R1. Overview

The 38 commentaries on the target article span a broad range of disciplines and perspectives. I have organized my response to the commentaries around three broad questions: First, how serious are the problems discussed in the target article? Second, are there are other, potentially more productive, ways to think about the issues that the target article framed in terms of generalizability? And third, what, if anything, should we collectively do about these problems? Each of these three sections is in turn divided into a number of subsections, each of which summarizes a particular answer to the question given by one or more commentaries.

It should go without saying that the assignment of commentaries to groups is pretty loose. In some cases you kind of have to squint to see it. Also, several commentaries show up in multiple groups, because, despite being under 1,000 words each, they are large; they contain multitudes.

R2. How serious is the problem?

A sensible place to start a review of 38 commentaries on a fairly polemical article is to ask to what extent those commentaries agreed or disagreed with the article's general characterization of affairs. The target article's central claim was that psychology research presently suffers from widespread failure to adequately align verbal hypotheses with their presumed statistical operationalizations, resulting in pervasive generalizability failures: researchers often know very little about what universe of observations their statistical results actually refer to, and consequently draw overly broad conclusions that the statistics do not support on any obvious reading. While the great majority of commentaries expressed substantive agreement with this claim, several did not—and even among those that agreed, there were differences in commentators' positions. Here I characterize four different positions, ranging from outright rejection of the target article's central premise to wholesale acceptance of the argument and exploration of some of the more severe consequences.

R2.1. No big problems

Two commentaries argue that the generalizability-related problems highlighted in the target article focuses are already widely appreciated and have well-established solutions. The stronger position is taken by **Lakens, Tunç, & Tunç**[1], who argue that researchers already have two perfectly sound ways to justify generalizabilty claims—falsificationism and confirmationism. In Lakens, Tunç, & Tunç's view, the argument laid out in the target article constitutes "a third approach built on the impossible ideal of verifying (i.e. conclusively confirming) generalizability claims through random-effect modelling." To be frank, I am not sure how the authors arrive at

[1] For context, the concerns raised here Lakens, Tunç, & Tunç largely recycle ones brought up by Lakens in a much longer open commentary on an earlier draft of the target article (Lakens, ???). I wrote a detailed online rebuttal to that commentary (Yarkoni, ???), which the present commentators appear to have entirely ignored, but that interested readers are encouraged to look at.

this conclusion. So far as I can see, my paper says nothing that could be reasonably construed as a new philosophy of science. It simply points out the direct implication of what seems to me an incontrovertible fact: one cannot pair up verbal claims with statistical models arbitrarily and still claim one is doing science. There is no serious philosophical view under which one may freely draw whatever verbal conclusion one wishes to from a given statistic, irrespective of the latter's consensual meaning. Neither falsificationists nor confirmationists get to pretend otherwise. For the falsificationist, a deductively sound conclusion requires true premises—and how could the truth of a premise like "my theory is falsified if I observe that A > B, $p < .05$" not depend on the specification of the model that produced the *p* value? Correspondingly, how could a confirmationist ever conclude, as Lakens, Tunç, & Tunç suggest, that "subsequent observations enlarge the set of positive instances predicted by the theory" unless the confirmationist understands what set of instances a given statistical model plausibly refers to? When an author observes A > B in a particular experiment, should they write in their Discussion that the effect is present for all possible populations, for only the specific observations in the sample, or for some intermediate universe in between? And just how do Lakens, Tunç, & Tunç think any researcher—be they a falsificationist, confirmationist, or anarchist—could make such a determination without thinking carefully about their model specification?

Gilead argues that the generalizability of a finding is often only a secondary concern, as researchers often operate in other modes of investigation. There are two ways to read this concern. One reading is that Gilead is arguing that psychologists don't always need to lean on inferential statistics so heavily; that when they are doing what Gilead calls *naming* or *causal ontology*, they can rely on other methods of inference. If this is Gilead's point, I agree with it—indeed, several of my recommendations are to exactly this effect. But there is another reading under which Gilead is saying something much stronger—something more like, *hey, lighten up—the way people use inferential quantities like p-values is fine, even if those quantities don't map onto reality in quite the way the authors' words seem to suggest*. The latter view seems implied by Gilead's assertion that "...the generalizability of a pattern is fully independent from the claims made about its generalizability"—by which he means, I think, that it does not matter much what verbal conclusions authors draw from their statistical results, because readers are always free to draw their own conclusions, and so "...it is irrelevant whether an author is grandstanding." If *this* is the intended force of Gilead's argument, it seems to me clearly wrong. The point of writing scientific papers is, I think, to clearly and accurately communicate one's findings to others. The fact that a very motivated reader with enough expertise and free time on their hands could in principle carefully pore over the methods and results of every paper they read before drawing their own conclusions doesn't seem like a good reason to be lackadaisical about the strong claims authors routinely make in their manuscripts.

R2.2. Big problems, but no crisis

Two of the commentaries—**Medaglia & Fernandez** and **Watson—**acknowledge the severity of the problems I draw attention to, but argue that they don't rise to the level of a crisis. The concern, as Medaglia & Fernandez express it, is that "[t]he recent trend to label dilemmas in psychology as 'crises' is insidious," and risks "contributing to bandwagoneering negativity, cynicism, indifference, and antiscientific sentiments." I am sympathetic to this argument in

principle, inasmuch as one can clearly cause harm by exaggerating the implications of a situation—that is, after all, one of the central claims of the target article! But the crisis label seems to me wholly appropriate in this case. Medaglia & Fernandez do not dispute the arguments made in the paper; on the contrary, they explicitly endorse them. Yet the direct implication of these arguments is that psychologists routinely make claims that are not only spurious on a reasonable reading of the marshaled statistics, but often have no meaningful connection to the empirical data at all. If this doesn't constitute a crisis for a field, personally, I have a hard time imagining what would. But no matter: if some readers prefer to think of the target article as describing only Very Serious Problems, and not an actual crisis, they are welcome to do so. Nothing much hangs on it, since Very Serious Problems are presumably also things we should move to rectify with haste.

R2.3. Big problems, *but*...

Several commentaries agree with the general tenor of the target article, but in a qualified way: they argue that the problems discussed in the target article are downstream symptoms of some more fundamental failing, and that the situation is unlikely to improve much until the root cause (whatever it may be) is addressed. I discuss most of these commentaries in more detail in R.3, as they generally include some argument to the effect that the so-called generalizability crisis is better understood or conceptualized in different terms. **Dacey**'s commentary is perhaps unique in that the author embraces my characterization of the problems but nevertheless argues that many of them could be readily eliminated if researchers were to simply "...recognize the distinction in statistics between statistical hypotheses and substantive hypotheses, and to treat them differently from one another". While I don't exactly disagree with this suggestion, I'm not sure what it adds to the analysis. It goes without saying that substantive and statistical expressions are not the same thing, and should be treated separately; indeed, if they *were* the same thing, there would have been no point in the first place in my arguing that researchers should take greater pains to align the two. So what Dacey sees as a solution is to my mind simply a restatement of one of the target article's central premises.

Three other commentaries (**Braver & Braver, Sievers & DeFilippis**, and **Iliev, Medin, & Bang**) argue that the target article is, despite the soundness of most of its arguments, too pessimistic in outlook. Braver & Braver don't like my suggestion that some psychologists may wish to consider a different career or focus more heavily on qualitative research. Sievers & DeFilippis argue that the target article makes much of social science sound hopelessly difficult, and suggest that such pessimism is unwarranted given that there are numerous examples of robust psychology findings in the literature. Iliev, Medin, & Bang only mention that my outlook is "gloomy" in passing, so it's probably unfair of me to include them here; but this way, you see, I have three examples instead of two.

In any case, I am unmoved. **Braver & Braver** presumably do not mean to suggest that it is *never* appropriate for researchers to question whether they could or should be doing something different with their lives; it seems to me irresponsible for publicly-funded researchers putatively working in the public interest *not* to occasionally ask themselves whether what they're doing is worthwhile. I don't doubt that most of the time, the answer will be "yes", and that's great. But if

the answer for some people turns out to be “no” (as it has been for me on more than one occasion), there is no shame in that either.

The same logic applies when determining whether a given research problem is or is not tractable. **Sievers & DeFilippis** would surely agree that the mere fact that many psychology findings are robust does not mean that *every* research question that pops into one’s head must be worth pursuing. It should be obvious that I think there are plenty of sound, widely-generalizable psychology findings out there, or else I would have simply suggested that everyone should quit doing psychology and ended my paper there. The point is, it falls on each individual researcher to ask whether *their* particular question seems likely to yield fruit; the mere fact that there once was a gentleman named Stroop who studied a very robust effect doesn’t absolve every other psychologist of doing an honest cost-benefit analysis. So if my outlook seems gloomy to some, so be it. I don’t *feel* gloomy myself, but hopefully even readers who do find my paper gloomy can allow that the mere fact that it makes them feel gloomy does not mean that its arguments are bad.

R2.4. Big problems, and...

The final, and largest, sub-group consists of commentaries that accept the target article’s central premises more or less as-is, and focus their discussion either on potential solutions to the problems or on further exploration of the implications. I discuss the former set of commentaries in R4; here I focus on the latter—i.e., those commentaries that expand on the issues raised in my article. Several of these commentaries draw attention to the implications for applied research: **Grubbs** focuses on clinical psychology applications; **Wiernik et al.** focus on issues in industrial-organizational settings; **de Leeuw et al.** focus on implications for education; and **Lewin** discusses implications in the legal sphere. I found all of these commentaries lucid and compelling, but lack of expertise in these areas precludes me from adding much of substance. So I will simply double down on what I see as the shared message of these commentaries—namely, a recognition that the rampant overgeneralization common to many areas of psychology is not just a distasteful but benign consequence of systemic pressures and warped incentives, but can and does routinely lead practitioners and policy-makers to deploy suboptimal and even dangerous real-world interventions. Grubbs makes probably the strongest claim in this respect—though one that I think is entirely justified—when he points out that, in clinical psychology (though the point applies to many other applied fields), “...the costs of a generalizability crisis are measured in human lives, not wasted resources.”[2]

[2] Care should of course be taken not to fall into the trap of thinking that the dire consequences described by this group of commentaries apply *only* to applied researchers. So long as basic and applied researchers share departments, students, and journals, any questionable research practices endemic to one side cannot help but seep into the other—producing deleterious downstream societal impacts even when originated by communities that might sincerely believe they owe society nothing more than the satisfaction of their own intellectual curiosity.

Two commentaries focus on implications for basic research in specific domains of psychology. **Visser et al.** describe key features of the ManyBabies initiative (Frank et al., 2017), and illustrate how these can help address many of the problems described in the target article. I liked this commentary, and have nothing else to say about it except *holy hell, that's a lot of authors for a 1,000-word commentary*. **Harris et al.** discuss implications for moral and political psychology.

Lastly, several commentaries suggest that the target article may actually have *under*stated the severity of the problems it describes. **Turner & Smaldino** point out that theories in psychology are often so underspecified as to be essentially untestable—in which case, what does it even matter which variance components are or are not included in a model? **Gelman** observes that the problems I discuss are not limited to psychology, and also pervade many other sciences. The latter point is also echoed by **Maniadis**, who notes that similar troubles afflict experimental economics, a field that is (at least on its face) far more quantitatively rigorous than most of psychology.

R3. Other ways to conceptualize the problem

R3.1 Lack of theory

Several commentaries view the root problem underlying the issues the target article describes as a lack of adequate theory. Appeals for more theory in psychology are, of course, an old phenomenon—though they do seem to be experiencing something of a renaissance recently. I am not, I confess, a fan of such calls. Actually, I hate them. I think they tend to suffer from a number of serious flaws. For one thing, many never bother to tell us what they actually mean by *theory*, and on the most plausible reading, the term often approaches vacuousness. Several of the present commentaries (e.g., **Maniadis**; **Harris et al.**; **Visser et al.**; **Lakens, Tunç, & Tunç**; **Davidson et al.**; and **Turner & Smaldino**) fall into this category. Most of these commentaries call for more theory only tangentially, so it is perhaps unfair to expect a detailed explication. But in a couple of cases, lack of theory is the primary focus of the commentary, and still the reader is given no clear definition. For example, **Hensel, Miłkowski, & Nowakowski**'s titular claim is that "Without more theory, psychology will be a headless rider." The reader is never actually told what Hensel et al. *mean* by theory, but the definition must be an inclusive one indeed, for the authors take pains to note that, despite the absence of any explicit discussion of theory in the target article, "[Yarkoni] wouldn't have been able to make his case without appealing to theoretical insights." In other words: *Yarkoni thinks he can talk about problems in psychology without ever mentioning theory, and yet his argument is itself heavily theoretical. Hoist by his own petard!*

Let us suppose that **Hensel et al.** are right. Well, what then? So far as I can see, my argument relies almost entirely on a bit of statistics and some common sense. It makes virtually no appeal to any domain expertise in psychology. If even an argument like *that* is to be considered theoretical, then it appears all the authors really mean by *theory* is, roughly, *careful thinking*.

This is, admittedly, a difficult position to argue against; who would dare suggest that what psychology needs more of is *sloppy* thinking? The trouble is, Hensel et al. don't tell us how to differentiate good theory from bad theory, or how a bad theorist might go about becoming a better one. "Theorizing," we are simply told, "is an activity integral to any scientific approach regardless of its specific aims and methods. It transcends the difference between qualitative and quantitative research." But when theory is everything, it is also nothing. No surprise if Hensel et al. believe that "all the shortcomings of current practice discussed by Yarkoni come from a common source: researchers' inadequate appreciation of how various theoretical considerations should inform the decisions made at every stage of scientific investigation." It could hardly be otherwise, seeing as virtually every research error stems in whole or part from a failure to reason correctly about *something*. But pointing this out seems about as helpful as suggesting to a figure skater who just took a nasty fall that their problems all stem from "poor technique".

A similar concern applies to **Turner & Smaldino**'s call for greater use of mechanistic models in psychology. Here, again, the feel-good conclusion (that psychologists should use more mechanistic models) is belied by a lack of clarity about what concrete approaches the authors are actually advocating. We are told that mechanistic explanations "can help by forcing the researcher to articulate their guiding assumptions, decomposing their study system into the parts, properties, and relationships", but we are not told what a mechanistic explanation actually *is*. When a researcher puts forward what they claim to be a mechanistic model of a phenomenon, on what grounds should we evaluate that claim? Is it the amount of math involved that matters? The biological plausibility of the claims? The subjective sense of satisfaction a putative explanation elicits? The precision of its predictions? And how good is good enough, for any of these criteria? We simply don't know.

One might be tempted to dismiss this concern by saying that even if it's hard to give a principled definition of *theory* or *mechanism*, day-to-day usage is sufficiently consistent that it doesn't really matter. In other words, *we all know it when we see it*. That view is intuitively appealing, but also wrong. I have argued elsewhere that efforts to unpack such terms almost invariably reveal them to depend heavily on authors' particular intellectual and aesthetic preferences—which, unsurprisingly, tend to differ widely across individuals (Yarkoni, 2020). But you don't have to take my word for it; we can observe this phenomenon in the present commentaries. Consider the pieces by **Dickins** and **Donkin, Szollosi, & Bramley.** Both argue that psychology needs better theory, yet their concrete prescriptions diverge in ways that are not obviously reconcilable. Dickins asserts that grounding psychology in deeper theory requires "seeking some unity with biology, through the adoption of highly corroborated theories such as evolutionary theory"; Donkin, Szollosi, & Bramley, by contrast, make no appeal at all to biology or evolution, and instead suggest that the "primary explicanda of psychology are people's capacities", and hence, "[p]sychological explanations should not only account for what people did in some experiment, but also for what they could have done". Of course, these are only two particular positions in a literature replete with differing views as to what skill set or body of knowledge is conducive to good theory. A naive reader could be forgiven for concluding they need to be a polymath before they can start to develop "good" theory.

To be clear, I am not suggesting that pro-theory arguments like these are *wrong*, exactly—only that they are unhelpful. There are innumerably many reasons why any given statistical result might fail to support a particular verbal claim, and there is little reason to suppose that, say, a social psychologist's failure to consider stimulus variability in an IAT task has much in common with an educational psychologist's failure to consider variability in instructor quality in a study of flipped classrooms. It would be pleasant to think we could eliminate most, or even many, generalizability-related problems simply by convincing psychologists to think more about evolution or biology or culture or whatever—or just to think more carefully in general. But I think this would be an exercise in wishful thinking.

R3.2. Generalizability from a construct validity perspective

Two commentaries approach the issues raised in the target article from a traditional psychometric perspective. **Flake, Luong & Shaw** argue that a productive way forward is to emphasize large-scale construct validation—that is, to conduct extensive descriptive research aimed at ensuring that one's measures are actually measuring what they're supposed to be measuring. **King & Wright** echo this suggestion and further point out that the problems the target article describes in terms of generalizability can be equivalently construed in terms of construct validity—specifically, the assertion that statistical expressions ought to map closely onto verbal/theoretical expressions can be restated as saying that measures should be valid operationalizations of the constructs they are meant to represent.

I am broadly sympathetic to these commentaries—though I do have some concerns about **Flake, Luong, & Shaw**'s motives (see their conflict of interest statement). My only (minor) reservation is a practical one: framing things in terms of construct validity carries a certain amount of psychometric baggage that in my view can be counterproductive. Both Flake, Luong, & Shaw and **King & Wright** view construct validity as something one ought to establish *before* one starts computing inferential statistics, making predictions, and so on. I think this is good advice for researchers with a realist orientation who construe their research as a search for the latent causes of people's behaviors (for discussion, see Yarkoni, 2020). But this is not the only view one can take. I have argued previously that the data-generating processes underlying many psychological phenomena may simply be too complex and messy for traditional psychometric models to have much utility, so that in practice, the most effective way to make progress may be to largely set aside psychometric concerns about (internal) validity and instead focus more on developing predictively useful models, however complex or uninterpretable they may be (Yarkoni & Westfall, 2017; Rocca & Yarkoni, in press). I won't defend the latter position here, but am simply observing that the framing I adopted in the target article deliberately sought to minimize theoretical commitments and describe the problem in a maximally general way.

R4. Solutions

The fourth and largest group of commentaries focused on describing one or more solutions to the problems identified in the target article. I have organized these into 4 sub-groups. Respectively, they include commentaries that focus on (1) formal methodologies (either the

general need for greater formalism, or specific techniques); (2) benefits afforded by big data and associated technical developments; (3) various methodological procedures, several of which expand on suggestions made in the target article; and (4) bird's-eye or "meta" perspectives that focus on how resources and incentives organize researchers' efforts at a communal level.

R.4.1. Formal methods

Several of the commentaries call for an increased role for formal methodologies in psychological science. **Turner & Smaldino**'s call is the most general; the authors argue for increased emphasis on mechanistic explanation and formal modeling throughout psychology. I have already explained why I find the mechanistic part of their appeal unconvincing; on the other hand, I enthusiastically agree with their call for greater adoption of formal/computational methods. The main reason for this is that I think there's far greater transfer between computational skills than between substantive bodies of domain knowledge, so it matters less precisely *which* computational and mathematical skills are taught in training programs. I'm not sure I've ever heard a psychologist say they regret learning linear algebra or programming, whereas I've heard any number of psychologists lament the time they wasted taking theory-heavy courses in other areas of psychology just to meet distribution requirements.

Ross echoes Turner & Smaldino's call for more formal methodology in psychology, and argues in particular for greater adoption of methods widely used in experimental econometrics—e.g., larger-scale (and more expensive) experiments, and Bayesian estimation. I am sympathetic to many of Ross's specific recommendations, which overlap to some degree with those I made in the target article. That said, I don't think Ross's commentary should be read (or is intended) as an injunction against the use of other modeling techniques and strategies. As **Gelman** points out in his commentary, it may be helpful for scientists to think of statistics as a box of heterogeneous tools, where "[d]ifferent models and statistical tests capture different aspects of the data we observe and the underlying structure we are trying to study". A central message of both commentaries is that it is hubristic to suppose that mindless application of statistical significance tests could produce meaningful answers to most of the questions psychologists pose—so authors should be prepared to develop a broader toolset.

Braver & Braver and **Bonifay** focus on more specific techniques. Braver & Braver echo my call for a greater focus on variance decomposition approaches, and offer valuable design recommendations (e.g., to try and systematically vary at least one purportedly irrelevant factor in every study). They also take issue with what they see as my unwarranted dismissal of conceptual replications, pointing out that it's entirely possible to aggregate conceptual-related experiments that don't share common design elements via meta-analysis. But I don't think I *dismissed* conceptual replications so much as observed that it is difficult to integrate the results of conceptual replications in a principled way. I stand by this claim. It is of course true that one can aggregate estimates from *any* set of studies via meta-analysis. The trouble is that meta-analyses of conceptual replications suffer from the same garbage-in-garbage-out (GIGO) principle as all other meta-analysis applications: it is extraordinarily easy for authors publishing a string of conceptual replications to selectively report heterogeneous studies and analyses that support their favored story, at which point a meta-analysis cannot help but reify those biases

already baked in by selective reporting and construal. By contrast, explicitly varying multiple design factors within a single study (a strategy that, to be clear, Braver & Braver also endorse) makes it more difficult—though certainly not impossible—for authors to fool themselves and others.

Bonifay focuses on the minimum description length (MDL) principle as a means of reducing overfitting, and hence (indirectly) also generalization errors. The MDL is one in a class of information theoretic techniques that formally attempt to mitigate overfitting by penalizing models for complexity. Bonifay argues that the MDL "...offers insights into overfitting and generalizability that are not possible using traditional methods", and this may be true to some degree. At the same time, I think Bonifay's commentary somewhat oversells the benefits of MDL. Quoting from Grünwald (2004), Bonifay suggests that the MDL "automatically and inherently protects against overfitting." This is somewhat misleading: the MDL is an idealization, and prevents overfitting only in principle. In practice, there is no way to deterministically compute the shortest possible description of a dataset, or even verify that a given proposal is optimal. Specific MDL algorithms *are* computable, but necessarily introduce inductive biases, and hence can and do overfit (they are also restricted to certain classes of models). Moreover, it's important to remember that Occam's Razor (which MDL is a formalization of) is only a heuristic, not a law. The MDL principle offers no guarantee that a favored model adequately captures the true data-generating process, but only that it compactly describes the data. As always, there is no free lunch: specific MDL algorithms will sometimes perform better than other approaches and sometimes worse, but blanket statements to the effect that the MDL principle overcomes standard problems of model comparison seem to me hard to justify.

Bear & Phillips take issue with the target article's advocacy of more expansive mixed-effects models. They argue that the use of random effects is problematic in many common designs, as the inclusion of such terms depends on the assumption that the levels of the random factors are being sampled randomly from well-defined underlying populations, which is clearly false in most cases (e.g., most researchers don't really sample their stimuli at random from some well-defined space). I think this argument runs afoul of Box's famous aphorism that "all models are wrong, but some are useful". The point of including random effects is to adjust parameter estimates to account for presumed sources of variance in the data. It should go without saying that in cases where researchers are able to write down a deterministic expression that more closely approximates the true data-generating process, they should do so (see also footnote 11 in the target article). But such scenarios are extremely rare in psychology. The vastly more common scenario involves a choice between a model that makes *no* effort to account for obvious sources of variability in the data, and one that makes an effort to do so imperfectly. So while Bear & Phillips' titular claim that "random effects won't solve the problem of generalizability" is trivially true—after all, *nothing* can ever guarantee safe generalization (even the assumption that the world will exist when we wake up tomorrow requires an inductive leap of faith!)—this is hardly a reason to forsake random effects, because the conventional alternative is still worse. A charitable reading of Bear & Phillips is that they are simply pointing out that there can be more suitable formalisms than mixed-effects models in many cases—a view I agree with, and

explicitly endorsed in the target article (e.g., “Of course, inclusion of additional random effects is only one of many potential avenues for sensible model expansion…”).

Finally, **Maniadis** (and to some degree also **Gelman)** provides an important counterpoint to the other commentaries in this group by observing that increased formalism alone will not suffice to solve the generalizability crisis. Maniadis points out that there are other fields (e.g., experimental economics) that already emphasize formal methods to a far greater extent than psychology, yet suffer from very similar problems. This is a point worth reaffirming: while there can be little doubt that greater statistical sophistication in psychology would improve the state of affairs, it is clearly neither necessary nor sufficient to ensure that researchers produce defensible scientific inferences. Maniadis puts it well in emphasizing the need for caution, observing that “while formalism makes excessive ad hoc theorising more difficult, it does not rule it out”.

R.4.2. Benefits of larger, richer datasets

Several commentaries highlight the utility of large, rich datasets in addressing concerns about generalizability, and emphasize the critical role of technology in facilitating the acquisition or analysis of such datasets. I enthusiastically endorse the approaches promoted in these commentaries, and have made similar arguments myself in the past (e.g., Yarkoni, 2012). Three of the commentaries focus on the utility of closely-related crowdsourcing (**Cyrus-Lai et al.**), “citizen science” (**Hilton & Mehr**), and “many labs” (**Visser et al.**) approaches. The key point here is that establishing the generality of an effect usually requires datasets that sample from a broad universe of observations, and acquiring such datasets is far easier when researchers leverage the scale of internet-based data collection, or join forces and form research consortia. The common goal is to maximize variation in the data—whether by randomly assigning large samples to diverse conditions; by allowing investigators to operationalize hypotheses as they see fit; or by acquiring data at multiple sites, from multiple populations, using multiple methods.

Davidson et al. illustrate how modern technologies—in particular, digital traces of behavior obtained from smartphone sensors and interactions with mobile applications—can be used to expand the scope of measurement of behavior beyond the traditional emphasis on self-report. There is much to like about Davidson et al’s advocacy for the study of digital traces and large-scale data sharing—though, for reasons already alluded to above (see R3.2), I’m less enamored their assertion that “it is critically important psychology shifts away from predictive validity alone as evidence for successful operationalization and parameterization”. I think there are many scenarios in which psychologists would probably do themselves a favor by focusing *more* heavily on predictive validity, and correspondingly less on traditional indicators of validity. But that position has little bearing on the one advanced in the target article, and is also incidental to Davidson et al.’s central position, so I won’t defend it here (interested readers can see Yarkoni & Westfall, 2017; Rocca & Yarkoni, in press).

Lastly, **Van de Velde, Pascale, & Speelman** discuss the strengths and limitations of corpus linguistics approaches—which emphasize large, naturalistic datasets over small factorial experiments—when used in pursuit of stronger, more generalizable inferences. Van de Velde,

Pascale & Speelman's commentary is notable and refreshing in that the authors emphasize the complexities and tradeoffs involved in adopting corpus linguistics methods, and caution against treating such approaches as a panacea. The point is well taken, and applies well beyond the study of language. The target article provided only a brief sketch of a few modeling strategies that can help close the gap between authors' generalization intentions and their statistical operationalizations; it goes without saying that the central lesson is not that linear mixed-effect models can solve all problems, but rather, that ritualistic reliance on statistical defaults (e.g., the conventional subject-as-random-effect model) rarely leads to good outcomes. Once researchers acknowledges this point, then the difficult work of selecting and specifying a model appropriate to the specific domain and problem at hand can begin—and Van de Velde, Pascale, & Speelman's commentary provides a nice case study of the kinds of considerations that may arise.

R.4.3. Methodological recommendations

A third group of solution-focused commentaries consists of what I'll call "methodological recommendations", reflecting their emphasis on a particular type of methodological practice (generally non-statistical in nature, in contrast to the first subgroup of commentaries in this section). A cynical reader might object that the commentaries I've lumped together here are pretty heterogeneous, so maybe a better characterization would be *all the commentaries that don't fit neatly into any of the other boxes*. And my response to such a charge would be: *hey, look, what's that shiny thing over there?*

West et al. focus on the role of strictly descriptive work in psychology—that is, research that makes no claim to establish causal relationships, but simply seeks to characterize the relationships between various measured variables. The authors describe several guiding principles that can help improve the quality of descriptive research. I broadly agree with their recommendations. My only minor quibble is that West et al. encourage researchers to thoroughly explore their data before performing inferential tests. While data exploration is certainly desirable, conditioning one's choice of inferential procedures on prior examination of one's data is an excellent way to procedural overfit that data (Yarkoni & Westfall, 2017). Researchers who wish to follow West et al's advice should take pains to maintain a clear separation between exploration and confirmation (e.g., via use of preregistration, hold-out datasets, etc.).

Blersch et al. argue for the use of formal causal frameworks as a means of bridging between qualitative and quantitative analysis in psychology. Their commentary echoes other recent appeals for psychologists to embrace causal analysis (e.g., Rohrer, 2018; Grosz, Rohrer, & Thoemmes, 2020). I have mixed feelings about such calls. On the one hand, I agree with the present authors that greater familiarity with the dominant causal approaches (e.g., Rubin's potential outcomes framework and Pearl's work on causal graphs) might help many psychologists better understand the limitations of their models. On the other hand, I think Blersch et al considerably overestimate the power of formalisms like directed acyclic graphs (DAGs) to, as they put it, "bridge between qualitative and quantitative research". The toy example they present in Figure 1 fails to convey what is actually difficult about formalizing

causal relationships in most areas of psychology: it isn't the ability to express one's qualitative hypotheses in terms of nodes or edges (witness the rise of closely-related structural equation models in psychology over the past few decades), but rather, the ability to justify the assumption that *this* particular graph adequately represents the causal phenomena it is intended to stand in for, in the face of innumerable viable alternatives. Contra Blersch et al., such assumptions are usually impossible to test empirically, and strictly logical considerations often dictate their prima facie absurdity anyway-and yet they nevertheless proliferate in the literature unchecked. This is not, I submit, because psychologists haven't read enough Rubin or Pearl, and lack the causal understanding needed to identify such flaws; I suspect it's because asking the obvious questions-for example, "is this single artificial stimulus really an adequate stand-in for a broad construct like *verbal overshadowing*?", or, "isn't it blindingly obvious that the environment, personality, and behavior *all* causally influence one another?"-tends to diminish one's ability, and perhaps also desire, to publish one's work.

Syed & McLean echo the target article's call for more serious consideration of qualitative approaches. A key point the authors emphasize is that a huge amount of the work psychologists currently engage in is already qualitative in nature. Discussion sections unpack the qualitative implications of quantitative results; measurement studies assign qualitative interpretations to factors that are, at bottom, mathematical abstractions; and much of the coded data that enters statistical analyses reflects qualitative assessments. As Syed & McLean observe, "[i]t appears that even qualitative analysis is permissible in mainstream psychology so long as we do not call too much attention to the practice, and do not engage in the intentionality and rigor of best practices in qualitative methods". I strongly endorse Syed & McLean's argument that "qualitative methods can also play a key role in testing, applying, and exemplifying theoretical claims". Indeed, one way to read many of my claims in the target article is to observe that many psychologists now rely on routinely sloppy quantitative methods to justify claims that readers might deem ludicrous if they were presented strictly on the basis of their qualitative merits.

Wilford et al. argue that concerns about generalizability largely stem from the dominance of the stimulus-response (S-R) paradigm within psychology; they advocate for a different paradigm—the *perturbation experiment*—that avoids these issues. It wasn't clear to me what features the authors think define perturbation experiments, or how such experiments manage to avoid the need to ensure an alignment between one's verbal and statistical statements. On one reading, Wilford et al. are reiterating Popper and Meehl's call for "risky" predictions—for example, they write that perturbation experiments "...aim to identify the precise variable or variables implicated in the ongoing control of a complete activity". Accomplishing such a feat would presumably require an experiment to be so carefully operationalized that the outcome rules dispositively in favor of one particular interpretation of a phenomenon. If this is the intended conclusion, I am supportive (and argue as much in the target article). But perturbation as Wilford et al. discuss it doesn't seem either necessary or sufficient for producing risky predictions (e.g., some of the methods Wilford et al. list as intrinsically perturbative in nature, like TMS and lesion studies, are routinely used to draw rather silly conclusions). Moreover, even in the best-case scenario, there remains no escape from the need to align statistical and substantive expressions. To see this, one need only consider the statistics reported in the elegant Adolph, Eppler, & Gibson (1992)

study Wilford et al. hold up as an example of a successful perturbation experiment. Would Wilford et al. continue to argue that the Adolph et al. study provides “unambiguous evidence” for its conclusion if it were later discovered that the statistical model had been incorrectly specified, or if the effect were shown to obtain only in the hands of one particular experimenter? It seems doubtful.

R.4.4. Bird’s eye views

The last set of solutions-focused commentaries take a bird’s eye view of the issues discussed in the target article. Instead of focusing on the mechanics of specific solutions, these commentaries focus on broader cultural issues and incentives, historical perspectives, and cross-field comparisons. Two of the commentaries—**Schiavone, Bottesini, & Vazire**, and **Sievers & DeFilippis**—argue that the problems described in the target article would be more effectively addressed by focusing on community-level practices and incentives rather than on individual researchers’ behavior. I take no position on this claim; the prescriptions I outlined were largely agnostic with respect to implementation. I do, however, think we should be generally wary of arguments to the effect that major cultural changes would, as Schiavone, Bottesini & Vazire write, “...follow swiftly if a small group of gatekeepers decided to make it a priority”. It is true that power is disproportionately concentrated in the hands of a relatively few gatekeepers; but gatekeepers generally do not attain their status by operating outside of mainstream mores: they are typically people who have benefited *more* from the status quo than rank and file researchers, and consequently are both less likely to share a belief in the need for reform and more likely to lose status in the event that reform takes place. Similarly, Siever & DeFilippis’s suggestion that we should “foster a diverse scholarly community that is incentivized to reveal what those who came before them have missed” sounds laudable, but it is not immediately clear that it is any easier a goal to achieve than simply saying “researchers should stop missing things.” The standard trade-off between individual and collective action applies here: individuals acting alone have less power to influence their community, but they are also able to initiate action directly, without waiting for anyone else’s consent or encouragement. I don’t pretend to know what the optimal course of action is in this case—indeed, I am not sure such a thing is knowable, *a priori*—so perhaps letting a thousand roses bloom is the optimal strategy.

The **Gigerenzer** and **Alzahawi & Monin** commentaries provide historically-oriented perspectives on the field, and travel different paths to arrive at a similar (at least superficially) conclusion: we should think more carefully about how we conduct quantitative research in psychology. Gigerenzer observes that many of our current statistical conventions (e.g., modeling subjects but not stimuli as random effects) reflect historical accidents and misunderstanding of statistics, and suggests we “liberate research practice from methodological rituals”. I am very much in agreement with this conclusion, though Gigerenzer’s observation that logical fallacies and misunderstandings like the “replication delusion” are widespread even among psychology professors and statisticians raises some serious concerns about the scope of the task at hand.

Alzahawi & Monin's conclusion is, on the surface, similar to **Gigerenzer**'s: the authors suggest that we should work to highlight "how inferential statistics can be more thoughtfully applied". While the general conclusion is again easy to agree with (no one would argue for *less* thoughtful application), the argument Alzahawi & Monin offer in its support is, in my view, self-defeating and rather cynical. The authors specifically reject any effort to move away from quantitative methods in psychology, arguing that such a thing is "unlikely to obtain", because quantitative methods are presently "core to psychology's social and scientific status". This position conflates explanation and justification. It may be true that psychologists historically rushed to adopt quantitative methods in part because doing so conferred prestige and resources on the field; but surely we should not accept it as axiomatic that misaligned incentives cannot ever change, or reform of almost any kind would become impossible. Ironically, Alzahawi & Monin's closing recommendation to "draw more accurate—if more modest—conclusions from our data" is susceptible to their very own argument. There are presently few cultural incentives for psychologists to be more modest in their conclusions or more thoughtful in their inferences; by Alzahawi & Monin's reasoning, shouldn't this doom their own prescription to failure? Should we construe psychological scientists as no more than prestige-maximizing automatons, incapable of ever privileging what is right over what is expedient?

Ioannidis takes up the question of how to optimally sequence research activities; specifically, he asks whether it is better to focus on replication first and generalization second, or to do the converse. Ioannidis ascribes to me the latter view—i.e., he takes me to favor a sequence that goes "....discover-generalize-replicate, i.e. don't waste time with replication unless a promising research finding has been probed in a sufficiently large variety of settings to have some sense that it is generalizable." He then argues that this strategy has downsides, and that there are many scenarios in which it makes sense to try to replicate narrow findings ahead of any attempt to demonstrate their broader generalizability. I agree with this. In writing that "the current focus on reproducibility and replicability risks distracting us from more important, and logically antecedent, concerns about generalizability," I was not suggesting that establishing generality is a more important *empirical* goal than replicability, only that the decision to replicate a given finding should presuppose an adequate understanding of its plausible implications. Researchers who believe it is more important to directly replicate Experiment 1 of Schooler & Engstler-Schooler (1990) than to expand the scope of its design are welcome to privilege the former. But that decision should be made with full, explicit recognition of the experiment's limitations, rather than implicitly or explicitly equating a very narrow operationalization with the broad construct of interest.

Lastly, **Lampinen et al.** compare and contrast publishing norms in psychology with those in the field of artificial intelligence (AI), and suggest that each field would benefit from adopting some of the habits of the other. I lack the expertise necessary to evaluate this recommendation with respect to AI, but I largely agree with Lampinen et al,'s suggestion that psychology would benefit from increased adoption of informal, rapid publication streams (cf. Yarkoni, 2012). Of course, it's hard to know how much of AI's rapid progress can be attributed specifically to its publishing conventions; I have previously suggested that much of machine learning and AI's success may stem from its emphasis on evaluating models against standardized benchmarks—

a practice that contrasts markedly with psychologists' tendency to choose their own idiosyncratic evaluation metrics on a case-by-case basis (Rocca & Yarkoni, in press). But either way, I agree with Lampinen et al.'s concrete recommendations.

References

Adolph, K. E., M. A. Eppler, and E. J. Gibson. 1993. "Crawling versus Walking Infants' Perception of Affordances for Locomotion over Sloping Surfaces." Child Development 64 (4): 1158–74.

Frank, Michael C., Elika Bergelson, Christina Bergmann, Alejandrina Cristia, Caroline Floccia, Judit Gervain, J. Kiley Hamlin, et al. 2017. "A Collaborative Approach to Infant Research: Promoting Reproducibility, Best Practices, and Theory-Building." Infancy: The Official Journal of the International Society on Infant Studies 22 (4): 421–35.

Grosz, Michael P., Julia M. Rohrer, and Felix Thoemmes. 2020. "The Taboo against Explicit Causal Inference in Nonexperimental Psychology." Perspectives on Psychological Science: A Journal of the Association for Psychological Science 15 (5): 1243–55.

Grunwald, Peter. 2004. "A Tutorial Introduction to the Minimum Description Length Principle." ArXiv [Math.ST]. arXiv. http://arxiv.org/abs/math/0406077.

Rocca, Roberta, and Tal Yarkoni. 2020. "Putting Psychology to the Test: Rethinking Model Evaluation through Benchmarking and Prediction." PsyArXiv. https://doi.org/10.31234/osf.io/e437b.

Rohrer, Julia M. 2018. "Thinking Clearly about Correlations and Causation: Graphical Causal Models for Observational Data." Advances in Methods and Practices in Psychological Science 1 (1): 27–42.

Schooler, J. W., and T. Y. Engstler-Schooler. 1990. "Verbal Overshadowing of Visual Memories: Some Things Are Better Left Unsaid." Cognitive Psychology 22 (1): 36–71.

Yarkoni, Tal. 2012. "Psychoinformatics: New Horizons at the Interface of the Psychological and Computing Sciences." Current Directions in Psychological Science 21 (6): 391–97.

Yarkoni, Tal. 2012. "Designing Next-Generation Platforms for Evaluating Scientific Output: What Scientists Can Learn from the Social Web." Frontiers in Computational Neuroscience 6 (October): 72.

Yarkoni, Tal, and Jacob Westfall. 2017. "Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning." Perspectives on Psychological Science: A Journal of the Association for Psychological Science 12 (6): 1100–1122.

Yarkoni, Tal. 2020. “Implicit Realism Impedes Progress in Psychology: Comment on Fried (2020).” Psychological Inquiry 31 (4): 326–33.