The generalizability crisis

Tal Yarkoni[1*]

[1]Department of Psychology, University of Texas at Austin
[*]Email: tyarkoni@gmail.com

Most theories and hypotheses in psychology are verbal in nature, yet their evaluation overwhelmingly relies on inferential statistical procedures. The validity of the move from qualitative to quantitative analysis depends on the verbal and statistical expressions of a hypothesis being closely aligned—that is, that the two must refer to roughly the same set of hypothetical observations. Here I argue that many applications of statistical inference in psychology fail to meet this basic condition. Focusing on the most widely used class of model in psychology—the linear mixed model—I explore the consequences of failing to statistically operationalize verbal hypotheses in a way that respects researchers' actual generalization intentions. I demonstrate that whereas the "random effect" formalism is used pervasively in psychology to model inter-subject variability, few researchers accord the same treatment to other variables they clearly intend to generalize over (e.g., stimuli, tasks, or research sites). The under-specification of random effects imposes far stronger constraints on the generalizability of results than most researchers appreciate. Ignoring these constraints can dramatically inflate false positive rates, and often leads researchers to draw sweeping verbal generalizations that lack a meaningful connection to the statistical quantities they are putatively based on. I argue that failure to take the alignment between verbal and statistical expressions seriously lies at the heart of many of psychology's ongoing problems (e.g., the replication crisis), and conclude with a discussion of several potential avenues for improvement.

Keywords: generalization, statistics, psychology, random effects, inference, philosophy of science

Introduction

Modern psychology is—at least to superficial appearances—a quantitative discipline. Evaluation of most claims proceeds by computing statistical quantities that are thought to bear some important relationship to the theories or practical applications psychologists care about. This observation may seem obvious, but it's worth noting that things didn't have to turn out this way. Given that the theories and constructs psychologists are interested in usually have qualitative origins, and are almost invariably expressed verbally, a naive observer might well wonder why psychologists bother with numbers at all. Why take the trouble to compute p-values, Bayes Factors, or confidence intervals when evaluating qualitative theoretical claims? Why don't psychologists simply look at the world around them, think deeply for a while, and then state—again in qualitative terms—what they think they have learned?

The standard answer to this question is that quantitative analysis offers important benefits that qualitative analysis cannot (e.g., Steckler, McLeroy, Goodman, Bird, & McCormick, 1992)—perhaps most notably, greater objectivity and precision. Two observers can disagree over whether a crowd of people should be considered "big" or "small", but if a careful count establishes that the crowd contains exactly 74 people, then it is at least clear what the facts on the ground are, and any remaining dispute is rendered largely terminological.

Unfortunately, the benefits of quantitation come at a steep cost: verbally expressed psychological constructs[1]—things like cognitive dissonance, language acquisition, and working memory capacity—cannot be directly measured with an acceptable level of objectivity and precision. What *can* be measured objectively and precisely are operationalizations of those constructs—for example, a performance score on a particular digit span task, or the number of English words an infant has learned by age 3. Trading vague

[1]I avoid the conventional habit of describing psychological constructs as latent variables, as such language is often taken to imply a realist philosophical stance towards theoretical entities (e.g., Borsboom, Mellenbergh, & van Heerden, 2003). For present purposes, it's irrelevant whether one thinks psychological constructs objectively exist in some latent or platonic realm, or are merely pragmatic fictions.

verbal assertions for concrete measures and manipulations is what enables researchers to draw precise, objective, quantitative inferences; however, the same move also introduces new points of potential failure, because the validity of the original verbal assertion now depends not only on what happens to be true about the world itself, but also on the degree to which the chosen proxy measures successfully capture the constructs of interest—what psychometricians term *construct validity* (Cronbach & Meehl, 1955; Guion, 1980; O'Leary-Kelly & J. Vokurka, 1998).

When the construct validity of a measure or manipulation is low, any conclusions one draws at the operational level run a high risk of failing to generalize to the construct level. An easy way to appreciate this is to consider an extreme example. Suppose I hypothesize that high social status makes people behave dishonestly. If I claim that I can test this hypothesis by randomly assigning people to either read a book or watch television for 10 minutes, and then measuring their performance on a speeded dishwashing task, nobody is going to take me very seriously. It doesn't even matter how the results of my experiment turn out: there is no arrangement of numbers in a table, no p-value I could compute from my data, that could possibly turn my chosen experimental manipulation into a sensible proxy for social status. And the same goes for the rather questionable use of speeded dishwashing performance as a proxy for dishonesty.

The absurdity of the preceding example exposes a critical assumption that often goes unnoticed: for an empirical result to have bearing on a verbal assertion, the measured variables must be suitable operationalizations of the verbal constructs of interest, and the relationships between the measured variables must parallel those implied by the logical structure of the verbal statements. Equating the broad construct of honesty with a measure of speeded dishwashing is so obviously nonsensical that we immediately reject such a move out of hand. What may be less obvious is that exactly the same logic implicitly applies in virtually every case where researchers lean on statistical quantities to justify their verbal claims. Statistics is not, as many psychologists appear to view it, a rote, mechanical procedure for turning data into conclusions. It is better understood as a parallel, and more precise, *language* in which one can express one's hypotheses or beliefs. Every statistical model is a description of some real or hypothetical state of affairs in the world. If its mathematical expression fails to capture roughly the same state of affairs as the verbal hypothesis the researcher began with, then the statistical quantities produced by the model cannot serve as an adequate proxy for the verbal statements—and consequently, the former cannot be taken as support for the latter.

Viewed from this perspective, the key question is how closely the verbal and quantitative expressions of one's hypothesis align with each other. When a researcher verbally expresses a particular proposition—be it a theoretically-informed hypothesis or a purely descriptive characterization of some data—she is implicitly defining a set of hypothetical measurements (or *admissible observations*; Brennan, 1992) that would have to come out a certain way in order for the statement to be corroborated. If the researcher subsequently asserts that a particular statistical procedure provides a suitable operationalization of the verbal statement, she is making the tacit but critical assumption that the universe of hypothetical measurements implicitly defined by the chosen statistical procedure, in concert with the experimental design and measurement model, is well aligned with the one implicitly defined by the qualitative statement. Should a discrepancy between the two be discovered, the researcher will then face a choice between (a) working to resolve the discrepancy in some way (i.e., by modifying either the verbal statement or the quantitative procedure(s) meant to provide an operational parallel); or (b) giving up on the link between the two and accepting that the statistical procedure does not inform the verbal expression in a meaningful way.

The next few sections explore this relationship with respect to the most widely used class of statistical model in psychology—linear mixed models containing fixed and random effects (though the broader conceptual points I will make apply to *any* use of statistical quantities to evaluate verbal claims). The exploration begins with an examination of the standard random-subjects model—a mainstay of group-level inferences in psychology—and then progressively considers addi-

tional sources of variability whose existence is implied by most verbal inferences in psychology, but that the standard model fails to appropriately capture. The revealed picture is that an unknown but clearly very large fraction of statistical hypotheses described in psychology studies cannot plausibly be considered reasonable operationalizations of the verbal hypotheses they are meant to inform. (While I deliberately restrict the focus of my discussion to the field of psychology, with which I am most familiar, I expect that researchers in various social and biomedical disciplines will find that the core arguments I lay out generalize well to many other areas.)

Fixed vs. random effects

Let us begin with a scenario that will be familiar to many psychologists. Suppose we administer a cognitive task—say, the color-word Stroop (MacLeod, 1991; Stroop, 1935)—to a group of participants (the reader is free to mentally substitute almost any other experimental psychology task into the example). Each participant is presented with a series of trials, half in a congruent condition and half in an incongruent condition. We are tasked with fitting a statistical model to estimate the canonical Stroop effect—i.e., the increase in reaction time (RT) observed when participants are presented with incongruent color-word information relative to congruent color-word information.

A naive, though almost always inappropriate, model might be the following:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 X_{ij} + e_{ij} \\ e_{ij} &\sim \mathcal{N}(0, \sigma_e^2) \end{aligned} \tag{1}$$

In this linear regression, y_{ij} denotes the i'th subject's response on trial j, X_{ij} indexes the experimental condition (congruent or incongruent) of subject i's j'th trial, β_0 is an intercept, β_1 is the effect of congruency, and e_{ij} captures the errors, which are assumed to be normally distributed.

What is wrong with this model? Well, one rather serious problem is that the model blatantly ignores sources of variance in the data that we know on theoretical grounds must exist. Notably, because the model includes only a single intercept parameter and a single slope parameter across all subjects and trials, it predicts exactly the same reaction time value for all trials in each condition, no matter which subject a given trial is drawn from. Such an assumption is clearly untenable: it's absurd to suppose that the only source of trial-to-trial RT variability within experimental conditions is random error. We know full well that people differ systematically from one another in performance on the Stroop task (and for that matter, on virtually every other cognitive task). Any model that fails to acknowledge this important source of variability is clearly omitting an important feature of the world as we understand it.

From a statistical standpoint, the model's failure to explicitly acknowledge between-subject variability has several deleterious consequences for our Stroop estimate. The most salient one, given psychologists' predilection towards dichotomous conclusions (e.g., whether or not an effect is statistically significant), is that the estimated uncertainty surrounding the parameter estimates of interest will tend to be biased—typically downwards (i.e., in our Stroop example, the standard error of the Stroop effect will usually be underestimated)[2]. The reason is that, lacking any concept of a "person", our model cannot help but assume that any new set of trials—no matter who they come from—must have been generated by exactly the same set of processes that gave rise to the trials the model has previously seen. Consequently, the model cannot adjust the uncertainty around the point estimate to account for variability between subjects, and will usually produce an overly optimistic estimate of its own performance when applied to new subjects whose data-generating process is at least somewhat different from the process that generated the data the model was trained on.

The deleterious impact of using Model (1) to es-

[2]The precise effect of failing to include random factors depends on a number of considerations, including the amount of variance between versus within the random effects, the covariance with other variables, and the effective sample sizes of different factors. But in most real-world settings, the inclusion of random effects will lead to (often much) larger uncertainty estimates and smaller inferential test statistics.

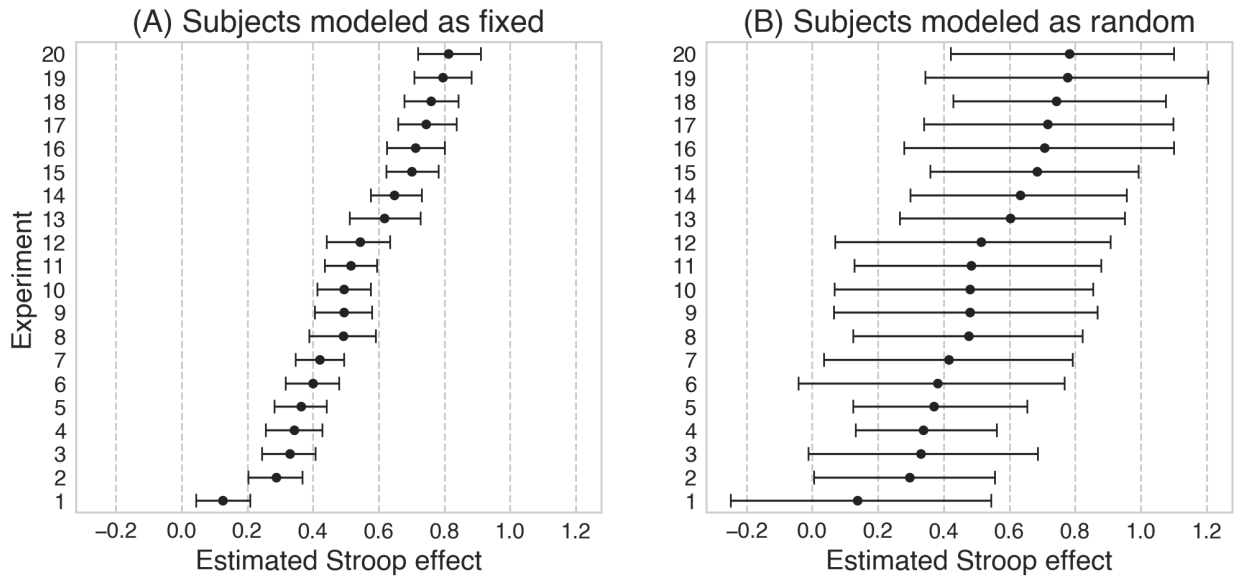

Figure 1: Consequences of mismatch between model specification and generalization intention. Each row represents a simulated Stroop experiment with $n = 20$ new subjects randomly drawn from the same global population (the ground truth for all parameters is constant over all experiments). Bars display the estimated Bayesian 95% highest posterior density (HPD) intervals for the (fixed) condition effect of interest in each experiment. Experiments are ordered by the magnitude of the point estimate for visual clarity. (A) The fixed-effects model specification in Eq. (1) does not account for random subject sampling, and consequently underestimates the uncertainty associated with the effect of interest. (B) The random-effects specification in Eq. (2) takes subject sampling into account, and produces appropriately calibrated uncertainty estimates.

timate the Stroop effect when generalization to new subjects is intended is illustrated in Figure 1A. The figure shows the results of a simulation of 20 random Stroop experiments, each with 20 participants and 200 trials per participant (100 in each condition). The true population effect—common to all 20 experiments—is assumed to be large. As expected, fitting the simulated data with the fixed-effects model specification in Eq. (1) produces an unreasonably narrow estimate of the uncertainty surrounding the point estimates—observe that, for any given experiment, most of the estimates from the other experiments are well outside the 95% highest posterior density (HPD) interval. Researchers who attempt to naively generalize the estimates obtained using the fixed-effects model to data from new subjects are thus setting themselves up for an unpleasant surprise.

How might we adjust Model (1) to account for the additional between-subject variance in the data introduced by the stochastic sampling of individuals from a broader population? One standard approach is to fit a model like the following:

$$
\begin{aligned}
y_{ij} &= \beta_0 + \beta_1 X_{ij} + u_{0i} + u_{1i} X_{ij} + e_{ij} \qquad (2)\\
u_{0i} &\sim \mathcal{N}(0, \sigma^2_{u_0})\\
u_{1i} &\sim \mathcal{N}(0, \sigma^2_{u_1})\\
e_{ij} &\sim \mathcal{N}(0, \sigma^2_{e})
\end{aligned}
$$

Here, we expand Model (1) to include two new terms: u_0 and u_1, which respectively reflect a set of intercepts and a set of slopes—one pair of terms per subject[3]. The u parameters are assumed (like the error e) to

[3]To keep things simple, I ignore the question of how one ought to decide whether or not to include both random slopes

follow a normal distribution centered at zero, with the size of the variance components (i.e., the variances of the groups of random effects) $\sigma^2_{u_k}$ estimated from the data.

Conventionally, the u parameters in Model (2) are referred to as *random* (or sometimes, *varying* or *stochastic*) effects, as distinct from the *fixed* effects captured by the β terms[4]. There are several ways to conceptualize the distinction between random and fixed effects (Gelman & Hill, 2006), but since our focus here is on generalizability, we will define them this way: fixed effects are used to model variables that must remain constant in order for the model to preserve its meaning across replication studies; random effects are used to model indicator variables that are assumed to be stochastically sampled from some underlying population and can vary across replications without meaningfully altering the research question. In the context of our Stroop example, we can say that the estimated Stroop effect β_1 is a fixed effect, because if we were to run another experiment using a different manipulation (say, a Sternberg memory task), we could no longer reasonably speak of the second experiment being a replication of the first. By contrast, psychologists almost invariably think of experimental subjects as a random factor: we are rarely interested in the particular people we happen to have in a given sample, and it would be deeply problematic if two Stroop experiments that differed only in their use of different subjects (randomly sampled from the same population) had to be treated as if they provided estimates of two conceptually distinct Stroop effects.[5]

Note that while the model specified in (2) is a substantial improvement over the one specified in (1) if our goal is to draw inferences over populations of subjects, it is not in any meaningful sense the "correct" model. Model (2) is clearly still an extremely simplistic approximation of the true generative processes underlying Stroop data, and, even within the confines of purely linear models, there are many ways in which we could further elaborate on (2) to account for other potentially important sources of variance (e.g., practice or fatigue effects, stimulus-specific effects, measured individual differences in cognitive ability, etc.). Moreover, the fact that Model (2) supports inference over *some* broader population of subjects provides no guarantee that that population is one the researcher is interested in. If, for example, our subjects are all sampled from a Western undergraduate population aged 18 - 23, then Model (2) may license generalization of the results to other undergraduates like the ones we studied, but we would be leaning very heavily on auxiliary assumptions not explicitly included in our model if we were to generalize our conclusions to the broader population of human beings.

In highlighting the difference between models (1) and (2), I simply wish to draw attention to two important and interrelated points. First, inferences about model parameters are always tied to a particular model specification. A claim like "there is a statistically significant effect of Stroop condition" is not a claim about the world per se; rather, it is a claim about the degree to which a specific model accurately describes the world under certain theoretical assumptions and measurement conditions. Strictly speaking, a statistically significant effect of Stroop condition in Model (1) tells us only that the data we observe would be unlikely to occur under a null model that considers all trials to be completely exchangeable. By contrast, a statistically significant effect in Model (2) for what nominally appears to be the "same" β_1 parameter would have a different (and somewhat stronger) interpretation, as we are now entitled to conclude that the data we observe would be unlikely if there were no effect (on average) at the level of individuals randomly

and random intercepts (for discussion, see (Barr, Levy, Scheepers, & Tily, 2013; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). The goal here is simply to elucidate the distinction between fixed and random effects.

[4]Note that in econometrics, the term *fixed effect* has a narrower meaning, and refers specifically to a group mean parameter (rather than just any predictor variable) modeled as non-random.

[5]A reasonable argument could be made that since no experimental context is ever *exactly* the same across two measurement occasions, in a technical sense, no design factor is ever truly fixed. Readers who are sympathetic to such an argument (as I also am) should remember Box's dictum that "all models are false, but some are useful", and are invited to construe the choice between fixed and random effects as a purely pragmatic one that amounts to deciding which of two idealizations better approximates reality.

drawn from some population.

Second, the validity of an inference depends not just on the model itself, but also on the analyst's (typically implicit) intentions. As discussed earlier, to support valid inference, a statistical model must adequately represent the universe of observations the analyst intends to implicitly generalize over when drawing qualitative conclusions. In our example above, what makes Model (1) a bad model is not the model specification alone, but the fact that the specification aligns poorly with the universe of observations that researchers typically care about. In typical practice, researchers intend their conclusions to apply to entire populations of subjects, and not just to the specific individuals who happened to walk through the laboratory door when the study was run. Critically, then, *it is the mismatch between our generalization intention and the model specification that introduces an inflated risk of inferential error, and not the model specification alone.* The reason we model subjects as random effects is not that such a practice is objectively better, but rather, that this specification more closely aligns the meaning of the quantitative inference with the meaning of the qualitative hypothesis we're interested in evaluating (for discussion, see Cornfield & Tukey, 1956).

Beyond random subjects

The discussion in the preceding section may seem superfluous to some readers given that, in practice, psychologists almost universally already model subject as a random factor in their analyses. Importantly, however, there is nothing special about subjects. In principle, what goes for subjects also holds for any other factor of an experimental or observational study whose levels the authors intend to generalize over. The reason that we routinely inject extra uncertainty into our models in order to account for between-subject variability is that we want our conclusions to apply to a broader population of individuals, and not just to the specific people we randomly sampled. But the same logic also applies to a large number of other factors that we do not routinely model as random effects—stimuli, experimenters, research sites, and so on. Indeed, as Brunswik long ago observed, "...proper sampling of situations and problems may in the end be more important than proper sampling of subjects, considering the fact that individuals are probably on the whole much more alike than are situations among one another" (Brunswik, 1947, p. 179). As we shall see, extending the random-effects treatment to other factors besides subjects has momentous implications for the interpretation of a vast array of published findings in psychology.

The stimulus-as-fixed-effect fallacy

A paradigmatic example of a design factor that psychologists almost universally—and inappropriately—model as a fixed rather than random factor is experimental stimuli. The tendency to ignore stimulus sampling variability has been discussed in the literature for over 50 years (Baayen, Davidson, & Bates, 2008; Clark, 1973; Coleman, 1964; Judd, Westfall, & Kenny, 2012), and was influentially dubbed the *fixed-effect fallacy* by (Clark, 1973). Unfortunately, outside of a few domains such as psycholinguistics, it remains rare to see psychologists model stimuli as random effects—despite the fact that most inferences researchers draw are clearly meant to generalize over populations of stimuli. The net result is that, strictly speaking, the inferences routinely drawn throughout much of psychology can only be said to apply to a specific—and usually small—set of stimuli. Generalization to the broader class of stimuli like the ones used is not licensed.

It is difficult to overstate how detrimental an impact the stimulus-as-fixed-effect fallacy has had— and continues to have—in psychology. Empirical studies in domains ranging from social psychology to functional MRI have demonstrated that test statistic inflation of up to 300% is not uncommon, and that, under realistic assumptions, false positive rates in many studies could easily exceed 60% (Judd et al., 2012; Westfall, Nichols, & Yarkoni, 2016; Wolsiefer, Westfall, & Judd, 2017). In cases where subject sample sizes are very large, stimulus samples are very small, and stimulus variance is large, the false positive rate theoretically approaches 100%.

The clear implication of such findings is that many literatures within psychology are likely to be populated by studies that have spuriously misattributed statistically significant effects to fixed effects of interest when they should actually be attributed to stochastic variation in uninteresting stimulus properties. Moreover, given that different sets of stimuli are liable to produce effects in opposite directions (e.g., when randomly sampling 20 nouns and 20 verbs, some samples will show a statistically significant noun > verb effect, while others will show the converse), it is not hard to see how one could easily end up with entire literatures full of "mixed results" that seem statistically robust in individual studies, yet cannot be consistently replicated across studies.

Generalizing the generalizability problem

The stimulus-as-fixed-effect fallacy is but one special case of a general tradeoff between precision of estimation and breadth of generalization. Each additional random factor one adds to a model licenses generalization over a corresponding population of potential measurements, expanding the scope of inference beyond only those measurements that were actually obtained. However, adding random factors to one's model also typically increases the uncertainty with which the fixed effects of interest are estimated. The fact that most psychologists have traditionally modeled only subject as a random factor—and have largely ignored the variance introduced by stimulus sampling—is probably best understood as an accident of history (or, more charitably perhaps, of technological limitations, as the software and computing resources required to fit such models were hard to come by until fairly recently).

Unfortunately, just as the generalizability problem doesn't begin and end with subjects, it also doesn't end with subjects and stimuli. Exactly the same considerations apply to all other aspects of one's experimental design or procedure that could in principle be varied without substantively changing the research question. Common design factors that researchers hardly ever vary, yet almost invariably intend to generalize over, include experimental task, between-subject instructional manipulation, research site, experimenter (or, in clinical studies, therapist; e.g., Crits-Christoph & Mintz, 1991), instructions, laboratory testing conditions (e.g., Crabbe, Wahlsten, & Dudek, 1999; Wahlsten et al., 2003), weather, and so on and so forth effectively ad infinitum.

Naturally, the degree to which each such factor matters will vary widely across domain and research question. I'm not suggesting that most statistical inferences in psychology are invalidated by researchers' failure to explicitly model what their participants ate for breakfast three days prior to participating in a study. Collectively, however, unmodeled factors almost always contribute substantial variance to the outcome variable. Failing to model such factors appropriately (or at all) means that a researcher will end up either (a) running studies with substantially higher-than-nominal false positive rates, or (b) drawing inferences that technically apply only to very narrow, and usually uninteresting, slices of the universe the researcher claims to be interested in.

Case study: Verbal overshadowing

To illustrate the problem, it may help to consider an example. Alogna and colleagues (2014) conducted a large-scale "registered replication report" (RRR; Simons, Holcombe, & Spellman, 2014) involving 31 sites and over 2,000 participants. The study sought to replicate an influential experiment by Schooler and Engstler-Schooler (1990) in which the original authors showed that participants who were asked to verbally describe the appearance of a perpetrator caught committing a crime on video showed poorer recognition of the perpetrator following a delay than did participants assigned to a control task (naming as many countries and capitals as they could). Schooler & Engstler-Schooler (1990) dubbed this the *verbal overshadowing effect*. In both the original and replication experiments, only a single video, containing a single perpetrator, was presented at encoding, and only a single set of foil items was used at test. Alogna et al. successfully replicated the original result in one of two tested conditions, and concluded that their findings revealed "a robust verbal overshadowing effect" in that condition.

Let us assume for the sake of argument that there is a genuine and robust causal relationship between the manipulation and outcome employed in the Alogna et al study. I submit that there would still be essentially no support for the authors' assertion that they found a "robust" verbal overshadowing effect, because the experimental design and statistical model used in the study simply cannot support such a generalization. The strict conclusion we are entitled to draw, given the limitations of the experimental design inherited from Schooler and Engstler-Schooler (1990), is that there is at least one particular video containing one particular face that, when followed by one particular lineup of faces, is more difficult for participants to identify if they previously verbally described the appearance of the target face than if they were asked to name countries and capitals. This narrow conclusion does not preclude the possibility that the observed effect is specific to this one particular stimulus, and that many other potential stimuli the authors could have used would have eliminated or even reversed the observed effect. (In later sections, I demonstrate that the latter conclusion is statistically bound to be true given even very conservative background assumptions about the operationalization, and also that one can argue from first principles—i.e., *without any data at all*—that there must be *many* stimuli that show a so-called verbal overshadowing effect.)

Of course, stimulus sampling is not the only unmodeled source of variability we need to worry about. We also need to consider any number of other plausible sources of variability: research site, task operationalization (e.g., timing parameters, modality of stimuli or responses), instructions, and so on. On any reasonable interpretation of the *construct* of verbal overshadowing, the corresponding universe of intended generalization should clearly also include most of the operationalizations that would result from randomly sampling various combinations of these factors (e.g., one would expect it to still count as verbal overshadowing if Alogna et al. had used live actors to enact the crime scene, instead of showing a video)[6]. Once we accept this assumption, however, the critical question researchers should immediately ask themselves is: are there other psychological processes besides verbal overshadowing that could plausibly be influenced by random variation in any of these uninteresting factors, *independently of the hypothesized psychological processes of interest*? A moment or two of consideration should suffice to convince one that the answer is a resounding yes. It is not hard to think of dozens of explanations unrelated to verbal overshadowing that could explain the causal effect of a given manipulation on a given outcome in any single operationalization[7].

This verbal overshadowing example is by no means unusual. The same concerns apply equally to the broader psychology literature containing tens or hundreds of thousands of studies that routinely adopt similar practices. In most of psychology, it is standard operating procedure for researchers employing just one experimental task, between-subject manipulation, experimenters, testing room, research site, etc., to behave as though an extremely narrow operationalization is an acceptable proxy for a much broader universe of admissible observations. It is instructive—and somewhat fascinating from a sociological perspective—to observe that while no psychometrician worth their salt would ever recommend a default strategy of measuring complex psychological constructs using a single unvalidated item, the majority of psychology studies

[6]That even small differences in such factors can have large impacts on the outcome is clear from the Alogna et al. (2014) study itself: due to an error in the timing of different components of the procedure, Alogna et al actually conducted *two* large replication studies. They observed a markedly stronger effect when the experimental task was delayed by 20 minutes than when it immediately followed the video.

[7]For example, perhaps participants in Alogna et al's experimental condition felt greater pressure to produce the correct answer (having previously spent several minutes describing their perceptions), and it was the stress rather than the treatment *per se* that resulted in poorer performance. Or, perhaps the effect had nothing at all to do with the treatment condition, and instead reflected a poor choice of control condition (say, because naming countries and capitals incidentally activates helpful memory consolidation processes). And so on and so forth. (A skeptic might object that each such explanation is individually not as plausible as the verbal overshadowing account, but this misses the point: safely generalizing the results of the narrow Schooler & Engstler-Schooler (1990) design to the broad construct of verbal overshadowing implies that one can rule out the influence of *all* other confounds in the aggregate—and reality is not under any obligation to only manifest sparse causal relationships that researchers find intuitive!)

do precisely that with respect to *multiple* key design factors. The modal approach is to stop at a perfunctory demonstration of face validity—that is, to conclude that if a particular operationalization *seems* like it has something to do with the construct of interest, then it is an acceptable stand-in for that construct. Any measurement-level findings are then uncritically generalized to the construct level, leading researchers to conclude that they've learned something useful about broader phenomena like verbal overshadowing, working memory, ego depletion, etc., when in fact such sweeping generalizations typically obtain little support from the reported empirical studies.

Unmeasured factors

In an ideal world, generalization failures like those described above could be addressed primarily via statistical procedures—e.g., by adding new random effects to models. In the real world, this strategy is a non-starter: in most studies, the vast majority of factors that researchers intend to implicitly generalize over don't actually observably vary in the data, and therefore can't be accounted for using traditional mixed-effects models. Unfortunately, the fact that one has failed to introduce or measure variation in one or more factors doesn't mean those factors can be safely ignored. Any time one samples design elements into one's study from a broader population of possible candidates, one introduces sampling error that is likely to influence the outcome of the study to some unknown degree.

Suppose we generalize our earlier Model (2) to include all kinds of random design factors that we have no way of directly measuring:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 X_{1ij} + u_{0ij} + u_{1ij} + \ldots + u_{kij} + e_{ij} \\ u_{kij} &\sim \mathcal{N}(0, \sigma^2_{u_k}) \\ e_{ij} &\sim \mathcal{N}(0, \sigma^2_e) \end{aligned} \tag{3}$$

Here, $u_0 \ldots u_k$ are placeholders for all of the variance components that we implicitly consider part of the universe of admissible observations, but that we have no way of measuring or estimating in our study. It should be apparent that our earlier Model (2) is just a special case of (3) where the vast majority of the u_k and $\sigma^2_{u_k}$ terms are fixed to 0. That is—and this is arguably the most important point in this paper—the conventional "random effects" model (where in actuality only subjects are modeled as random effects) *assumes exactly zero effect of site, experimenter, stimuli, task, instructions, and every other factor except subject*—even though in most cases it's safe to assume that such effects exist and are non-trivial, and even though authors almost invariably start behaving as if their statistical models did in fact account for such effects as soon as they reach the Discussion section.

Estimating the impact

We do not have to take the urgency of the above exhortation on faith. While it's true that we can't directly estimate the population magnitude of variance components that showed no observable variation in our sample, we can still simulate their effects under different assumptions. Doing so allows us to demonstrate empirically that even modest assumptions about the magnitude of unmeasured variance components may be sufficient to completely undermine many conventional inferences about fixed effects of interest.

To illustrate, let's return to Alogna et al (2014)'s verbal overshadowing RRR. Recall that the dataset included data from over 2,000 subjects sampled at 31 different sites, but used exactly the same experimental protocol (including the same single stimulus sequence) at all sites. Since most of the data are publicly available, we can fit a mixed-effects model to try and replicate the reported finding of a "robust verbal overshadowing effect". Both the dataset and the statistical model used here differ somewhat from the ones in Alogna et al. (2014)[8], but the differences are

[8]The model differs in that I fit a single mixed-effects linear probability model with random intercepts and slopes for sites, whereas Alogna et al. first computed the mean difference in response accuracy between conditions for each site, and then performed a random-effects meta-analysis (note that a logistic regression model would be appropriate here given the binary outcome, but I opted for the linear model for the sake of consistency with Alogna et al. and simplicity of presentation). The data differ because (a) some sites' datasets were not publicly available, (b) I made no attempt to adhere

immaterial for our purposes. As Figure 2 illustrates (top row, labeled $\sigma^2_{unmeasured} = 0$), we can readily replicate the key finding from Alogna et al. (2014): participants assigned to the experimental condition were more likely to misidentify the perpetrator seen in the original video.

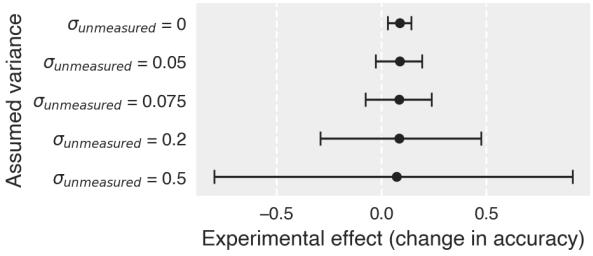

Figure 2: Effects of unmeasured variance components on the putative "verbal overshadowing" effect. Error bars display the estimated Bayesian 95% highest posterior density (HPD) intervals for the experimental effect reported in Alogna et al. (2014). Positive estimates indicate better performance in the control condition than in the experimental condition. Each row represents the estimate from the model specified in Eq. (4), with only the size of $\sigma^2_{unmeasured}$ (corresponding to $\sigma^2_{u_2}$ in Eq. (4)) varying as indicated. This parameter represents the assumed contribution of all variance components that are unmeasured in the experiment, but fall within the universe of intended generalization conceptually. The top row ($\sigma^2_{u_2} = 0$) can be interpreted as a conventional model analogous to the one reported in Alogna et al (2014)—i.e., it assumes that no unmeasured sources have any impact on the putative verbal overshadowing effect.

We now ask the following question: how would the key result depicted in the top row of Fig. 2 change if we *knew* the size of the variance component associated with random stimulus sampling? This question cannot be readily answered using classical inferential procedures (because there's only a single stimulus in the dataset, so the variance component is non-identifiable), but is trivial to address using a Bayesian estimation framework. Specifically, we fit the following model:

$$
\begin{aligned}
y_{ps} &= \beta_0 + \beta_1 X_{ps} + u_{0s} + u_{1s} X_{ps} + u_2 X_{ps} + e_{ps} \\
u_{0s} &\sim \mathcal{N}(0, \sigma^2_{u_0}) \\
u_{1s} &\sim \mathcal{N}(0, \sigma^2_{u_1}) \\
u_2 &\sim \mathcal{N}(0, \sigma^2_{u_2}) \\
e_{ps} &\sim \mathcal{N}(0, \sigma^2_e)
\end{aligned}
\tag{4}
$$

Here, p indexes participants, s indexes sites, X_{ps} indexes the experimental condition assigned to participant p at site s, the β terms encode the fixed intercept and condition slope, and the u terms encode the random effects (site-specific intercepts u_0, site-specific slopes u_1, and the stimulus effect u_2). The novel feature of this model is the inclusion of u_2, which would ordinarily reflect the variance in outcome associated with random stimulus sampling, but is constant in our dataset (because there's only a single stimulus).

Unlike the other parameters, we cannot estimate u_2 from the data. Instead, we fix its prior during estimation, by setting $\sigma^2_{u_2}$ to a specific value. While the posterior estimate of u_2 is then necessarily identical to its prior (because the prior makes no contact with the data), and so is itself of no interest, the inclusion of the prior has the incidental effect of (appropriately) increasing the estimation uncertainty around the fixed effect(s) of interest. Conceptually, one can think of the added prior as a way of quantitatively representing our uncertainty about whether any experimental effect we observe should really be attributed to verbal overshadowing *per se*, as opposed to irrelevant properties of the specific stimulus we happened to randomly sample into our experiment. By varying the amount of variance injected in this way, we can study the conditions under which the conclusions obtained from the "standard" model (i.e., one that assumes zero effect of stimuli) would or wouldn't hold.

As it turns out, injecting even a small amount of stimulus sampling variance to the model has momentous downstream effects. If we very conservatively set $\sigma^2_{u_2}$ to 0.05, the resulting posterior distribution for the

closely to the reported preprocessing procedures (e.g., inclusion/exclusion criteria), and (c) I used only the data from the (more successful) second RRR reported in the paper. All data and code used in the analyses reported here are available at https://github.com/tyarkoni/generalizability.

condition effect expands to include negative values within the 95% HPD (Fig. 2). For perspective, 0.05 is considerably lower than the between-site variance estimated from these data ($\sigma^2_{u_1} = 0.075$)—and it's quite unlikely that there would be less variation between different stimuli at a given site than between different sites for the same stimulus (as reviewed above, in most domains where stimulus effects have been quantitatively estimated, they tend to be large). Thus, even under very conservative assumptions about how much variance might be associated with stimulus sampling, there is little basis for concluding that there is a *general* verbal overshadowing effect. To draw Alogna et al's conclusion that there is a "robust" verbal overshadowing effect, one must effectively equate the construct of verbal overshadowing with *almost exactly* the operationalization tested by Alogna et al. (and Schooler & Schooler-Engstler before that), down to the same single video.

Of course, stimulus variance isn't the only missing variance component we ought to worry about. As Eq. (3) underscores, many other components are likely to contribute non-negligible variance to outcomes within our universe of intended generalization. We could attempt to list these components individually and rationally estimate their plausible magnitudes if we like, but an alternative route is to invent an omnibus parameter, $\sigma^2_{unmeasured}$, that subsumes *all* of the unmeasured variance components we expect to systematically influence the condition estimate β_1. Then we can repeat our estimation of the model in Eq. (4) with larger values of $\sigma^2_{u_2}$ (for the sake of convenience, I treat $\sigma^2_{u_2}$ and $\sigma^2_{unmeasured}$ interchangeably, as the difference is only that the latter is larger than the former).

For example, suppose we assume that the hypothetical aggregate influence of all the unmodeled variance components roughly equals the residual within-site variance estimated in our data (i.e., $\sigma^2_{unmeasured} = \sigma^2_{e_{ps}} \approx 0.5$). This is arguably still fairly conservative when one considers that the aggregate $\sigma^2_{unmeasured}$ now includes not only stimulus sampling effects, but also the effects of differences in task operationalization, instructions, etc. In effect, we are assuming that the net contribution of all of the uninteresting factors that vary across the entire universe of observations we consider "verbal overshadowing" is no bigger than the residual error we observe for this one particular operationalization. Yet fixing $\sigma^2_{unmeasured}$ to 0.5 renders our estimate of the experimental effect essentially worthless: the 95% HPD interval for the putative verbal overshadowing effect now spans values between -0.8 and 0.91—almost the full range of possible values! The upshot is that, even given very conservative background assumptions, the massive Alogna et al. study—an initiative that drew on the efforts of dozens of researchers around the world—does not tell us much about the general phenomenon of verbal overshadowing. Under more realistic assumptions, it tells us essentially nothing. The best we can say, if we are feeling optimistic, is that it might tell us something about one particular *operationalization* of verbal overshadowing[9].

The rather disturbing implication of all this is that, in any research area where one expects the aggregate contribution of the missing σ^2_u terms to be large—i.e., anywhere that "contextual sensitivity" (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016) is high—the inferential statistics generated from models like (2) will often underestimate the true uncertainty surrounding the parameter estimates to such a degree as to make an outright mockery of the effort to learn something from the data using conventional inferential tests. Recall that the nominal reason we care about whether subjects are modeled as fixed or random effects is that the latter specification allows us to generalize to theoretically exchangeable observations (e.g., new subjects sampled from the same population), whereas the former does not. In practice, however, the majority of psychologists have no compunction about verbally generalizing their results not only to previously unseen subjects, but also to all kinds of other factors that have not explicitly been modeled—to new stimuli, experimenters, research sites, and so

[9] We should probably be cautious in drawing even this narrow conclusion, however, because the experimental procedure in question could very well be producing the observed effect due to idiosyncratic and uninteresting properties, and not because it induces verbal overshadowing *per se*.

on.

Under such circumstances, it's unclear why anyone should really care about the inferential statistics psychologists report in most papers, seeing as those statistics bear only the most tenuous of connections to authors' sweeping verbal conclusions. Why take pains to ensure that subjects are modeled in a way that affords generalization beyond the observed sample—as nearly all psychologists reflexively do—while raising no objection whatsoever when researchers freely generalize their conclusions across all manner of variables that weren't explicitly included in the model at all? Why not simply model *all* experimental factors, including subjects, as fixed effects—a procedure that would, in most circumstances, substantially increase the probability of producing the sub-.05 p-values psychologists so dearly crave? Given that we've already resolved to run roughshod over the relationship between our verbal theories and their corresponding quantitative specifications, why should it matter if we sacrifice the sole remaining sliver of generality afforded by our conventional "random effects" models on the altar of the Biggest Possible Test Statistic?

It's hard to think of a better name for this kind of behavior than what Feynman famously dubbed *cargo cult science* (Feynman, 1974)—an obsessive concern with the superficial form of a scientific activity rather than its substantive empirical and logical content. Psychologists are trained to believe that their ability to draw meaningful inferences depends to a large extent on the production of certain statistical quantities (e.g., p-values below .05, BFs above 10, etc.), so they go to great effort to produce such quantities. That these highly contextualized numbers typically have little to do with the broad verbal theories and hypotheses that researchers hold in their heads, and take themselves to be testing, does not seem to trouble most researchers much. The important thing, it appears, is that *the numbers have the right form.*

A crisis of replicability or of generalizability?

It is worth situating the above concerns within the broader ongoing "replication crisis" in psychology and other sciences (Lilienfeld, 2017; Pashler & Wagenmakers, 2012; Shrout & Rodgers, 2018). My perspective on the replicability crisis broadly accords with other commentators who have argued that the crisis is real and serious, in the sense that there is irrefutable evidence that questionable research practices (Gelman & Loken, 2013; John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011) and strong selection pressures (Francis, 2012; Kühberger, Fritz, & Scherndl, 2014; Smaldino & McElreath, 2016) have led to the publication of a large proportion of spurious or inflated findings that are unlikely to replicate (J. Ioannidis, 2008; J. P. A. Ioannidis, 2005; Yarkoni, 2009). Accordingly, I think the ongoing shift towards practices such as preregistration, reporting checklists, data sharing, etc. is a welcome development that will undoubtedly help improve the reproducibility and replicability of psychology findings.

At the same time, the current focus on reproducibility and replicability risks distracting us from more important, and logically antecedent, concerns about generalizability. The root problem is that when the manifestation of a phenomenon is highly variable across potential measurement contexts, it simply does not matter very much whether any single realization is replicable or not (cf. Gelman, 2015, 2018). Ongoing efforts to ensure the superficial reproducibility and replicability of effects—i.e., the ability to obtain a similar-looking set of numbers from independent studies—are presently driving researchers in psychology and other fields to expend enormous resources on studies that are likely to have very little informational value even in cases where results can be consistently replicated. This is arguably clearest in the case of large-scale "registered replication reports" (RRRs) that have harnessed the enormous collective efforts of dozens of labs (e.g., Acosta et al., 2016; Alogna et al., 2014; Cheung et al., 2016; Eerland et al., 2016)—only to waste that collective energy on direct replications of a handful of poorly-validated experimental paradigms.

While there is no denying that large, collaborative efforts could have enormous potential benefits (and there are currently a number of promising initiatives, e.g., the Psychological Science Accelerator (Moshontz et al., 2018) and ManyBabies Consortium (Bergelson et al., 2017)), realizing these benefits will require a willingness to eschew direct replication in cases where the experimental design of the to-be-replicated study is fundamentally uninformative. Researchers must be willing to look critically at previous studies and flatly reject—on logical and statistical, rather than empirical, grounds—assertions that were never supported by the data in the first place, even under the most charitable methodological assumptions. A recognition memory task that uses just one video, one target face, and one set of foils simply cannot provide a meaningful test of a broad construct like verbal overshadowing, and it does a disservice to the field to direct considerable resources to the replication of such work. The appropriate response to a study like Schooler & Engstler-Schooler (1990) is to point out that the very narrow findings the authors reported did not—and indeed, could not, no matter how the data came out—actually support the authors' sweeping claims. Consequently, the study does not deserve any follow-up until such time as its authors can provide more compelling evidence that a phenomenon of any meaningful generality is being observed.

The same concern applies to many other active statistical and methodological debates. Is it better to use a frequentist or a Bayesian framework for hypothesis testing (Kruschke & Liddell, 2017; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers, 2007)? Should we move the conventional threshold for statistical significance from .05 to .005 (Benjamin et al., 2018; Lakens et al., 2018; McShane, Gal, Gelman, Robert, & Tackett, 2019)? A lot of ink continues to be spilled over such issues, yet in any research area where effects are highly variable (i.e., in most of psychology), the net contribution of such methodological and analytical choices to overall inferential uncertainty is likely to be dwarfed by the bias introduced by implicitly generalizing over unmodeled sources of variance in the data. There is little point in debating the merits of a statistical significance cut-off of .005 rather than .05 in a world where even a trivial change in an unmodeled variable—e.g., a random choice between two nominally equivalent cognitive tasks, or the use of a slightly different stimulus sample—can routinely take one from $p = .5$ to $p = .0005$ or vice versa (cf. Crits-Christoph & Mintz, 1991; Westfall et al., 2016; Wolsiefer et al., 2017). Yet this root problem continues to go largely ignored in favor of efforts to treat its downstream symptoms. It appears that, faced with the difficulty of stating what the complex, multicausal effects we psychologists routinely deal in actually *mean*, we have collectively elected to instead pursue superficially precise answers to questions none of us really care much about.

To be clear, my suggestion is not that researchers should stop caring about methodological or statistical problems that presently limit reproducibility and replicability. Such considerations are undeniably important. My argument, rather, is that these considerations should be reserved for situations where the verbal conclusions drawn by researchers demonstrably bear some non-trivial connection to the reported quantitative analyses. The mere fact that a previous study has had a large influence on the literature is not a sufficient reason to expend additional resources on replication. On the contrary, the recent movement to replicate influential studies using more robust methods risks making the situation worse, because in cases where such efforts superficially "succeed" (in the sense that they obtain a statistical result congruent with the original), researchers then often draw the incorrect conclusion that the new data corroborate the original claim (e.g., Alogna et al., 2014)—when in fact the original claim was never supported by the data in the first place. A more appropriate course of action in cases where there are questions about the internal coherence and/or generalizability of a finding is to first focus a critical eye on the experimental design, measurement approach, and model specification. Only if a careful review suggests that these elements support the claims made by a study's authors should researchers begin to consider conducting a replication.

Where to from here?

A direct implication of the arguments laid out above is that a huge proportion of the quantitative inferences drawn in the published psychology literature are so weak as to be at best questionable and at worst utterly nonsensical. The difficult question I take up now is what we ought to do about this. I suggest three broad and largely disjoint courses of action researchers can pursue that would, in the aggregate, considerably improve the quality of research in psychological science.

Do something else

One perfectly reasonable course of action when faced with the difficulty of extracting meaningful, widely generalizable conclusions from effects that are inherently complex and highly variable is to opt out of the enterprise entirely. There is an unfortunate cultural norm within psychology (and, to be fair, many other fields) to demand that every research contribution end on a wholly positive or "constructive" note. This is an indefensible expectation that I won't bother to indulge. In life, you often can't have what you want, no matter how hard you try. In such cases, I think it's better to recognize the situation for what it is sooner rather than later. The fact that a researcher is able to formulate a question in his or her head that seems sensible (for example, "does ego depletion exist"?) doesn't mean that the question really *is* sensible. Moreover, even when the question is a sensible one to *ask* (in the sense that it's logically coherent and seems theoretically meaningful), it doesn't automatically follow that it's worth trying to obtain an empirical answer. In many research areas, if generalizability concerns were to be taken seriously, the level of effort required to obtain even minimally informative answers to seemingly interesting questions would likely so far exceed conventional standards that I suspect many academic psychologists would, if they were dispassionate about the matter, simply opt out. I see nothing wrong with such an outcome, and think it is a mistake to view a career in psychology (or any other academic field) as a higher calling of some sort.

Admittedly, the utility of this advice depends on one's career stage, skills, and interests. It should not be terribly surprising if few tenured professors are eager to admit (even to themselves) that they have, as Paul Meehl rather colorfully put it, "achieved some notoriety, tenure, economic security and the like by engaging, to speak bluntly, in a bunch of nothing" (P. E. Meehl, 1990b, p. 230). The situation is more favorable for graduate students and postdocs, who have much less to lose (and potentially much more to gain) by pursuing alternative careers. To be clear, I'm not suggesting that a career in academic psychology isn't a worthwhile pursuit *for anyone*; for many people, it remains an excellent choice. But I do think all psychologists, and early-career researchers in particular, owe it to themselves to spend some time carefully and dispassionately assessing the probability that the work they do is going to contribute meaningfully—even if only incrementally—to our collective ability either to understand the mind or to practically improve the human condition. There is no shame whatsoever in arriving at a negative answer, and the good news is that, for people who have managed to obtain a PhD (or have the analytical skills to do so), career prospects outside of academia have arguably never been brighter.

Embrace qualitative analysis

A second approach one can take is to keep doing psychological research, but to largely abandon inferential statistical methods in favor of qualitative methods. This may seem like a radical prescription, but I contend that a good deal of what currently passes for empirical psychology is already best understood as insightful qualitative analysis trying to quietly pass for quantitative science. Careful consideration of the logical structure of a psychological theory often makes it clear that there is little point in subjecting the theory to quantitative analysis. Sometimes this is because the theory appears logically incoherent, or is so vague as to make falsification via statistical procedures essentially impossible. Very often, however, it is because careful inspection reveals that the theory is actually *too* sensible. That is, its central postulates are so obviously true that there is nothing to be gained by subjecting it to further empirical tests—effectively constituting what Smedslund (1991) dubbed "pseu-

doempirical research".

To see what I mean, let's return to our running example of verbal overshadowing. To judge by the accumulated literature (for reviews, see Meissner & Brigham, 2001; Meissner & Memon, 2002), the question of whether verbal overshadowing is or is not a "real" phenomenon seems to be taken quite seriously by many researchers. Yet it's straightforward to show that some phenomenon like verbal overshadowing must exist given even the most basic, uncontroversial facts about the human mind. Consider the following set of statements:

1. The human mind has a finite capacity to store information.
2. There is noise in the information-encoding process.
3. Different pieces of information will sometimes interfere with one another during decision-making—either because they directly conflict, or because they share common processing bottlenecks.

None of the above statements should be at all controversial, yet the conjunction of the three logically entails that there will be (many) situations in which something we could label verbal overshadowing will predictably occur. Suppose we take the set of all situations in which a person witnesses, and encodes into memory, a crime taking place. In some subset of these cases, that person will later reconsider, and verbally re-encode, the events they observed. Because the encoding process is noisy, and conversion between different modalities is necessarily lossy, some details will be overemphasized, underemphasized, or otherwise distorted. And because different representations of the same event will conflict with one another, it is then guaranteed that there will be situations in which the verbal re-consideration of information at Time 2 will lead a person to incorrectly ignore information they may have correctly encoded at Time 1. We can call this *verbal overshadowing* if we like, but there is nothing about the core idea that requires any kind of empirical demonstration. So long as it's framed strictly in broad qualitative terms, the "theory" is trivially true; the only way it could be false is if at least one of the 3 statements listed above is false—which is almost impossible to imagine. (Note too, that the inverse of the theory is also trivially true: there must be many situations in which lossy re-encoding of information across modalities actually ends up being accidentally beneficial.)

To be clear, I am not suggesting that there's no point in quantitatively studying broad putative constructs like verbal overshadowing. On the contrary: if our goal is to develop models detailed enough to make useful real-world predictions, quantitative analysis may be indispensable. It would be difficult to make real-world predictions about when, where, and to what extent verbal overshadowing will manifest unless one has systematically studied and modeled the putative phenomenon under a broad range of conditions—including extensive variation of the perceptual stimuli, viewing conditions, rater incentives, timing parameters, and so on and so forth. But taking this quantitative objective seriously requires much larger and more complex datasets, experimental designs, and statistical models than have typically been deployed in most areas of psychology. As such, psychologists intent on working in "soft" domains who are unwilling to learn potentially challenging new modeling skills—or to spend months or years trying to meticulously address "minor" methodological concerns that presently barely rate any mention in papers—may need to accept that their work is, at root, qualitative in nature, and that the inferential statistics so often reported in soft psychology articles primarily serve as a ritual intended to convince one's colleagues and/or one's self that something very scientific and important is taking place.

What would a qualitative psychology look like? In many sub-fields, almost nothing would change. The primary difference is that researchers would largely stop using inferential statistics, restricting themselves instead to descriptive statistics and qualitative discussion. Such a policy is not without precedent: in 2014, the journal Basic and Applied Social Psychology banned the reporting of p-values from all submitted manuscripts (Trafimow, 2014; Trafimow & Marks, 2015). Although the move was greeted with derision

by many scientists (Woolston, 2015), what is problematic about the BASP policy is, in my view, only that the abolition of inferential statistics was made mandatory. Framed as a strong recommendation that psychologists should avoid reporting inferential statistics that they often do not seem to understand, and that have no clear implications for our understanding of, or interaction with, the world, I think there would be much to like about the policy.

For many psychologists, fully embracing qualitative analysis would provide an explicit license to do what they are already most interested in doing—namely, exploring big ideas, generalizing conclusions far and wide, and moving swiftly from research question to research question. The primary cost would be the reputational one: in a world where most psychology papers are no longer accompanied by scientific-looking inferential statistics, journalists and policy-makers would probably come knocking on our doors less often. I don't deny that this is a non-trivial cost, and I can understand why many researchers would be hesitant to pay such a toll. But such is life. I don't think it requires a terribly large amount of intellectual integrity to appreciate that one shouldn't portray one's self as a serious quantitative scientist unless one is actually willing to do the corresponding work.

Lest this attitude seem overly dismissive of qualitative approaches, it's worth noting that the core argument made in this paper is itself a qualitative one. I do not rely on inferential statistical results to support my conclusions, and all of the empirical data I quantitatively analyze are used strictly to illustrate general principles. Put differently, I am not making a claim of the form "87% of psychology articles draw conclusions that their data do not support"; I am observing that under modest assumptions that seem to me almost impossible to dispute in most areas of psychology (e.g., that the aggregate contribution of random variation in factors like experimental stimuli, task implementation, experimenter, site, etc., is (i) large, and (ii) almost never modeled), it is logically entailed that the conclusions researchers draw verbally will routinely deviate markedly from what the reported statistical analyses can strictly support. Researchers are, of course, free to object that this sweeping conclusion might not apply to *their* particular study, or that the argument would be more persuasive if accompanied by a numerical estimate of the magnitude of the problem in different areas[10]. But the mere fact that an argument is qualitative rather than quantitative in nature does not render it inferior or dismissible. On the contrary, as the verbal overshadowing example above illustrates, even a relatively elementary qualitative analysis can often provide more insightful answers to a question than a long series of ritualistic quantitative analyses. So I mean it sincerely when I say that an increased emphasis on qualitative considerations would be a welcome development in its own right in psychology, and should not be viewed as a consolation prize for studies that fail to report enough numbers.

Adopt better standards

The previous two suggestions are not a clumsy attempt at dark humor; I am firmly convinced that many academic psychologists would be better off either pursuing different careers, or explicitly acknowledging the fundamentally qualitative nature of their work (I lump myself into the former group much of the time, and this paper itself exemplifies the latter). For the remainder—i.e., those who would like to approach their research from a more quantitatively defensible perspective—there are a number of practices that, if deployed widely, could greatly improve the quality and reliability of quantitative psychological inference.

Draw more conservative inferences

Perhaps the most obvious, and arguably easiest, solution to the generalizability problem is for authors to draw much more conservative inferences in their manuscripts—and in particular, to replace the sweep-

[10] For what it's worth, it's unclear how much utility global quantitative estimates of this kind could actually have given the enormous variation across studies, and the relative ease of obtaining directly relevant local estimates. Individual researchers who want to know whether or not it is safe to assume zero stimulus, experimenter, or task effects in their statistical models do not have to wait for someone else to conduct a comprehensive variance-partitioning meta-analysis in their general domain; they can simply calculate the variance over such factors in their own prior datasets!

ing generalizations pervasive in contemporary psychology with narrower conclusions that hew much more closely to the available data. Concretely, researchers should avoid extrapolating beyond the universe of observations implied by their experimental designs and statistical models without clearly indicating that they are engaging in speculation. Potentially relevant design factors that are impractical to measure or manipulate, but that conceptual considerations suggest are likely to have non-trivial effects (e.g., effects of stimuli, experimenter, research site, culture, etc.), should be identified and disclosed to the best of authors' ability. Papers should be given titles like "Transient manipulation of self-reported anger influences small hypothetical charitable donations", and not ones like "Hot head, warm heart: Anger increases economic charity". I strongly endorse the recent suggestion by Simons and colleagues that most manuscripts in psychology should include a *Constraints on Generality* statement that explicitly defines the boundaries of the universe of observations the authors believe their findings apply to (Simons, Shoda, & Lindsay, 2017)—as well as earlier statements to similar effects in other fields (e.g., sociology; Walker & Cohen, 1985).

Correspondingly, when researchers evaluate results reported by others, credit should only be given for what the empirical results of a study *actually* show—and not for what its authors claim they show. Continually emphasizing the importance of the distinction between verbal constructs and observable measurements would go a long way towards clarifying which existing findings are worth replicating and which are not. If researchers develop a habit of mentally reinterpreting a claim like "we provide evidence of ego depletion" as "we provide evidence that crossing out the letter *e* slightly decreases response accuracy on a subsequent Stroop task", I suspect that many findings would no longer seem important enough to warrant any kind of follow-up—at least, not until the original authors have conducted considerable additional work to demonstrate the generalizability of the claimed phenomenon.

Take descriptive research more seriously

Traditionally, purely descriptive research—where researchers seek to characterize and explore relationships between measured variables without imputing causal explanations or testing elaborate verbal theories—is looked down on in many areas of psychology. This stigma discourages modesty, inhibits careful characterization of phenomena, and often leads to premature and overconfident efforts to assess simplistic theories that are hopelessly disconnected from the complexity of the real world (Cronbach, 1975; Rozin, 2001). I suspect it stems to a significant extent from a failure to recognize and internalize just how fragile many psychological phenomena truly are. Acknowledging the value of empirical studies that do nothing more than carefully describe the relationships between a bunch of variables under a wide range of conditions would go some ways towards atoning for our unreasonable obsession with oversimplified causal explanations.

We know that a large-scale shift in expectations regarding the utility of careful descriptive work is possible, because other fields have undergone such a transition to varying extents. Perhaps most notably, in statistical genetics, the small-sample candidate gene studies that made regular headlines in the 1990s (e.g., Ebstein et al., 1996; Lesch et al., 1996)—virtually all of which later turned out to be spurious (Chabris et al., 2012; Colhoun, McKeigue, & Davey Smith, 2003; Sullivan, 2007), and were motivated by elegant theoretical hypotheses that seem laughably simplistic in hindsight—have all but disappeared in favor of massive genome-wide association studies (GWAS) involving hundreds of thousands of subjects (Nagel et al., 2018; Savage et al., 2018; Wray et al., 2018). The latter are now considered the gold standard even in cases where they do little more than descriptively identify novel statistical associations between gene variants and behavior. In much of statistical genetics, at least, researchers seem to have accepted that the world is causally complicated, and attempting to obtain a reasonable descriptive characterization of some small part of it is a perfectly valid reason to conduct large, expensive empirical studies.

Fit more expansive statistical models

To the degree that authors intend for their conclusions to generalize over populations of stimuli, tasks, experimenters, and other such factors, they should develop a habit of fitting more expansive statistical models. As noted earlier, nearly all statistical analyses of multi-subject data in psychology treat subject as a varying effect. The same treatment should be accorded to other design factors that researchers intend to generalize over and that vary controllably or naturally in one's study. Of course, inclusion of additional random effects is only one of many potential avenues for sensible model expansion (Draper, 1995; Gelman & Shalizi, 2013)[11]. The good news is that improvements in statistical computing over the past few years have made it substantially easier for researchers to fit arbitrarily complex mixed-effects models within both Bayesian and frequentist frameworks. Models that were once intractable for most researchers to fit due to either mathematical or computational limitations can now often be easily specified and executed on modern laptops using mixed-effects packages (e.g., lmer or MixedModels.jl; (Bates, Maechler, Bolker, Walker, & Others, 2014)) or probabilistic programming frameworks (e.g., Stan or PyMC; (Carpenter et al., 2017; Salvatier, Wiecki, & Fonnesbeck, 2016)).

This recommendation conveniently sidesteps the question of *which* varying factors researchers should choose to focus on. A number of commentators on earlier drafts of this paper have suggested that the general prescription to fit bigger models, while technically reasonable, is too vague to be helpful. I am sympathetic to this concern, but nevertheless think that attempting to make generic statements about the relative importance of different sources of variation in "typical" psychology studies would be a mistake. There are two reasons for this. First, I see little reason to think that any brief domain-general summary of the relative magnitudes of different variance components would have much utility for almost any individual study. How important is it to consider the role of different task operationalizations? Do cross-cultural differences have a small or large impact on observed effect sizes? And what about experimenter effects, how big are those? The only answer one can give to such questions that is both honest and concise is "it depends".

Second, the sense of discomfort some readers might feel at the realization that they don't know what to do next is, in my view, a feature, not a bug. It *should* bother researchers to discover that they don't have a good sense of what the major sources of variance are in the data they routinely work with. What does it say about a researcher's ability to update their belief in a hypothesis if they cannot even roughly state the conditions under which the obtained statistical results would or would not constitute an adequate test of the hypothesis? I would not want to give researchers the impression that there is some generic list of factors one can rely on here; there is simply no substitute for careful and critical consideration of the data-generating processes likely to underlie each individual effect of interest.

Design with variation in mind

In most areas of psychology, there is a long-dominant tradition of trying to construct randomized experiments that are as tightly controlled as possible—even at the cost of decreased generalizability. Though calls for researchers to emphasize the opposite side of the precision-generalization tradeoff—i.e., to embrace naturalistic, ecologically valid designs that embrace variability—have a long history in psychology (Brunswik, 1947; Cronbach, 1975), they have intensified considerably in recent years. For example, in neuroimaging, researchers are increasingly fitting sophisticated models to naturalistic stimuli such as coherent narratives or movies (Hamilton & Huth, 2018; Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Huth, Nishimoto, Vu, & Gallant, 2012; Spiers & Maguire, 2007). In psycholinguistics, large-scale analyses involving databases of thousands of words and subjects have superseded traditional small-n factorial studies for many applications (Balota, Yap, Hutchison, & Cortese, 2012; Keuleers & Balota, 2015). Even

[11] In a sense, the very idea of a random effect is just a convenient fiction—effectively a placeholder for a large number of hypothetical fixed variables (or functions thereof) that we presently do not know how to write, or lack the capacity to measure and/or estimate.

in domains where many effects traditionally display little sensitivity to context, some researchers have advocated for analysis strategies that emphasize variability. For example, Baribault and colleagues (2018) randomly varied 16 different experimental factors in a large multi-site replication (6 sites, 346 subjects, and nearly 5,000 "microexperiments") of a subliminal priming study (Reuss, Kiesel, & Kunde, 2015). The "radical randomization" strategy the authors adopted allowed them to draw much stronger conclusions about the generalizability of the priming effect (or lack thereof) than would have otherwise been possible.

The deliberate introduction of variance into one's studies can also be construed as a more principled version of the *conceptual replication* strategy already common in many areas of psychology. In both cases, researchers seek to determine the extent to which an effect generalizes across the levels of one or more secondary design factors. The key difference is that traditional conceptual replications do not lend themselves well to a coherent modeling strategy: when authors present a series of discrete conceptual replications in Studies 2 through N of a manuscript, it is rarely obvious how one can combine the results to obtain a meaningful estimate of the robustness or generalizability of the common effect. By contrast, explicitly modeling the varying factors as components of a single overarching design makes it clear what the putative relationship between different measurements is, and enables stronger quantitative inferences to be drawn.

Naturally, variation-enhancing designs come at a cost: they will often demand greater resources than conventional approaches that seek to minimize extraneous variation. But if authors intend for their conclusions to hold independently of variation in uninteresting factors, and to generalize to broad classes of situations, there is no good substitute for studies whose designs make a serious effort to respect and capture the complexity of real-world phenomena. Large-scale, collaborative projects of the kind pioneered in RRRs (Simons et al., 2014) and recent initiatives such as the Psychology Accelerator (Moshontz et al., 2018) are arguably the natural venue for such an approach—but to maximize their utility, the substantial resources they command must be used to directly measure and model variability rather than minimizing and ignoring it.

Emphasize variance estimates

An important and underappreciated secondary consequence of the widespread disregard for generalizability is that researchers in many areas of psychology rarely have good data—or even just strong intuitions—about the relative importances of different sources of variance. One way to mitigate this problem is to promote analytical approaches that emphasize the estimation of variance components rather than focusing solely on point estimates. For primary research studies, Generalizability Theory (Brennan, 1992; Cronbach, Rajaratnam, & Gleser, 1963; Shavelson & Webb, 1991) provides a well-developed (and underused) framework for computing and applying such estimates. At the secondary level, meta-analysts could similarly work to quantify the magnitudes of different variance components—either by meta-analyzing reported within-study variance estimates, or by meta-analytically computing between-study variance components for different factors. Such approaches could provide researchers with critically important background estimates of the extent to which a new finding reported in a particular literature should be expected to generalize to different samples of subjects, stimuli, experimenters, research sites, and so on. Notably, such estimates would be valuable irrespective of the presence or absence of a main effect of putative interest. For example, even if the accumulated empirical literature is too feeble to allow us to estimate anything approximating a single overall *universe score* for ego depletion, it would still be extremely helpful when planning a new study to know roughly how much of the observed variation in the existing pool of studies is due to differences in stimuli, subjects, tasks, etc.

Make riskier predictions

There is an important sense in which most of the other recommendations made in this section could be obviated simply by making theoretical predictions that assume a high degree of theoretical risk. I have

approached the problem of generalizability largely from a statistical perspective, but there is a deep connection between the present concerns and a long tradition of philosophical commentary focusing on the logical relationship (or lack thereof) between theoretical hypotheses and operational or statistical ones.

Perhaps the best exposition of such ideas is found in the seminal work of Paul Meehl, who, beginning in the 1960s, argued compellingly that many of the methodological and statistical practices routinely applied by psychologists and other social scientists are logically fallacious (e.g., P. E. Meehl, 1967, 1978, 1990b). Meehl's thinking was extremely nuanced, but a recurring theme in his work is the observation that most hypothesis tests in psychology commit the logical fallacy of affirming the consequent. A theory T makes a prediction P, and when researchers obtain data consistent with P, they then happily conclude that T is corroborated. In reality, the confirmation of P provides no meaningful support for T unless the prediction was relatively specific to T—that is, there are no readily available alternative theories $T'_1 \dots T'_k$ that also predict P. Unfortunately, in most domains of psychology, there are pervasive and typically very plausible competing explanations for almost every finding (Cohen, 2016; Lykken, 1968; P. E. Meehl, 1967, 1986).

The solution to this problem is, in principle, simple: researchers should strive to develop theories that generate risky predictions (P. Meehl, 1997; P. E. Meehl, 1967, 1990a; Popper, 2014)—or, in the terminology popularized by Deborah Mayo, should subject their theories to *severe tests* (Mayo, 1991, 2018). The canonical way to accomplish this is to derive from one's theory some series of predictions—typically, but not necessarily, quantitative in nature—sufficiently specific to that theory that they are inconsistent with, or at least extremely implausible under, other accounts. As Meehl put it:

> If my meteorological theory successfully predicts that it will rain sometime next April, and that prediction pans out, the scientific community will not be much impressed. If my theory enables me to correctly predict which of 5 days in April it rains, they will be more impressed. And if I predict how many millimeters of rainfall there will be on each of these 5 days, they will begin to take my theory very seriously indeed (P. E. Meehl, 1990a, p. 110).

The ability to generate and corroborate a truly risky prediction strongly implies that a researcher must already have a decent working model (even if only implicitly) of most of the contextual factors that could potentially affect a dependent variable. If a social psychologist were capable of directly deriving from a theory of verbal overshadowing the prediction that target recognition should decrease 1.7% +/- 0.04% in condition A relative to condition B in a given experiment, concerns about the generalizability of the theory would dramatically lessen, as there would rarely be a plausible alternative explanation for such precision other than that the theories in question were indeed accurately capturing something important about the way the world works.

In practice, it's clearly wishful thinking to demand this sort of precision in most areas of psychology (potential exceptions include, e.g., parts of psychophysics, mathematical cognitive psychology, and behavioral genetics). The very fact that most of the phenomena psychologists study are enormously complex, and admit a vast array of causal influences in even the most artificially constrained laboratory situations, likely precludes the production of quantitative models with anything close to the level of precision one routinely observes in the natural sciences. This does not mean, however, that vague directional predictions are the best we can expect from psychologists. There are a number of strategies that researchers in such fields could adopt that would still represent at least a modest improvement over the status quo (for discussion, see Gigerenzer, 2017; Lilienfeld, 2004; P. E. Meehl, 1990a; Roberts & Pashler, 2000). For example, researchers could use equivalence tests (Lakens, 2017); predict specific orderings of discrete observations; test against compound nulls that require the conjunctive rejection of many independent directional predictions; and develop formal mathematical models that posit non-trivial functional forms between the input and

ouput variables (Marewski & Olsson, 2009; Smaldino, 2017). While it is probably unrealistic to expect truly severe tests to become the norm in most fields of psychology, severity is an ideal worth keeping perpetually in mind when designing studies—if only as a natural guard against undue optimism.

Focus on practical predictive utility

An alternative and arguably more pragmatic way to think about the role of prediction in psychology is to focus not on the theoretical risk implied by a prediction, but on its practical utility. Here, the core idea is to view psychological theories or models not so much as statements about how the human mind *actually* operates, but as convenient approximations that can help us intervene on the world in useful ways (Breiman, 2001; Hofman, Sharma, & Watts, 2017; Shmueli, 2010; Yarkoni & Westfall, 2017). For example, instead of asking the question *does verbal overshadowing exist*?, we might instead ask: *can we train a statistical model that allows us to meaningfully predict people's behaviors in a set of situations that superficially seem to involve verbal overshadowing*? The latter framing places emphasis primarily on what a model is able to *do* for us rather than on its implied theoretical or ontological commitments.

One major advantage of an applied predictive focus is that it naturally draws attention to objective metrics of performance that can be easier to measure and evaluate than the relatively abstract, and often vague, theoretical postulates of psychological theories. A strong emphasis on objective, communal measures of model performance has been a key driver of rapid recent progress in the field of machine learning (Jordan & Mitchell, 2015; LeCun, Bengio, & Hinton, 2015; Russakovsky et al., 2015)—including numerous successes in domains such as object recognition and natural language translation that arguably already fall within the purview of psychology and cognitive science. A focus on applied prediction would also naturally encourage greater use of large samples, as well as of cross-validation techniques that can minimize overfitting and provide alternative ways of assessing generalizability outside of the traditional inferential statistical framework (e.g., Woo, Chang, Lindquist, & Wager, 2017). Admittedly, a large-scale shift towards instrumentalism of this kind would break with a century-long tradition of explanation and theoretical understanding within psychology; however, as I have argued elsewhere (Yarkoni & Westfall, 2017), there are good reasons to believe that psychology would emerge as a healthier, more reliable discipline as a result.

Conclusion

Most contemporary psychologists view the use of inferential statistical tests as an integral part of the discipline's methodology. The ubiquitous reliance on statistical inference is the source of much of the perceived objectivity and rigor of modern psychology—the very thing that, in many people's eyes, makes it a quantitative science. I have argued that, for most research questions in most areas of psychology, this perception is illusory. Closer examination reveals that the inferential statistics reported in psychology articles typically have only a tenuous correspondence to the verbal claims they are intended to support. The overarching conclusion is that many fields of psychology currently operate under a kind of collective self-deception, using a thin sheen of quantitative rigor to mask inferences that remain, at their core, almost entirely qualitative.

Such concerns are not new, of course. Commentators have long pointed out that, viewed dispassionately, an enormous amount of statistical inference in psychology (and, to be fair, other sciences) has a decidedly ritualistic character: rather than improving the quality of scientific inference, the use of universalized testing procedures serves mainly to increase practitioners' subjective confidence in broad verbal assertions that would otherwise be difficult to defend on logical grounds (e.g., Gelman, 2016; Gigerenzer, 2004; Gigerenzer & Marewski, 2015; P. E. Meehl, 1967, 1990b; Tong, 2019). What I have tried to emphasize in the present treatment is that such critiques are not, as many psychologists would like to believe, pedantic worries about edge cases that one can safely ignore most of the time. The problems in question are fundamental, and follow directly from foundational assump-

tions of our most widely used statistical models. The central point is that the degree of support a statistical analysis lends to a verbal proposition derives not just from some critical number that the analysis does or doesn't pop out (e.g., $p < .05$), but also (and really, *primarily*), from the ability of the statistical model to implicitly define a universe that matches the one defined by the verbal proposition.

When the two diverge markedly—as I have argued is extremely common in psychology—one is left with a difficult choice to make. One possibility is to accept the force of the challenge and adjust one's standard operating procedures accordingly—by moderating one's verbal claims, narrowing the scope of one's research program, focusing on making practically useful predictions, and so on. This path is effort-intensive and incurs a high risk that the results one produces post-remediation will, at least superficially, seem less impressive than the ones that came before. But it is the intellectually honest road, and has the secondary benefit of reducing the probability of making unreasonably broad claims that are unlikely to stand the test of time.

The alternative is to simply brush off these concerns, recommit one's self to the same set of procedures that have led to prior success by at least *some* measures (papers published, awards received, etc.), and then carry on with business as usual. No additional effort is required here; no new intellectual or occupational risk is assumed. The main cost is that one must live with the knowledge that many of the statistical quantities one routinely reports in one's papers are essentially just an elaborate rhetorical ruse used to mathematize people into believing claims they would otherwise find logically unsound.

I don't pretend to think this is an easy choice. I have little doubt that the vast majority of researchers have good intentions, and genuinely want to do research that helps increase understanding of human behavior and improve the quality of people's lives. I am also sympathetic to objections that it's not fair to expect individual researchers to pro-actively hold themselves to a higher standard than the surrounding community, knowing full well that a likely cost of doing the right thing is that one's research may become more difficult to pursue, less exciting, and less well received by others. Unfortunately, the world we live in isn't always fair. I don't think anyone should be judged very harshly for finding major course correction too difficult an undertaking after spending years immersed in an intellectual tradition that encourages rampant overgeneralization. But the decision to stay the course should at least be an informed one: researchers who opt to ignore the bad news should recognize that, in the long term, such an attitude hurts the credibility both of their own research program and of the broader profession they have chosen. One is always free to pretend that small p-values obtained from extremely narrow statistical operationalizations can provide an adequate basis for sweeping verbal inferences about complex psychological constructs. But noone else—not one's peers, not one's funders, not the public, and certainly not the long-term scientific record—is obligated to honor the charade.

Acknowledgments

This paper was a labor of pain that took an inexplicably long time to produce. It owes an important debt to Jake Westfall for valuable discussions and comments—including the idea of fitting the "unmeasured variances" model illustrated in Figure 2—and also benefited from conversations with dozens of other people (many of them on Twitter, because it turns out you can say a surprisingly large amount in hundreds of 280-character tweets).

Funding statement

This work was supported by NIH awards R01MH096906 and R01MH109682.

Conflicts of Interest statement

None.

References

Acosta, A., Adams (Jr.), R. B., Albohn, D. N., Allard, E. S., Beek, T., Benning, S. D., … Zwaan, R. A. (2016, November). Registered replication report: Strack, martin, & stepper (1988). *Perspect. Psychol. Sci.*, *11*(6), 917–928.

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., … Zwaan, R. A. (2014, September). Registered replication report: Schooler and Engstler-Schooler (1990). *Perspect. Psychol. Sci.*, *9*(5), 556–578.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.*, *59*(4), 390–412.

Balota, D. A., Yap, M. J., Hutchison, K. A., & Cortese, M. J. (2012). Megastudies: What do millions (or so) of trials tell us about lexical processing? In J. S. Adelman (Ed.), *Visual word recognition volume 1: Models and methods, orthography and phonology*. Psychology Press.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., … Vandekerckhove, J. (2018, March). Metastudies for robust tests of theory. *Proc. Natl. Acad. Sci. U. S. A.*, *115*(11), 2607–2612.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013, April). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.*, *68*(3).

Bates, D., Maechler, M., Bolker, B., Walker, S., & Others. (2014). lme4: Linear mixed-effects models using eigen and S4. *R package version*, *1*(7), 1–23.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Others (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6.

Bergelson, E., Bergmann, C., Byers-Heinlein, K., Cristia, A., Cusack, R., Dyck, K., … others (2017). Quantifying sources of variability in infancy research using the infant-directed speech preference.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003, April). The theoretical status of latent variables. *Psychol. Rev.*, *110*(2), 203–219.

Breiman, L. (2001). Statistical modeling: The two cultures. *Stat. Sci.*, *16*(3), 199–215.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27–34.

Brunswik, E. (1947). Systematic and representative design of psychological experiments. In *Proceedings of the berkeley symposium on mathematical statistics and probability* (pp. 143–202).

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.*, *76*(1).

Chabris, C. F., Hebert, B. M., Benjamin, D. J., Beauchamp, J., Cesarini, D., van der Loos, M., … Laibson, D. (2012, September). Most reported genetic associations with general intelligence are probably false positives. *Psychol. Sci.*, *23*(11), 1314–1323.

Cheung, I., Campbell, L., LeBel, E. P., Ackerman, R. A., Aykutog˘lu, B., Bahník, Š., … Yong, J. C. (2016, September). Registered replication report: Study 1 from finkel, rusbult, kumashiro, & hannon (2002). *Perspect. Psychol. Sci.*, *11*(5), 750–764.

Clark, H. H. (1973, August). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359.

Cohen, J. (2016). The earth is round (p<. 05). In *What if there were no significance tests?* (pp. 69–82). Routledge.

Coleman, E. B. (1964). Generalizing to a language population. *Psychol. Rep.*, *14*(1), 219–226.

Colhoun, H. M., McKeigue, P. M., & Davey Smith, G. (2003, March). Problems of reporting genetic associations with complex outcomes. *Lancet*, *361*(9360), 865–872.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *The Annals of Mathematical Statistics*, *27*(4), 907–949.

Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science*, *284*(5420), 1670–1672.

Crits-Christoph, P., & Mintz, J. (1991, February). Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *J. Consult. Clin. Psychol.*, *59*(1), 20–26.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American psychologist*, *30*(2), 116.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.*, *52*(4), 281–302.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *Br. J. Math. Stat. Psychol.*, *16*(2), 137–163.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 45–70.

Ebstein, R. P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., … Belmaker, R. H. (1996). Dopamine D4 receptor (D4DR) exon III polymorphism associated with the human personality trait of novelty seeking. *Nat. Genet.*, *12*(1), 78–80.

Eerland, A. S., Magliano, A. M., Zwaan, J. P., Arnal, R. A., Aucoin, J. D., & Crocker, P. (2016). Registered replication report: Hart & albarracín (2011). *Perspect. Psychol. Sci.*, *11*(1), 158–171.

Feynman, R. P. (1974). Cargo cult science. *Eng. Sci.*, *37*(7), 10–13.

Francis, G. (2012, December). Publication bias and the failure of replication in experimental psychology. *Psychon. Bull. Rev.*, *19*(6), 975–991.

Gelman, A. (2015, February). The connection between varying treatment effects and the crisis of unreplicable research: A bayesian perspective. *J. Manage.*, *41*(2), 632–643.

Gelman, A. (2016). The problems with p-values are not just with p-values. *Am. Stat.*, *70*(supplemental material to the ASA statement on p-values and statistical significance), 10.

Gelman, A. (2018, January). The failure of null hypothesis significance testing when studying incremental changes, and what to do about it. *Pers. Soc. Psychol. Bull.*, *44*(1), 16–23.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and Multilevel/Hierarchical models*. Cambridge University Press.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis. *Downloaded January*, 1–17.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38.

Gigerenzer, G. (2004, November). Mindless statistics. *J. Socio Econ.*, *33*(5), 587–606.

Gigerenzer, G. (2017). A theory integration program. *Decision*, *4*(3), 133.

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, *41*(2), 421–440.

Guion, R. M. (1980, June). On trinitarian doctrines of validity. *Prof. Psychol.*, *11*(3), 385–398.

Hamilton, L. S., & Huth, A. G. (2018, July). The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 1–10.

Hofman, J. M., Sharma, A., & Watts, D. J. (2017, February). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016, April). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012, December). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224.

Ioannidis, J. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med.*, *2*(8), e124.

John, L. K., Loewenstein, G., & Prelec, D. (2012, May). Measuring the prevalence of question-

able research practices with incentives for truth telling. *Psychol. Sci.*, *23*(5), 524–532.

Jordan, M. I., & Mitchell, T. M. (2015, July). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). *Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem* (Vol. 103) (No. 1).

Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Q. J. Exp. Psychol.*, *68*(8), 1457–1468.

Kruschke, J. K., & Liddell, T. M. (2017, February). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychon. Bull. Rev.*.

Kühberger, A., Fritz, A., & Scherndl, T. (2014, September). Publication bias in psychology: a diagnosis based on the correlation between effect size and sample size. *PLoS One*, *9*(9), e105825.

Lakens, D. (2017, May). Equivalence tests: A practical primer for t tests, correlations, and Meta-Analyses. *Soc. Psychol. Personal. Sci.*, *8*(4), 355–362.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. A. (2018, February). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171.

LeCun, Y., Bengio, Y., & Hinton, G. (2015, May). Deep learning. *Nature*, *521*(7553), 436–444.

Lesch, K. P., Bengel, D., Heils, A., Sabol, S. Z., Greenberg, B. D., Petri, S., … Murphy, D. L. (1996). Association of anxiety-related traits with a polymorphism in the serotonin transporter gene regulatory region. *Science*, *274*(5292), 1527–1531.

Lilienfeld, S. O. (2004). Taking theoretical risks in a world of directional predictions. *Applied and Preventive Psychology*, *11*(1), 47–51.

Lilienfeld, S. O. (2017, July). Psychology's replication crisis and the grant culture: Righting the ship. *Perspect. Psychol. Sci.*, *12*(4), 660–664.

Lykken, D. T. (1968, September). Statistical significance in psychological research. *Psychol. Bull.*, *70*(3), 151–159.

MacLeod, C. M. (1991, March). Half a century of research on the stroop effect: an integrative review. *Psychol. Bull.*, *109*(2), 163–203.

Marewski, J. N., & Olsson, H. (2009). Beyond the null ritual: Formal modeling of psychological processes. *Zeitschrift für Psychologie/Journal of Psychology*, *217*(1), 49–60.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017, June). Balancing type I error and power in linear mixed models. *J. Mem. Lang.*, *94*, 305–315.

Mayo, D. G. (1991). Novel evidence and severe tests. *Philos. Sci.*, *58*(4), 523–552.

Mayo, D. G. (2018). *Statistical inference as severe testing*. Cambridge: Cambridge University Press.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245.

Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Erlbaum.

Meehl, P. E. (1967). Theory-Testing in psychology and physics: A methodological paradox. *Philos. Sci.*, *34*(2), 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir karl, sir ronald, and the slow progress of soft psychology. *J. Consult. Clin. Psychol.*, *46*(4), 806.

Meehl, P. E. (1986). 14 what social scientists don't understand. *Metatheory in social science: Pluralisms and subjectivities*, 315.

Meehl, P. E. (1990a, April). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychol. Inq.*, *1*(2), 108–141.

Meehl, P. E. (1990b, February). Why summaries of research on psychological theories are often uninterpretable. *Psychol. Rep.*, *66*(1), 195–244.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in

face identification. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *15*(6), 603–616.

Meissner, C. A., & Memon, A. (2002, December). Verbal overshadowing: A special issue exploring theoretical and applied issues. *Appl. Cogn. Psychol.*, *16*(8), 869–872.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Chartier, C. R. (2018, July). Psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*.

Nagel, M., Jansen, P. R., Stringer, S., Watanabe, K., de Leeuw, C. A., Bryois, J., … Posthuma, D. (2018, July). Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat. Genet.*, *50*(7), 920–927.

O'Leary-Kelly, S. W., & J. Vokurka, R. (1998). The empirical assessment of construct validity. *J. Oper. Manage.*, *16*(4), 387–405.

Pashler, H., & Wagenmakers, E.-J. (2012, November). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspect. Psychol. Sci.*, *7*(6), 528–530.

Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. routledge.

Reuss, H., Kiesel, A., & Kunde, W. (2015, January). Adjustments of response speed and accuracy to unconscious cues. *Cognition*, *134*, 57–62.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? a comment on theory testing. *Psychological review*, *107*(2), 358.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009, April). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.*, *16*(2), 225–237.

Rozin, P. (2001). Social psychology and science: Some lessons from solomon asch. *Personality and Social Psychology Review*, *5*(1), 2–14.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015, December). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, *115*(3), 211–252.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016, April). Probabilistic programming in python using PyMC3. *PeerJ Comput. Sci.*, *2*, e55.

Savage, J. E., Jansen, P. R., Stringer, S., Watanabe, K., Bryois, J., de Leeuw, C. A., … Posthuma, D. (2018, July). Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.*, *50*(7), 912–919.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990, January). Verbal overshadowing of visual memories: some things are better left unsaid. *Cogn. Psychol.*, *22*(1), 36–71.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. SAGE.

Shmueli, G. (2010). To explain or to predict? *Stat. Sci.*, 289–310.

Shrout, P. E., & Rodgers, J. L. (2018, January). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annu. Rev. Psychol.*, *69*, 487–510.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.*, *22*(11), 1359–1366.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014, September). An introduction to registered replication reports at perspectives on psychological science. *Perspect. Psychol. Sci.*, *9*(5), 552–555.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017, November). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspect. Psychol. Sci.*, *12*(6), 1123–1128.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In *Computational social psychology* (pp. 311–331). Routledge.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society open science*, *3*(9), 160384.

Smedslund, J. (1991). The pseudoempirical in psychology and the case for psychologic. *Psychological Inquiry*, *2*(4), 325–338.

Spiers, H. J., & Maguire, E. A. (2007). Decoding human brain activity during real-world experiences. *Trends Cogn. Sci.*, *11*(8), 356–365.

Steckler, A., McLeroy, K. R., Goodman, R. M., Bird, S. T., & McCormick, L. (1992). Toward integrating qualitative and quantitative methods: an introduction. *Health Educ. Q.*, *19*(1), 1–8.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *J. Exp. Psychol.*, *18*(6), 643.

Sullivan, P. F. (2007, May). Spurious genetic associations. *Biol. Psychiatry*, *61*(10), 1121–1126.

Tong, C. (2019). Statistical inference enables bad science; statistical thinking enables good science. *The American Statistician*, *73*(sup1), 246–261.

Trafimow, D. (2014, January). Editorial. *Basic Appl. Soc. Psych.*, *36*(1), 1–2.

Trafimow, D., & Marks, M. (2015, January). Editorial. *Basic Appl. Soc. Psych.*, *37*(1), 1–2.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016, June). Contextual sensitivity in scientific reproducibility. *Proc. Natl. Acad. Sci. U. S. A.*, *113*(23), 6454–6459.

Wagenmakers, E.-J. (2007, October). A practical solution to the pervasive problems ofp values. *Psychon. Bull. Rev.*, *14*(5), 779–804.

Wahlsten, D., Metten, P., Phillips, T. J., Boehm, S. L., Burkhart-Kasch, S., Dorow, J., … others (2003). Different data from different labs: lessons from studies of gene–environment interaction. *Journal of neurobiology*, *54*(1), 283–311.

Walker, H. A., & Cohen, B. P. (1985). Scope statements: Imperatives for evaluating theory. *American Sociological Review*, 288–301.

Westfall, J., Nichols, T. E., & Yarkoni, T. (2016, December). Fixing the stimulus-as-fixed-effect fallacy in task fMRI. *Wellcome Open Res*, *1*, 23.

Wolsiefer, K., Westfall, J., & Judd, C. M. (2017, August). Modeling stimulus variation in three common implicit attitude tasks. *Behav. Res. Methods*, *49*(4), 1193–1209.

Woolston, C. (2015, March). Psychology journal bans P values. *Nature News*, *519*(7541), 9.

Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., … Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2018, May). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.*, *50*(5), 668–681.

Yarkoni, T. (2009, May). Big correlations in little studies: Inflated fMRI correlations reflect low statistical Power-Commentary on vul et al. (2009). *Perspect. Psychol. Sci.*, *4*(3), 294–298.

Yarkoni, T., & Westfall, J. (2017, November). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.*, *12*(6), 1100–1122.