SOCIETY FOR THE TEACHING OF PSYCHOLOGY

Teaching of Psychology
2020, Vol. 47(2) 169-179

DOI: 10.1177/0098628320901372
journals.sagepub.com/home/top
SAGE

What Psychology Teachers Should Know About Open Science and the New Statistics

Beth Morling[1] and Robert J. Calin-Jageman[2]

Abstract

Psychology teachers have likely heard about the "replication crisis" and the "open science movement" in psychology, and they are probably aware that psychologists have proposed new standards for research practice. How should our psychology courses reflect these new standards? We describe several modern practices that have transformed our field and that seem likely to endure: preregistration of studies, transparency of reporting, norms for replication, and the new statistical focus on estimation and precision. We offer suggestions for how to integrate these new practices into psychology courses.

Keywords

college teaching, open science, research transparency, estimation thinking, new statistics

> As someone who has been doing research for nearly 20 years, I now can't help but wonder if the topics I chose to study are in fact real and robust. Have I been chasing puffs of smoke for all these years? . . . I'm in a dark place. I feel like the ground is moving from underneath me and I no longer know what is real and what is not.
>
> Inzlicht (2016)

Many of us can relate to Dr. Inzlicht's despair. Psychology's "replication crisis" has been at the forefront of science journalism, psychology conferences, and social media for several years now (Open Science Collaboration, 2012, 2015). Some of psychology's most well-known studies have been marked by a failure to replicate and have left us wondering which studies are "real and robust" and which are "puffs of smoke." For example, the facial feedback hypothesis (Strack et al., 1988) was not replicable in a large, multilab sample (Wagenmakers et al., 2016; see Strack, 2016). Similarly, although many studies support ego depletion theory (Baumeister, 2014), at least one large replication attempt failed (Hagger et al., 2016; but see Baumeister, 2019). The effects of "power posing" (Carney et al., 2010) have also been elusive except when using self-report measures (Ranehill et al., 2015; Simmons & Simonsohn, 2017; see also Cuddy et al., 2018). Even seemingly uncontroversial examples—such as the hypothesis that people eat more when using larger plates—have not survived replication attempts (Kosīte et al., 2019).

Importantly, a disappointing replication result does not mean the original finding is wrong. However, a failed replication suggests that these phenomena are more complex or nuanced than current textbooks might reflect.

Psychology's Methodological Upheaval: Problems and Solutions

Stories of failed replications have coevolved with a multidimensional "credibility revolution" (Vazire, 2018). As psychologists inquired why studies failed to replicate, they identified questionable research practices potentially behind the unstable findings (Chambers, 2017). Formerly common research practices such as HARKing (hypothesizing after the results are known), *p*-hacking, and small samples, are now questioned. In their place, new practices are becoming standard.

Publication practices are changing, too. For example, journals are devoting more space to replication studies (e.g., Lindsay, 2015, 2017; Makel et al., 2012). Similarly, because of the limitations of null hypothesis significance testing, many journals are now requiring confidence intervals (CIs) and effect sizes instead of *p* values (Cumming, 2014). Although not all psychologists agree about the existence or severity of the problem (e.g., Gilbert et al., 2016), one fact seems clear: Psychological science is experiencing a period of rapid methodological change.

Scientific and quantitative reasoning are core learning outcomes of many psychology courses (see, e.g., the American

[1] Department of Psychological and Brain Sciences, University of Delaware, Newark, DE, USA
[2] Department of Psychology, Dominican University, River Forest, IL, USA

Corresponding Authors:
Beth Morling, Department of Psychological and Brain Sciences, University of Delaware, Newark, DE 19716, USA; Robert J. Calin-Jageman, Department of Psychology, Dominican University, River Forest, IL 60305, USA.
Emails: morling@udel.edu; rcalinjageman@dom.edu

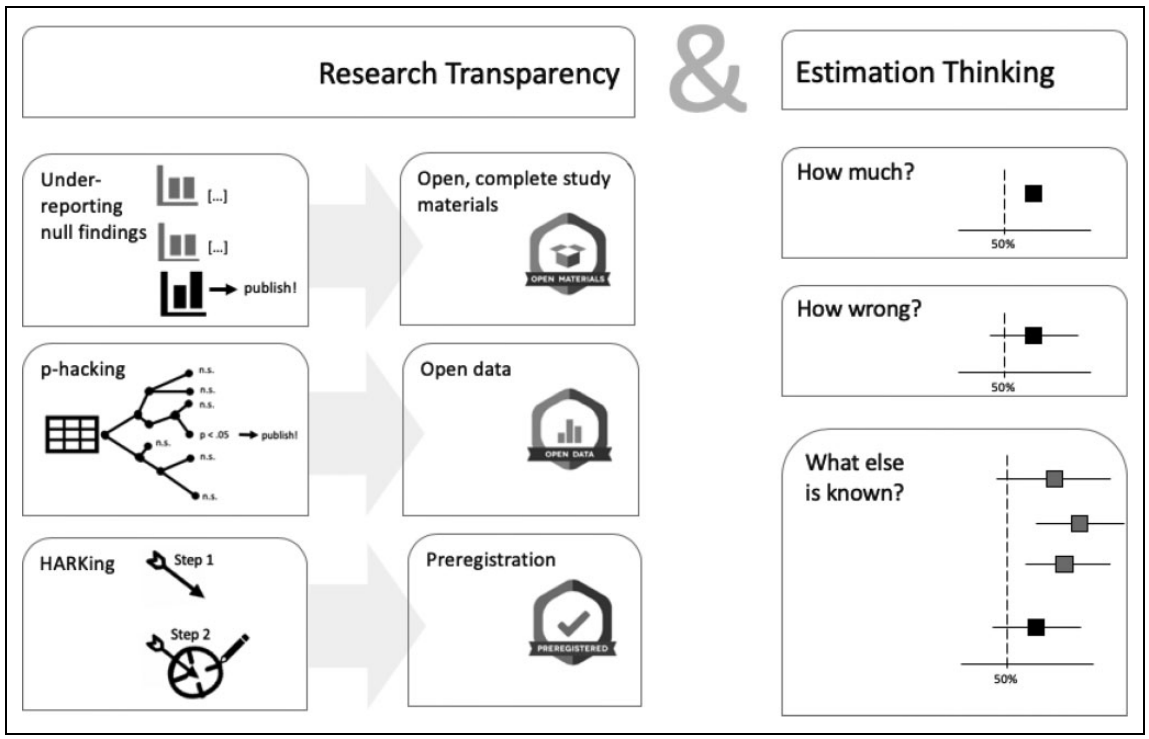

Figure 1. Concepts discussed in this article. Badge images in the center column were developed by the Association for Psychological Science (https://osf.io/tvyxz/wiki/home/).

Psychological Association [APA] Introductory Psychology Initiative, n.d.). Therefore, psychology teachers should help students practice scientific reasoning skills. Here we discuss two important themes: research transparency and estimation thinking. Figure 1 provides a pictorial introduction to the concepts in each theme.

Blurring the Line Between Confirmatory and Exploratory Research

The scientific method is meant to protect us from our own biases, to see the world as it truly is, not just as we would like to see it (to paraphrase Bacon, 1620/1889; see also Feynman, 1974). Science, though, is never complete—researchers are constantly discovering new ways biases can creep in and developing new protections against these biases (Nuzzo, 2015). The current period of rapid methodological change reflects how psychologists have identified new sources of error and quickly adopted potential remedies.

These remedies reassert that openness and transparency form the core of scientific practice and help redraw the bright line between confirmatory and exploratory research. Confirmatory research, depicted in psychology textbooks as the "theory-data cycle," or the hypothetico-deductive model (Chambers, 2017), typically proceeds as follows: Psychologists construct clear tests of their hypotheses, state these hypotheses in advance, collect and analyze their data objectively, and make results public even when the results fail to support the theory. Unfortunately, psychologists have too often unintentionally adopted questionable practices at each stage of this cycle (Chambers, 2017), resulting in a blurred line between confirmatory and exploratory research. These practices include underreporting nonsignificant effects, *p*-hacking, and HARKing. Before addressing how teachers might integrate these issues in their courses, we explain why each is problematic.

Blurring the Line by Underreporting

One questionable practice involves selective reporting of significant effects. For example, a researcher may include multiple dependent variables (DVs) in a study. When making results public, the researcher might report only the DVs that supported the hypothesis. There is nothing wrong with administering multiple dependent measures, but this practice becomes misleading when the researcher never reports outcomes that did not support predictions. The researcher has misrepresented an *exploratory* study as *confirmatory*.

An alleged example of this practice comes from Rohrer et al. (2015), who attempted to replicate a study on money and cognition. The original researchers found that exposure to money caused people to endorse attitudes that justify inequality (Caruso et al., 2013). While corresponding with Rohrer and colleagues, Caruso et al. revealed that they had reported only 9 of the 28 DVs they had measured. Rohrer et al. argued that by reporting only some of the DVs (only the ones that "worked"), the original authors misled readers about the strength of their

evidence. By underreporting, Caruso et al. presented their work as confirmatory when they had actually been exploring the effects of money on a wide variety of possible outcomes. (In response to Rohrer et al.'s failure to replicate, Vohs [2015] reviewed 10 years of studies on money primes and maintained their importance.)

Blurring the Line by p-hacking

A second practice that blurs the line between confirmatory and exploratory science is "*p*-hacking," or "exploiting researcher degrees of freedom" (Chambers, 2017; Simmons et al., 2018). During data analysis, researchers make dozens of decisions: whether to include outliers, whether to include covariates in a statistical test, and so on. Researchers may consciously or unconsciously make choices that lead to significance. The term "*p*-hacking" reflects that researchers' choices are driven by the desire to obtain a significant *p* value rather than by the goal of determining how well the data support the hypothesis (Simmons et al., 2011; Simonsohn et al., 2014). Many researchers argue—and rightfully so—that there is nothing inherently wrong with exploratory data analysis. However, such explorations become misleading when researchers fail to disclose all of the data permutations and statistical dead-ends they pursued.

Blurring the Line by HARKing

A third practice carries the disparaging name, HARKing, for *hypothesizing after the results are known* (Kerr, 1998). HARKing is the practice of presenting the data collection process as if the results were expected all along. To illustrate how widely accepted this practice once was, consider Bem's (2004) advice:

> There are two possible articles you can write: (a) the article you planned to write when you designed your study or (b) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (b). (pp. 171–172)

Bem argued that HARKing makes an empirical article easier to read. He is probably correct; however, as with underreporting and *p*-hacking, HARKed effects are less likely to be robust. HARKing, most clearly of all these practices, presents an exploratory process as if it were a confirmatory one.

The Transparency Solution

The solution to all three problems has been a call for radical transparency and openness in reporting of science (Table 1). This push for transparency is one reason people use the label Open Science Movement. Transparency applies to three areas: Method sections, hypotheses, and raw data. Journals have begun awarding "badges" to published work that uses these new standards (see Figure 1).

Transparent Methods Sections: Open Materials

Researchers are now expected to disclose every study detail. Some journals have eliminated word limits on Method sections so they can be as long as necessary for full disclosure. Many journals also enable the use of online supplementary materials, where researchers provide the complete list of conditions and variables used in their studies. An additional approach is the "21 Word Solution," with which researchers announce in their Method section that "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study" (Simmons et al., 2012). When researchers disclose all of the conditions and variables they included, readers are better able to evaluate the strength of the evidence and science makes more progress.

Transparent Results: Open Data

Another form of transparency is known as "open data," in which researchers publicly share the data files from published research. Open data also mean that researchers share how they prepared the data and how they computed any composite scores. Before sharing data, researchers remove any identifying information, and when data are impossible to anonymize, they might not be shared in their entirety. Some journals publish the data files with the published manuscript; other scholars publish their data separately on the Open Science Framework (osf.io).

There are multiple benefits to sharing data openly (Chambers, 2017). Open data can be reanalyzed to confirm the results. Open data help address underreporting and *p*-hacking. It is easier to detect fraud because independent scientists can look for patterns in the data that are consistent with data fabrication. Data are stored for posterity, and other scientists can use an old data set to test novel hypotheses. Finally, shared data seem ethically reasonable for studies funded by taxpayers who arguably deserve to have access to the knowledge gained by the investment they have made.

Transparent Hypotheses: Preregistration and Registered Reports

Preregistration—the antidote to HARKing—involves answering specific, structured questions about study design, hypotheses, and data analysis. Researchers document these decisions in advance, often in great detail, even down to drafting the code and posting the syntax files needed to analyze the data. When filed through public repositories such as aspredicted.org or the Open Science Framework, registrations carry a time stamp. Therefore, it can be clearer that researchers collected and analyzed their data only after preregistering.

Registered reports take preregistration a step closer by integrating it with peer review. Researchers write a plan (introduction, method, and data analysis) and submit it to a journal before commencing data collection. Peer reviewers can

Table 1. Summary of Practices Related to Research Transparency.

Term or Practice	Description	Context, Significance, or Impact	Teaching Notes for Psychology Courses
Questionable practices			
Underreporting of null findings	When a researcher conducts multiple studies or includes multiple dependent measures in a single study, they sometimes report only the studies or dependent measures that showed the predicted effect. This means that null findings are never reported or are underreported	This practice misleads the reader about the strength and consistency of the evidence. It also misleads by presenting exploratory research as confirmatory.	When teaching about the theory-data cycle, ask students, What should a scientist do if their study does not support the hypothesis? Should they hide that result or make it publicly available?
p-hacking	During data analysis, a researcher may unintentionally exploit flexibility in data decisions, such as which outliers or covariates to include or which statistical tests to run. Decisions that led to a significant *p* value remain in the research report, while less successful analyses go unreported	*p*-hacking inflates the Type I error rate, so *p*-hacked results are more likely to be flukes that cannot be replicated	Have your students try the *p*-hacking demonstration available at fivethirtyeight.com: "hack your way to scientific glory." Teach estimation thinking rather than significance testing (see Table 2)
HARKing: Hypothesizing after the results are known	After finding an unexpected result, researchers go back and revise the hypothesis, pretending they had predicted the result all along	Theoretically unexpected results are potentially interesting, but they can mislead about the importance or magnitude of an effect. Therefore, unexpected results should be treated with caution until they are replicated. HARKing misleads readers about the strength of the evidence	You can teach HARKing as the opposite of the theory-data cycle. A hypothesis should be like placing an advance bet: you bet that the study will come out a particular way and scientists accept the results no matter how they come out
New practices			
Full transparency via open materials and open data	Full transparency includes sharing every dependent variable and every statistical choice explored in the study. Open materials refers to the practice of sharing full experimental protocols and stimuli. Open data refers to the practice of sharing one's data files and codebooks so that other researchers can independently analyze the results.	Being transparent about all the steps taken in a particular study provides a more honest account of the strength of the evidence. Open materials allow other scientists to conduct direct replication studies. Open data allows other scientists to replicate a study's analyses independently. Open data also makes *p*-hacking more apparent. Open data allows people to explore other hypotheses in the data, ensuring that our participants' time and data has been fully used.	In an introductory class, discuss the benefits of open data for the general public, especially given that research is often supported by tax dollars. In upper-level statistics and methods courses, students can attempt to replicate published studies with open materials, or replicate published analyses with an open data set (see Open Stats Lab, n.d.).
Preregistration and registered reports	Preregistration is when researchers publicly submit their study's procedures, specific hypotheses, and planned data analyses in advance. The plan is posted online with a time- and date-stamp before data collection begins. In a registered report, the preregistered plan is peer reviewed, and the journal offers conditional acceptance of the resulting study, no matter how it turns out.	Preregistration makes the research team's hypotheses and data analytic decisions public in advance. It helps counter HARKing and *p*-hacking. Registered reports explicitly shift publication criteria away from the strength of the result and toward the importance of the question and the quality of the methods.	Introduce preregistration during theory-data cycle. Students can practice a form of preregistration by making their own predictions during class. After hearing about a study's method, students can submit a prediction (perhaps using a personal response system) before the instructor displays the results.

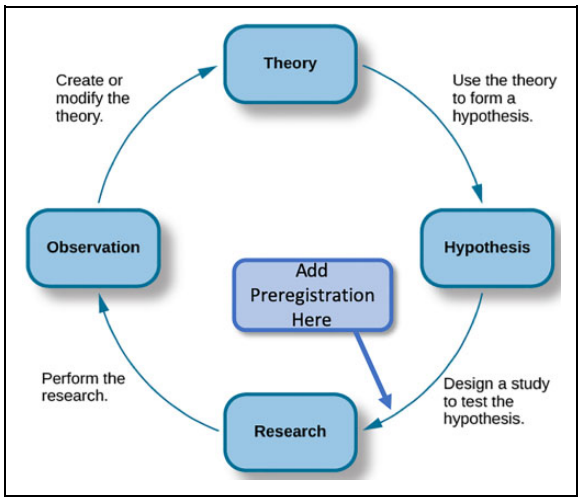

Figure 2. Preregistration can be added as a step in the theory-data cycle. Here, preregistration is added to an existing figure from an open introduction to psychology text. *Source:* https://courses.lumenlearning.com/wmopen-psychology/chapter/outcome-the-scientific-method/

recommend "conditional acceptance" if the study is deemed important, rigorous, and feasible. This process not only controls against HARKing and *p*-hacking; it also changes the incentives. Previously, a journal's evaluation of a paper might have been based, in part, on the strength of its results. Indeed, journals have long been accused of preferentially publishing only significant results—a bias that can lead to the "file drawer problem" (Rosenthal, 1979). In contrast, registered reports are evaluated on the importance of the research question and the quality of the proposed methods for testing that question (e.g., Benjamin, 2019; Lindsay, 2017).

Teaching About Transparency

The practices we have explained here—underreporting, *p*-hacking, HARKing, transparency, registered reports, and data sharing—all belong in the curriculum for the modern psychology major. The details may overwhelm students in the introductory course, but in even there, teachers should add the theme of transparent science.

Teach Preregistration

Teaching the theory-data cycle in psychology courses is essential: It reinforces that psychology is a science because researchers test their ideas with data. Teachers can modernize their coverage of the theory-data cycle by discussing the practice of preregistration (see Figure 2). Links to preregistration are usually available when an article is published.

Prompt Students to Predict

The modernized theory-data cycle merges nicely with the pedagogical technique of prediction. Prediction (rebranded in your classroom preregistration) is a simple and powerful teaching tool; in order to make a prediction, students have to activate what they already know, consolidate it, and apply it. Predictions create positive emotions, such as interest and anticipation, that can facilitate learning (Kornell et al., 2009; Ogan et al., 2009). Teachers can describe a theory, explain the study procedure, and then ask students to predict the results. Students can preregister their predictions in a notebook, turn to a classmate, use a clicker, or even sketch a graph to document their predictions.

Discuss Transparency

Another suggestion for teaching about transparency is through brief discussions, perhaps jump-started by clicker questions. Specifically, after discussing the scientific method, teachers can ask students,

> What should scientists do if they conduct a study and the results do not support their hypothesis?
>
> (a) ignore the result
> (b) make the result publicly available
> (c) revise the hypothesis as if it was expected all along
> (d) something else

Most students will *not* select (a) or (c), even though these were common practices in psychology before the dawn of the credibility revolution. The discussion that follows is likely to be engaging and informative about the processes of science.

Ideas for Upper Level Students

For statistics and methods students, the website fivethirtyeight.com produces an online tool called "[*p*-]hack your way to scientific glory." As students investigate the correlation between political party and economic growth, they can try different combinations of variable operationalizations until they get a *p* value under .05.

Upper level statistics and methods students could replicate published data analyses using open data. The Open Stats Lab (https://sites.trinity.edu/osl) provides articles, datasets, and activities that use open data to illustrate most of the tests taught in undergraduate statistics courses.

Supplemental Materials

We offer three suggestions for supplemental assignments to introduce the replication crisis. One is artist Miki Naro's graphic depiction of the key moments, called "Repeat after Me" (Naro, 2016). The second is a video from the HBO comedy show *Last Week Tonight* (search "Scientific Studies" but warn students of adult language). Finally, several journalists have covered the replication crisis. One excellent example is Engber (2017), who described how Bem's (2011) precognition studies arguably started it all.

Table 2. A Brief Comparison of the New Statistics to Null Hypothesis Significance Testing.

Concept	Null Hypothesis Significance Testing	Estimation (New Statistics)	Teaching Notes for Estimation
Research question	Binary and qualitative: Is there a difference?	Quantitative: How big is the difference?	Teach students to ask, "How much?" Supplement textbook coverage with effect size information
Statistical reporting	Test statistics and *p* value	Effect size with confidence interval	Teach students to ask, "How wrong?" Use simulations to show how point estimates and confidence intervals vary from sample to sample
Conclusions in context	Reject the null or fail to reject. A single result seems definitive, need for replication not clear.	The confidence interval indicates the range of values still compatible with the data. Highlights uncertainty, fostering replication, and meta-analysis.	Teach students to ask "What else is known?" Use forest plots to teach both quantitative reasoning and meta-analytic thinking.
Error framework	False positives (Type I) and false negatives (Type II)	Sign error (Type S: Is the difference in the wrong direction?) and magnitude error (Type M: Were we wrong about the strength of the effect?)	

Estimation Thinking

The second area of rapid change is in the area of statistical inference. Here, the key development is an increasing emphasis on psychology as a quantitative endeavor that accumulates knowledge through replication and synthesis. Often described as the "New Statistics" or the "Estimation Approach" (Cumming, 2012, 2014), it involves reporting and interpreting effect sizes ("How much?"), countenancing uncertainty in all statistical conclusions ("How wrong?"), and synthesizing results through meta-analysis ("What else is known?").

The estimation approach stands in contrast to the decision-making approach, which has dominated psychology (see Table 2). In the decision-making approach, researchers ask a qualitative, binary (yes/no) question ("Do antidepressants treat depression?") and use a *p* value to make a qualitative conclusion ("Yes, antidepressants reduce symptoms of depression," $p < .0001$). Researchers often treat their results as definitive, so there is little motivation to conduct replications. Moreover, there is often scant attention to practical significance. To be clear, this is *not* the way *p* values were meant to be used (e.g., Fisher, 1926), but this "one-and-done" approach to statistical inference has, until recently, been pervasive.

Estimation offers a different lens for interpreting data, one that should nudge researchers toward more thoughtful and nuanced conclusions. Rather than a binary question, estimation focuses on quantifying effects ("To what extent do antidepressants treat depression?"). Results focus not on a *p* value but rather on an effect size and an expression of uncertainty ("The average improvement with antidepressants was 10%, 95% CI [7%, 13%]"; Horder et al., 2011). Estimation—which emphasizes uncertainty—leads to meta-analysis, which combines data to reduce the uncertainty. Estimation thinking is a broad statistical tradition encompassing both parametric and nonparametric techniques as well as Bayesian and classical approaches to probability.

Teaching Estimation Thinking

Introducing estimation into the undergraduate curriculum is overdue. Since the 5th edition of the *APA Publication Manual*, the APA (2001) has recommended estimation as "the best reporting strategy" (p. 22; see Fidler, 2010). Specifically, the APA (2020) enjoins authors to "wherever possible, base discussion and interpretation of results on point and interval estimates" (p. 88; see Appelbaum et al., 2018). Following on these recommendations, many journals require the use of the estimation approach (see Giofrè et al., 2017) and textbooks are increasingly teaching estimation as the default choice for reporting and interpreting results (e.g., Harringon, 2020; Morling, 2020).

Introductory coursework is the ideal time to foster estimation thinking. Teachers can use the prompt, "How much?" to help students consider the magnitudes of effects and to seek context. Using the prompt, "How wrong?" can encourage students to embrace uncertainty and to introduce the key idea of sampling variation. Finally, prompting students with, "What else is known?" helps them see science as a cumulative and integrative process rather than as a series of "one-and-done" demonstrations. These three questions instill a nuanced view of science, where any one study is tenuous, and yet the cumulative evidence from a body of research can be compelling. This is a sophisticated epistemic viewpoint that avoids both excessive confidence and undue cynicism.

Certain psychology courses afford little time to cover statistical inference. Will students miss out if teachers emphasize estimation over decision-making with *p* values? No. In fact, the limited coverage of statistics provided in many textbooks is often incorrect, so it probably instills

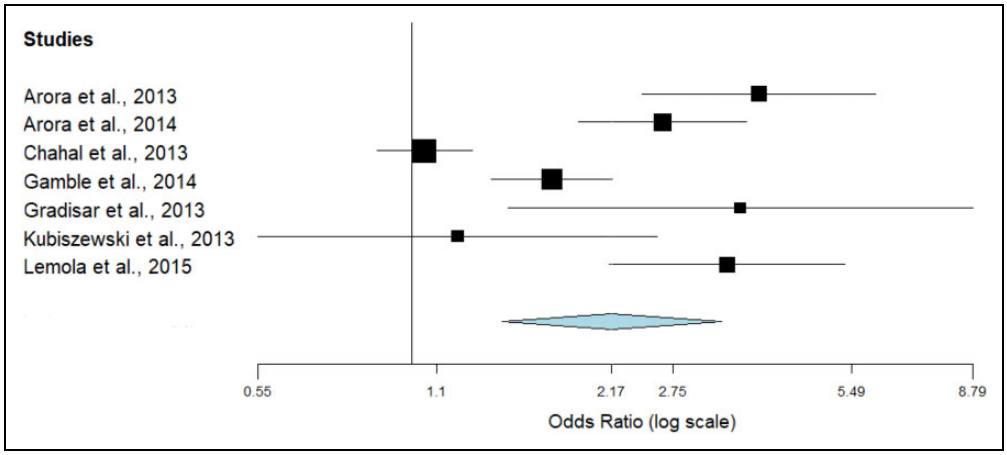

Figure 3. Forest plot of a meta-analysis of the relationship between poor sleep quality and access to a nighttime media device. The effect size here is an odds ratio; it is the chance of having poor sleep with access to a device compared to those without access. Values bigger than 1 indicate devices are associated with an increased risk of poor sleep, numbers less than 1 indicate devices are associated with a decreased risk, and 1 represents no association. Each square represents the effect size from a single study (with larger squares indicating larger samples); each line represents uncertainty from that study (95% confidence interval [CI]). The studies with bigger sample sizes have less uncertainty and therefore shorter lines. The diamond represents the meta-analytic effect size and uncertainty, with the middle of the diamond indicating the observed change in risk and the width of the diamond indicates the remaining uncertainty (95% CI). Overall, those with access to a device at bedtime were 2.17 times as likely to have poor sleep, but there is still considerable uncertainty about this risk (95% CI [1.42, 3.32]). This figure is replotted from a meta-analysis by Carter et al. (2016; all studies above fully cited there), using OpenMeta [Analyst] (Wallace et al., 2012).

misconceptions rather than accurate knowledge (Cassidy et al., 2019). By emphasizing estimation, students gain practice in nuanced quantitative reasoning and also gain a foundation for understanding decision-making. Specifically, students will come to learn that a significance test is equivalent to checking for the null hypothesis in the CI. Estimation gives students a foundation for understanding statistical significance that can help them avoid common misconceptions (e.g., thinking that statistical significance means a certain result is replicable).

Start With Opinion Polls

Teachers can introduce their students to the basics of estimation through political or attitudes polling. Students are used to seeing single polling results accompanied by a margin of error. The polling estimate (52% support a new referendum) is "how much," and the margin of error (plus or minus 4%) is "how wrong." Online poll aggregators synthesize polls conducted on the same topic, which address "what else is known."

Use Visualizations

Data visualizations from polling aggregators can help make the reality of sampling variation salient and illustrate how consensus can emerge across multiple polls. Teachers can go further by introducing the forest plot, which summarizes results across a meta-analysis. Forest plots allow students to visualize all three estimation-thinking points. The effect size of each study shows "how much." Each study includes a line indicating the CI or "how wrong." Finally, showing every study and plotting a meta-analytic effect size indicate "what else is known." In sum, a forest plot illustrates variation across studies as well as the emerging consensus across results.

As an example, Figure 3 shows a forest plot from a meta-analysis of the relation between sleep quality and children's access to media devices at bedtime (Carter et al., 2016). Despite considerable variability across studies, all of them show that access to a media device at bedtime is associated with inadequate sleep quality. Overall, children with access to a media device were 2.17 times more likely to have poor sleep. Even combining across studies, there is still uncertainty, with the 95% CI ranging from 1.41 to 3.32 times the risk. Values outside this CI are also possible. Students could debate the meaning of this relationship and whether parents should ban devices from the bedroom.

Overall, working with students to understand a variety of data visualizations is time well spent. Students can develop the ability to interpret, use, and think critically about quantitative information in an increasingly data-visualized world. A second advantage is that students can practice navigating through a figure using the caption and axis labels.

Explore Simulations

Interactive simulations can help students conceptualize sampling variation. "The Dance of the Means" lets students explore how samples drawn from the same population "dance" around the population mean (Cumming, 2012; https://tinyurl.com/danceofthemeans; see also https://rpsychologist.com/d3/CI/). Such simulations illustrate that even when scientists replicate a study, results vary due to sampling error, with the degree of variation determined in part by sample size. Moreover, students

witness how scientists can take this variation into account, reporting an estimate of uncertainty (a CI) that allows most studies to capture the truth despite sampling variation.

Integrate Estimation Thinking Throughout the Semester

How might teachers reinforce estimation thinking throughout the semester? First, seek opportunities to discuss effect sizes and CIs. For example, while discussing the relation between testosterone and aggression, rather than adopting qualitative language (testosterone *is* associated with aggression in humans), teachers can model estimation thinking by focusing on effect sizes (testosterone predicts about 2% of variation in human aggression; Book et al., 2001).

Second, highlight uncertainty by providing margins of error or CIs for classic studies. In fact, it can be useful to have students rate their confidence in a study from a qualitative textbook presentation and again after being presented with the effect size and uncertainty. For example, if teachers discuss the facial feedback study (Strack et al., 1988), they might share that the original effect was quite uncertain: On a 10-point scale, ratings increased by 0.82 with a margin of error of 0.87 points, giving a 95% CI of [−0.05, 1.69]. In this light, it is actually not surprising that a large-scale replication found an increase of only .03 points (Wagenmakers et al., 2016). Reflecting on uncertainty can help students understand the essential role of replication in science and also develop intuitions for which initial results are most likely to hold up to replication.

A third way to reinforce estimation thinking is to use forest plots and meta-analyses throughout the semester (e.g., Wagenmakers et al., 2016, included a forest plot of replication studies similar to Figure 3). Most textbooks mention only one or two key studies as evidence for a theory. A forest plot can meaningfully extend that coverage and show what happened next. In fact, if teachers help students understand uncertainty in the initial studies, students might even demand to see the meta-analysis!

One difficulty with these recommendations is that, for now, textbooks tend to omit effect sizes and uncertainty, presenting key studies in purely qualitative terms. Fortunately, they often provide key graphs or summary data from which teachers can extract effect sizes and uncertainty (where they do not, Google Scholar can help surface this information). There are also tools for obtaining effect sizes, CIs, and figures from summary data (see, e.g., http://thenewstatistics.com; Cumming & Calin-Jageman, 2017).

Critical thinking involves looking for context, and effect sizes are difficult to discuss in isolation. For example, a recent meta-analysis showed that electroconvulsive therapy produces more improvement than no treatment by an average of 9.7 points on the Hamilton Rating Scale for Depression (UK Electroconvulsive Therapy [ECT] Review Group, 2003). To contextualize these results, students need to learn about the scale: that scores range from 0 to 52 and that a score of at least 21 is required to indicate depression (Hamilton, 1960; Sharp, 2015).

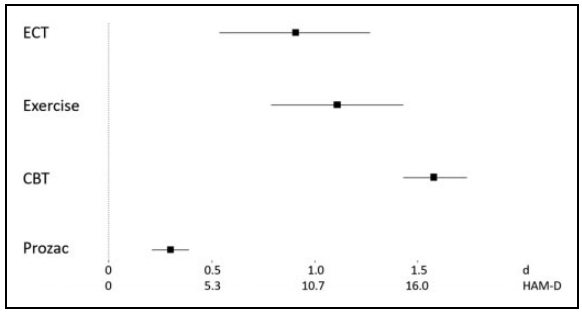

Figure 4. Effect size "zoo" for depression treatments. This figure shows meta-analytic effects of different interventions for depression: electroconvulsive therapy (ECT), exercise, cognitive behavioral therapy, and the antidepressant Prozac (Fluoxetine). Each square shows the meta-analytic effect size relative to control/placebo treatment; the lines show 95% confidence intervals. The meta-analyses reported effect sizes in standardized units (Cohen's *d*). In addition, the HAM-D axis converts these values into expected improvements in the Hamilton Depression Rating Scale (conversion based on the meta-analysis by UK ECT Group, 2003, which translated from d to HAM-D scores using a control group standard deviation of 10.68).

It also helps to provide context from other effect sizes (Figure 4). Providing context can help students think deeply about the scientific issues involved: Are these comparisons meaningful across different types of patients? Is treatment with the biggest effect size recommended or should costs and side effects also be considered?

Closing

In this piece, we have provided teachers with content updates about methodological changes of the past several years. We suggested that teachers should emphasize two fundamental themes: research transparency and estimation thinking. By introducing the value of transparency and the practice of estimation thinking early on, teachers can better prepare students to reason scientifically and quantitatively, readying them for a variety of future paths.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Author.

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Author.

American Psychological Association (n.d.) The APA Introductory Psychology Initiative. https://www.apa.org/ed/precollege/undergrad/introductory-psychology-initiative/

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. https://doi.org/10.1037/amp0000191

Bacon, F. (1889). *New organon, or true directions concerning the interpretation of nature*. Clarendon Press. (Original work published 1620)

Baumeister, R. F. (2019). *Self-control, ego depletion, and social psychology's replication crisis* [Preprint]. https://doi.org/10.31234/osf.io/uf3cn

Baumeister, R. F. (2014). Self-regulation, ego depletion, and inhibition. *Neuropsychologia*, *65*, 313–319. https://doi.org/10.1016/j.neuropsychologia.2014.08.012

Bem, D. J. (2004). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger, III (Eds.). *The compleat academic: A career guide* (2nd ed., pp. 185–219). American Psychological Association.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. https://doi.org/10.1037/a0021524

Benjamin, A. S. (2019). Editorial. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 193–195. https://doi.org/10.1037/xlm0000688

Book, A. S., Starzyk, K. B., & Quinsey, V. L. (2001). The relationship between testosterone and aggression: A meta-analysis. *Aggression and Violent Behavior*, *6*, 579–599. https://doi.org/10.1016/S1359-1789(00)00032-X

Carney, D. R., Cuddy, A. J. C., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, *21*, 1363–1368. https://doi.org/10.1177/0956797610383437

Carter, B., Rees, P., Hale, L., Bhattacharjee, D., & Paradkar, M. S. (2016). Association between portable screen-based media device access or use and sleep outcomes: A systematic review and meta-analysis. *JAMA Pediatrics*, *170*, 1202. https://doi.org/10.1001/jamapediatrics.2016.2341

Caruso, E. M., Vohs, K. D., Baxter, B., & Waytz, A. (2013). Mere exposure to money increases endorsement of free-market systems and social inequality. *Journal of Experimental Psychology: General*, *142*, 301–306. https://doi.org/10.1037/a0029288

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*, *2*, 233–239. https://doi.org/10.1177/2515245919858072

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press. https://doi.org/10.1515/9781400884940

Cuddy, A. J. C., Schultz, S. J., & Fosse, N. E. (2018). *P*-curving a more comprehensive body of research on postural feedback reveals clear evidential value for power-posing effects: Reply to Simmons and Simonsohn (2017). *Psychological Science*, *29*, 656–666. https://dx.doi.org/10.1177/0956797617746749

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

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

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics*. Routledge.

Engber, D. (2017, May 17). *Daryl Bem proved ESP is real: Which means science is broken*. https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html

Feynman, R. (1974). *Cargo cult science*. https://calteches.library.caltech.edu/51/2/CargoCult.htm

Fidler, F. (2010). The American Psychological Association publication manual sixth edition: Implications for statistics education. *Proceedings of eighth international conference on teaching statistics (ICOTS-8)*, Vol. 8.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, *33*, 503–513.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*, 1037.

Giofrè, D., Cumming, G., Fresc, L., Boedker, I., & Tressoldi, P. (2017). The influence of journal submission guidelines on authors' reporting of statistics and use of open research practices. *PLoS One*, *12*, e0175583. https://doi.org/10.1371/journal.pone.0175583

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., Elson, M., . . . Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*, 546–573. https://doi.org/10.1177/1745691616652873

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, & Psychiatry*, *23*, 56–62. https://doi.org/10.1136/jnnp.23.1.56

Harrington, M. (2020). *The design of experiments in neuroscience* (3rd ed.). Sage.

Horder, J., Matthews, P., & Waldmann, R. (2011). Placebo, Prozac and PLoS: Significant lessons for psychopharmacology. *Journal of Psychopharmacology*, *25*, 1277–1288. https://doi.org/10.1177/0269881110372544

Inzlicht, M. (2016, February 29) Reckoning with the Past. http://michaelinzlicht.com/getting-better/2016/2/29/reckoning-with-the-past

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

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 989–998. https://doi.org/10.1037/a0015729

Kosīte, D., König, L. M., De-loyde, K., Lee, I., Pechey, E., Clarke, N., Maynard, O., Morris, R. W., Munafò, M. R., Marteau, T. M., Fletcher, P. C., & Hollands, G. J. (2019). Plate size and food consumption: A pre-registered experimental study in a general population sample. *International Journal of Behavioral Nutrition and Physical Activity*, *16*, 75. https://doi.org/10.1186/s12966-019-0826-1

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*, 1827–1832. https://doi.org/10.1177/0956797615616374

Lindsay, D. S. (2017). Preregistered direct replications in psychological science. *Psychological Science*, *28*, 1191–1192. https://doi.org/10.1177/0956797617718802

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*, 537–542. https://doi.org/10.1177/1745691612460688

Morling, B. (2020). *Research methods in psychology: Evaluating a world of information* (4th ed.). W. W. Norton.

Naro, M. (2016). *Repeat after me*. https://thenib.com/repeat-after-me

Nuzzo, R. (2015). How scientists fool themselves—And how they can stop. *Nature News*, *526*, 182. https://doi.org/10.1038/526182a

Ogan, A., Aleven, V., & Jones, C. (2009). Advancing development of intercultural competence through supporting predictions in narrative video. *International Journal of Artificial Intelligence in Education*, *19*, 267–288.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*, 657–660. https://doi.org/10.1177/1745691612462588

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. https://doi.org/10.1126/science.aac4716

Open Stats Lab. (n.d.). *Open stats lab*. https://sites.trinity.edu/osl

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, *26*, 653–656. https://doi.org/10.1177/0956797614553946

Rohrer, D., Pashler, H., & Harris, C. R. (2015). Do subtle reminders of money change people's political views? *Journal of Experimental Psychology: General*, *144*, e73–e85. https://doi.org/10.1037/xge0000058

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Sharp, R. (2015). The Hamilton Rating Scale for Depression. *Occupational Medicine*, *65*, 340. https://doi.org/10.1093/occmed/kqv043

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. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2160588

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science*, *13*, 255–259. https://doi.org/10.1177/1745691617698146

Simmons, J. P., & Simonsohn, U. (2017). Power posing: *P*-curving the evidence. *Psychological Science*, *28*, 687–693. https://doi.org/10.1177/0956797616658563

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*, 666–681. https://doi.org/10.1177/1745691614553988

Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science*, *11*, 929–930. https://doi.org/10.1177/1745691616674460

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, *54*, 768–777. https://doi.org/10.1037/0022-3514.54.5.768

UK Electroconvulsive Therapy Review Group. (2003). Efficacy and safety of electroconvulsive therapy in depressive disorders: A systematic review and meta-analysis. *The Lancet*, *361*, 799–808. https://doi.org/10.1016/S0140-6736(03)12705-5

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

Vohs, K. D. (2015). Money priming can change people's thoughts, feelings, motivations, and behaviors: An update on 10 years of experiments. *Journal of Experimental Psychology: General*, *144*, e86–e93. https://doi.org/10.1037/xge0000091

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin- Hudon, E. M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., DeCicco, J. M., . . . Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*, 917–928. https://doi.org/10.1177/1745691616674458

Wallace, B. C., Dahabreh, I. J., Trikalinos, T. A, Lau, J., Trow, P., & Schmid, C. H. (2012). Closing the gap between methodologists and end-users: R as a computational back-end. *Wiley Interdisciplinary Reviews Computational*, *49*, 1–15.

Author Biographies

Beth Morling is a professor of Psychological and Brain Sciences at the University of Delaware, where she teaches research methods, cultural psychology, a seminar on the self-concept, and a graduate course in the teaching of psychology. She is a graduate of Carleton College and of the University of Massachusetts at Amherst. She contributes to APS's *Teaching Current Directions* and maintains a blog (everydayresearchmethods.com) about teaching psychological science in the news.

Robert J. Calin-Jageman is a professor of psychology and the neuroscience program director at Dominican University. He studies the neural mechanisms of forgetting and is active in promoting better Open Science and the New Statistics. He received his undergraduate degree in philosophy from Albion College, his PhD in physiological psychology from Wayne State University, and completed postdoctoral training in neurobiology at Georgia State University.