Introduction to the Special Section on Improving Research Practices:

Thinking Deeply Across the Research Cycle

Alison Ledgerwood

University of California, Davis

Address correspondence to:
Alison Ledgerwood
Department of Psychology
UC Davis
Davis, CA 95616
aledgerwood@ucdavis.edu
Introduction to the Special Section on Improving Research Practices:

Thinking Deeply Across the Research Cycle

The past five years have witnessed a profound shift in the way psychological scientists think about methods and practices. Whereas the prevailing sentiment was once a general contentment with the status quo, despite occasional rumblings from methodologists and statisticians (Cohen, 1992; Greenwald, 1975; Maxwell, 2004; Rosenthal, 1979), most psychologists now agree that we could do better. Our growing momentum has placed psychological science at the cutting edge of a broad movement to improve methods and practices across scientific disciplines. And as we move from debating whether we should change to investigating how best to do so, we are increasingly coming to grips with the fact that there are no magic bullet solutions.

Of course, as psychologists, we know that humans (including ourselves) often opt for cognitive shortcuts—that people tend to love a good heuristic, a simple decision rule, an easy answer (Chaiken & Ledgerwood, 2014; Kool, McGuire, Rosen, & Botvinick, 2010; Taylor, 1981). Yet we know, too, that oversimplified decision rules contributed to the problems with our methods and practices that we now face. For example, enshrining $p < .05$ as the ultimate arbiter of truth created the motivation to $p$-hack. Heuristics about sample sizes allowed us to ignore power considerations. Bean-counting publications created an immense pressure to build long CVs. The single most important lesson we can draw from our past in this respect is that we need to think more carefully and more deeply about our methods and our data. Heuristics got us into this mess. Careful thinking will help get us out.

After all, science is hard, and reality is messy. These are complex issues we are tackling, and they deserve nuanced and thoughtful solutions. To that end, the May 2014 and November
2014 special sections in *Perspectives on Psychological Science* provided a toolbox of concrete strategies that researchers can use to think more carefully about their methods and to learn more from their data (Ledgerwood, 2014a, 2014b). The current special section hammers home the importance of thinking carefully at every stage of the research process, from selecting among possible research strategies, to analyzing our data, to aggregating across studies to build a more comprehensive picture of a given topic area.

The special section begins by asking how we can think carefully about optimizing our choice of research strategy. Of course, selecting any one research strategy necessarily involves tradeoffs (Brewer & Crano, 2014; Campbell, 1957; Finkel, Eastwick, & Reis, in press). For instance, increasing the sample size of a given study boosts the power of that study, but at the cost of decreasing the total number of studies that can be conducted (given a finite pool of resources such as time, money, and/or available participants). Miller and Ulrich (this issue) propose a quantitative model that enables researchers to start weighing and integrating the costs and benefits of various research outcomes (true positives, false positives, true negatives, and false negatives) and to calculate an optimal sample size that maximizes what they call *total research payoff* across these study outcomes. Importantly, the model allows researchers to specify the value they place on different study outcomes, so that the calculation of optimal sample size can be tailored to a specific researcher’s values or to a particular research area (e.g., how important is it to this field at this time for true positives and true negatives to be correctly identified?). And, using such a model enables researchers to be explicit about the value they are placing on different study outcomes and the assumptions they are making about the effect size and base rate of true effects in a given research area. This kind of clarity helps reveal when conflicting recommendations about optimal research practices stem from variations in the
starting assumptions of their proponents, allowing us to move beyond intuition-based arguments and toward quantitatively-based discussions about best practices.

Just as we need to think carefully about how our assumptions can influence the conclusions we draw about optimal research strategies, the next two papers emphasize the need to think carefully about how the assumptions that we make as we construct our datasets and analyze our results can influence the conclusions that we draw from our studies. For instance, we all know that even the most common statistical analyses rely on certain assumptions, and yet we may have only a hazy memory left over from past statistics courses about what those assumptions are and how to check them. Ignoring these assumptions can lead us astray when interpreting our results—we might miss an important effect altogether, or conclude that our data provide support for one prediction when in fact they support another. To address this issue, Tay and colleagues (Tay, Parrigon, Huang, & LeBreton, this issue) provide a new tool called Graphical Descriptives, which allows researchers to easily visualize their data, check their statistical assumptions, and transparently communicate rich information about their dataset when writing up their results.

Even before we analyze our data, we make a series of choices about how to combine and/or transform our variables, what data to include or exclude, and how to code various participant responses. Sometimes these choices are arbitrary, or perhaps a researcher prefers one choice (e.g., including all participants who completed the study) but others are also defensible (e.g., excluding participants who failed an attention check). As Steegen, Tuerlinckx, Gelman, and Vanpaemel (this issue) point out, depending on the various data processing choices that a researcher makes, a given raw dataset can give rise to a multiplicity of potential datasets…and therefore a multiplicity of results. Steegen and colleagues offer a new analytic approach, called a
multiverse analysis, that researchers can use to unpack how the various data processing decisions they made could have influenced their results. By examining study results across the full range of plausible data processing scenarios, researchers can explore the extent to which their conclusions might fluctuate across the multiverse of data processing decisions, and they can identify which decisions matter—and which don’t matter—in shaping their results. This tool can also aid researchers in trying to deflate the multiverse: Researchers in a given subfield can work collectively to identify ahead of time whether certain data processing strategies are optimal. It can also highlight how—in the absence of such collective standards—reviewers and editors may unintentionally create sprawling multiverses over time by requesting alternative data processing decisions across different papers.

Finally, we need to be thoughtful about how we aggregate across studies to form a cumulative understanding of a given topic area. Meta-analytic techniques that perform well under a highly simplified set of starting assumptions may lead to incorrect conclusions when we enter the inevitably messier world of real data. For instance, a simplified set of starting assumptions might posit a single true population effect size, no p-hacking, and an oversimplified kind of publication bias where p-values less than .05 are published and those greater than .05 are not. In the real world, however, effect sizes are often heterogeneous, researchers may employ various kinds of p-hacking, and publication bias can take many complex forms.

The last two articles in this special section seek to improve our tools for aggregating research findings by investigating how various meta-analytic techniques perform when some of their idealized assumptions are violated. First, van Aert, Wicherts, and van Assen (this issue) explore how two recently proposed techniques—p-curve and p-uniform—perform under conditions that involve several forms of p-hacking and effect size heterogeneity. Their
simulations suggest that $p$-curve and $p$-uniform produce biased estimates of the average population effect size when various kinds of $p$-hacking have been employed or when there is substantial effect size heterogeneity. Given these issues, van Aert and colleagues caution against using $p$-curve or $p$-uniform when $p$-hacking and/or effect size heterogeneity may be present.

Next, McShane, Böckenholt, and Hansen (this issue) take a step back to think broadly about how meta-analysts can best tackle the thorny real-world complexities of (a) publication bias and (b) effect size heterogeneity. They point out that many meta-analytic techniques account for one while ignoring the other, which can lead these techniques to fall apart when both are present (as they often are in the real world). Regular meta-analysis, for instance, typically accounts for heterogeneity but not publication bias; when publication bias is present, regular meta-analysis not only gives upwardly biased effect size estimates but also typically gives a false impression of homogeneity (leading researchers to erroneously conclude that effect size heterogeneity is not present when it is). Meanwhile, $p$-curve, $p$-uniform, and an earlier version of these approaches (Hedges, 1984) each account for a very simple kind of publication bias while ignoring heterogeneity; when these highly idealized assumptions are violated, they perform poorly. McShane and colleagues therefore urge meta-analysts to fit models that account for both publication bias and heterogeneity (e.g., a basic three-parameter variant of Hedges, 1984, method) and point readers toward a user-friendly website and other resources that enable them to apply this recommendation when conducting meta-analyses of their own.

Together, these articles underscore the importance of questioning and probing the assumptions we make when selecting among research strategies, when interpreting the results of our own individual studies, and when aggregating across studies to draw conclusions about a literature. They suggest that the choices we make about how to create our datasets and the
choices we make about the starting assumptions for a simulation constrain the results we observe on the output side. Any set of results, whether empirical or simulated, give us only a partial picture of reality. Reality itself is always more complex. If we want to study it, we need to be honest and open about the simplifying choices that we make so that everyone—including ourselves—can evaluate these choices, question them, and explore what happens when different choices and assumptions are made.

More broadly, as our field continues to strive for better and better research practices, we need to remember to think carefully about every stage of the research process, and to be skeptical of any simple heuristic, cut-off value, or one-size-fits-all approach. As Spellman (2015) suggested, “after the revolution, we will come to a sensible middle ground.” For those interested in constructing that sensible, thoughtful middle ground, the articles in this section provide an excellent starting point.
References


