73*(Suppl. 1), 157–167. http://dx.doi.org/10.1080/00031305.2018.1537893

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9,* 333–342. http://dx.doi.org/10.1177/1745691614529796

Calin-Jageman, R. J., & Cumming, G. (2019). The new statistics for better science: Ask how much, how uncertain, and what else is known. *American Statistician, 73*(Suppl. 1), 271–280. http://dx.doi.org/10.1080/00031305.2018.1518266

Cheng, H. G., & Phillips, M. R. (2014). Secondary analysis of existing data: Opportunities and implementation. *Shanghai Archives of Psychiatry, 26,* 371–375. Retrieved from http://www.shanghaiarchivesofpsychiatry.org/

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. http://dx.doi.org/10.1037/0003-066X.49.12.997

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

Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 247–296). Amsterdam, the Netherlands: Elsevier.

Editorial statement on negative findings. (2015). *Health Economics, 24,* 505. http://dx.doi.org/10.1002/hec.3172

EQUATOR Network. (n.d.). *Enhancing the QUAlity and Transparency Of health Research*. Retrieved from http://www.equator-network.org/

Gannon, M. A. Pereira, C. A. d. B., & Polpo, A. (2019). Blending Bayesian and classical tools to define optimal sample-size dependent significance levels. *American Statistician, 73*(Suppl. 1), 213–222. http://dx.doi.org/10.1080/00031305.2018.1518268

Garfield, J., & Everson, M. (2009). Preparing teachers of statistics: A graduate course for future teachers. *Journal of Statistics Education*. Advance online publication. http://dx.doi.org/10.1080/10691898.2009.11889516

Gehlbach, H., & Robinson, C. D. (2018). Mitigating illusory results through preregistration in education. *Journal of Research on Educational Effectiveness, 11,* 296–315.

Gelman, A. (2018). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Personality and Social Psychology Bulletin, 44,* 16–23. http://dx.doi.org/10.1177/0146167217729162

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016, March 4). Comment on "Estimating the reproducibility of psychological science." *Science, 351,* 1037. http://dx.doi.org/10.1126/science.aad7243

Goodman, W. M., Spruill, S. E., & Komaroff, E. (2019). A proposed hybrid effect size plus p-value criterion: Empirical evidence supporting its use. *American Statistician, 73*(Suppl. 1), 168–185. http://dx.doi.org/10.1080/00031305.2018.1564697

Grant, S. P., Mayo-Wilson, E., Melendez-Torres, G. J., & Montgomery, P. (2013). Reporting quality of social and psychological intervention trials: A systematic review of reporting guidelines and trial publications. *PLoS ONE, 8*(5), e65442. http://dx.doi.org/10.1371/journal.pone.0065442

Green, J. L., & Blankenship, E. E. (2014, July). *Beyond calculations: Fostering conceptual understanding in statistics graduate teaching assistants*. Paper presented at the International Association for Statistical Education 9th International Conference on Teaching Statistics (ICOTS9), Flagstaff, AZ. Retrieved from https://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_3A3_GREEN.pdf

Greenland, S. (2017). Invited commentary: The need for cognitive science in methodology. *American Journal of Epidemiology, 186,* 639–645. http://dx.doi.org/10.1093/aje/kwx259

Gurnsey, R. (2018). *Statistics for research in psychology: A modern approach using estimation*. Thousand Oaks, CA: Sage.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7,* 1–17. Retrieved from https://www.dgps.de/fachgruppen/methoden/mpr-online/

Hoekstra, R., Kiers, H. A. L., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology, 3,* 137. http://dx.doi.org/10.3389/fpsyg.2012.00137

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21,* 1157–1164. http://dx.doi.org/10.3758/s13423-013-0572-3

Holdcroft, B. (2014). Student incivility, intimidation, and entitlement in academia. *Academe*. Retrieved from https://www.aaup.org/academe

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement, 60,* 661–681. http://dx.doi.org/10.1177/00131640021970808

Huck, S. W. (2016). *Statistical misconceptions*. New York, NY: Routledge.

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theory framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46,* 311–349. http://dx.doi.org/10.5735/086.046.0501

International Committee of Medical Journal Editors. (2016). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals*. Retrieved from http://www.icmje.org/recommendations/

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2,* e124. http://dx.doi.org/10.1371/journal.pmed.0020124

Isenberg, D. S. (2001, November 26). Amara's law [Web Blog post]. Retrieved from http://isen.com/archives/011126.html

Jannarone, R. J. (1986). Preparing incoming graduate students for statistics. *Teaching of Psychology, 13,* 156–157. http://dx.doi.org/10.1207/s15328023top1303_19

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524–532. http://dx.doi.org/10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217. http://dx.doi.org/10.1207/s15327957pspr0203_4

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., . . . Levin, J. R. (1998). Statistical practices of education researchers: An analysis of the ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350–368. http://dx.doi.org/10.3102/00346543068003350

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., . . . Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology, 14*(5), e1002456. http://dx.doi.org/10.1371/journal.pbio.1002456

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/14136-000

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Kline, R. B. (2020). *Becoming a behavioral science researcher* (2nd ed.). New York, NY: Guilford Press.

Kmetz, J. L. (2019). Correcting corrupt research: Recommendations for the profession to stop misuse of *p*-values. *American Statistician, 73*(Suppl. 1), 36–45. http://dx.doi.org/10.1080/00031305.2018.1518271

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121–1134. http://dx.doi.org/10.1037/0022-3514.77.6.1121

Kühberger, A., Fritz, A., Lermer, E., & Scherndl, T. (2015). The significance fallacy in inferential statistics. *BMC Research Notes, 8,* 84. http://dx.doi.org/10.1186/s13104-015-1020-4

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory & Psychology, 22,* 67–90. http://dx.doi.org/10.1177/0959354311429854

Lennon, J., & Ono, Y. (1971). *Imagine* [Recorded by John Lennon]. On Imagine [record]. London, UK: Apple Records.

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board Task Force report. *American Psychologist, 73,* 26–46. http://dx.doi.org/10.1037/amp0000151

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43,* 304–316. http://dx.doi.org/10.3102/0013189X14545513

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *The Quarterly Journal of Experimental Psychology, 65,* 2271–2279. http://dx.doi.org/10.1080/17470218.2012.711335

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70,* 487–498. http://dx.doi.org/10.1037/a0039400

McShane, B. B., & Gal, D. (2016). Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. *Management Science, 62,* 1707–1718. http://dx.doi.org/10.1287/mnsc.2015.2212

Meier, S. T. (1994). *The chronic crisis in psychological measurement and assessment: A historical survey*. New York, NY: Academic Press.

Moran, T. P. (2005). The sociology of teaching graduate statistics. *Teaching Sociology, 33,* 263–284. http://dx.doi.org/10.1177/0092055X0503300303

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS ONE, 7*(2), e32734. http://dx.doi.org/10.1371/journal.pone.0032734

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care, 41,* 582–592. http://dx.doi.org/10.1097/01.MLR.0000062554.74615.4C

Oakes, M. (1986). *Statistical inference*. New York, NY: Wiley.

Open Science Collaboration. (2015, August 28). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. http://dx.doi.org/10.1126/science.aac4716

Osborne, J. W. (2013). *Best practices in data screening*. Thousand Oaks, CA: Sage.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65,* 1–12. http://dx.doi.org/10.1037/a0018326

Shernoff, E. S., Bearman, S. K., & Kratochwill, T. R. (2017). Training the next generation of school psychologists to deliver evidence-based mental health practices: Current challenges and future directions. *School Psychology Review, 46,* 219–232. http://dx.doi.org/10.17105/SPR-2015-0118.V46.2

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Spellman, B. A. (2015). A short (personal) future history of Revolution 2.0. *Perspectives on Psychological Science, 10,* 886–899. http://dx.doi.org/10.1177/1745691615609918

Spence, J. R., Stanley, D., & Newby-Clark, I. (2018). *Why students are the answer to psychology's replication crisis*. Retrieved from https://theconversation.com/why-students-are-the-answer-to-psychologys-replication-crisis-90286

Steel, E. A., Liermann, M., & Guttorp, P. (2019). Beyond calculations: A course in statistical thinking. *American Statistician, 73*(Suppl. 1), 392–401. http://dx.doi.org/10.1080/00031305.2018.1505657

Szucs, D., & Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. *Frontiers in Human Neuroscience, 11,* 390. http://dx.doi.org/10.3389/fnhum.2017.00390

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science, 13,* 411–417. http://dx.doi.org/10.1177/1745691617751884

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$." *American Statistician, 73*(S1), 1–19. http://dx.doi.org/10.1080/00031305.2019.1583913

Waters, J. (2017). Benedict the dissident. *First Things*. Retrieved from https://www.firstthings.com/article/2017/02/benedict-the-dissident

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE, 11*(3), e0152719. http://dx.doi.org/10.1371/journal.pone.0152719

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604. http://dx.doi.org/10.1037/0003-066X.54.8.594

Williams, R. (2005). The sociology of teaching graduate statistics: Another perspective. A response to "The Sociology of Teaching Graduate Statistics." *Teaching Sociology, 33,* 277–279. http://dx.doi.org/10.1177/0092055X0503300306

Ziliak, S. T. (2016). Statistical significance and scientific misconduct: Improving the style of the published research paper. *Review of Social Economy, 74,* 83–97. http://dx.doi.org/10.1080/00346764.2016.1150730

Ziliak, S., & McCloskey, D. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor: University of Michigan Press.

Received May 27, 2019
Revision received October 4, 2019
Accepted October 8, 2019 ■