NOTE: This is a pre-publication manuscript version of a published article. This paper is not the copy of record and may not exactly replicate the authoritative document published in the journal. The final article is available at: https://doi.org/10.1016/j.jesp.2015.09.006

The Empirical Benefits of Conceptual Rigor: Systematic Articulation of Conceptual Hypotheses Can Reduce the Risk of Non-Replicable Results (and Facilitate Novel Discoveries Too)

Mark Schaller
University of British Columbia

Abstract

Most discussions of rigor and replication focus on *empirical* practices (methods used to collect and analyze data). Typically overlooked is the role of *conceptual* practices: The methods scientists use to arrive at and articulate research hypotheses in the first place. This article discusses how the conceptualization of research hypotheses has implications for methodological decision-making and, consequently, for the replicability of results. The article identifies three ways in which empirical findings may be non-replicable, and shows how all three kinds of non-replicability are more likely to emerge when scientists take an informal conceptual approach, in which personal predictions are equated with scientific hypotheses. The risk of non-replicability may be reduced if scientists adopt more formal conceptual practices, characterized by the rigorous use of "if-then" logic to articulate hypotheses, and to systematically diagnose the plausibility, size, and context-dependence of hypothesized effects. The article identifies benefits that are likely to arise from more rigorous and systematic conceptual practices, and identifies ways in which their use can be encouraged to be more normative within the scholarly culture of the psychological sciences.

Keywords: hypotheses, replication, false positives, false negatives, effect size, generalizability, research methods, research practices, scientific rigor, pedagogy

> "Ideas do not belong to anyone," he said. With his finger he sketched a series of continuous circles in the air and concluded: "They fly around up there like the angels." (Gabriel García Márquez, 1995, p. 56)

> "We must distinguish between, on the one hand, our subjective experiences or our feelings of conviction, which can never justify any statement … and, on the other hand, the objective logical relations subsisting among the various systems of scientific statements, and within each of them." (Karl Popper, 1959/2005, p. 22)

The twofold goal of any science is this: Maximize the production of novel empirical discoveries, while minimizing the production of erroneous and non-replicable results.[1] Within the psychological sciences there is currently renewed attention to the second half of that goal. Considerable efforts are currently being devoted to systematically diagnosing the extent to which previously obtained findings do—or don't—replicate (Brandt et al., 2014; Open Science Collaboration, 2012, 2015; Schweinsberg et al., this issue). This is a valuable enterprise, and it sure beats the haphazard ways in which non-replicable findings were (or weren't) detected in the past. Of course, as scientists, we want not only to identify previously published findings of

dubious replicability, we also want to limit the illusory "discovery" of non-replicable findings in the first place. There is real value associated with any methodological strategy that can help to reduce the likelihood that an empirical study will produce non-replicable results. Many such strategies and "best practices" have been recommended (Asendorpf et al., 2013; Finkel, Eastwick, & Reis, 2015; Funder, Levine, Mackie, Morf, Vazire, & West, 2014; Maner, this issue; Maruyama, Pekrun, & Fiedler, 2014; Sakaluk, this issue).

Although demonstrably beneficial, many of these strategies are also associated with limitations or countervailing costs. Consider, for example, the various empirical practices that are designed to decrease the risk of Type 1 inferential errors ("false positives") by restricting ad-hoc "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011). These practices can indeed reduce the possibility of Type 1 error in situations in which there is no real effect to be found; but, in situations in which there actually *is* a real effect to be found—and so the risk of making a Type 1 error is non-existent—employment of these strategies *increases* the risk of making Type 2 errors ("false negatives") and thus inhibits scientific discovery (Fiedler, Kutzner, & Krueger, 2012; Maruyama et al., 2014). Or consider the recommendation to use large sample sizes. Large sample sizes are unassailably beneficial: They enhance statistical power (thus reducing the risk of Type 2 error when there is a real effect to be found), limit the temptation to scientists to use potentially problematic ad-hoc strategies to increase power (thus reducing the risk of Type 1 error when there is no effect to be found), and enhance the accuracy of effect size estimates too. But, for some research questions, large samples just aren't feasible; and, for any methodology that poses a non-zero risk to research participants, any increase in sample size introduces additional ethical costs that must be weighed against the benefits. In general, enthusiasm for any potentially beneficial methodological practice must be balanced by careful consideration of associated costs and limitations too.

It is for this reason that the catalog of potentially beneficial methodological reforms remains a work in progress. In this article, I draw attention to a kind of methodological approach that has been largely overlooked in discussions of inference errors and replicability and "best practices" for the psychological sciences. It is an approach that pertains not to *empirical* practices, but instead to *conceptual* practices.

Generally speaking, most recommendations for methodological rigor and replicability focus on the methods that scientists use to collect and analyze data. Rarely discussed is the potentially important role of the methods that scientists use to identify research hypotheses in the first place. The purpose of this article is to fill that gap (or, at least, to stimulate a scientific conversation that will, over time, more completely fill that gap). I discuss how non-replicable results are more likely to emerge when scientists take an informal and idiosyncratic approach to research hypotheses—an approach characterized by the tendency to treat personal predictions as equivalent to scientific hypotheses. I show how this "personalized" approach to hypotheses increases the likelihood that scientists will make the kinds of decisions (in study design, data analysis, and reporting of results) that, in turn, increase the likelihood of non-replication. This informal approach is compared to a more analytically rigorous de-personalized approach, which is characterized by systematic use of if-then logic and which offers a systematic means of proactively appraising the plausibility, size, and generalizability of hypothesized effects. This more rigorous conceptual approach can help scientists make the kinds of practical decisions (in study design, data analysis, and reporting) that facilitate the production of empirical results that are real and replicable.

Although the focus here is on the minimization of non-replicable results, I also briefly consider other consequences that may follow from the deployment of an analytically rigorous approach to research hypotheses. Of particular note is that this conceptual approach may not only inhibit the production of false "discoveries," it also has the potential to facilitate discovery of novel phenomena that really do exist.

Three Categories of Non-Replicable Empirical Results

Before discussing how different conceptual approaches to hypotheses have different implications for replicability, it is useful to identify three kinds of empirical findings that resist replication: Erroneous effects, erroneously big effects, and erroneously broad effects.

Erroneous Effects (False Positives)

The most obvious kind of non-replicable findings are those in which some empirical effect is detected (e.g., some non-zero relation measured between two variables) when, in fact, no such effect exists in reality. In other words: a Type 1 inferential error or false positive error. These non-zero effects resist replication because they simply aren't real.

Erroneously Big Effects (Overestimated Effect Sizes)

Even if an empirically observed effect does correspond to an effect that actually exists in reality (e.g., there really is a non-zero relation between two variables), it may still be difficult to replicate if it overestimates the size of the actual effect. There are many non-zero relations in reality, and some of those effects are small. (Even many conceptually interesting and important effects have small effect sizes; Abelson, 1985; Prentice & Miller, 1992). Because these effect sizes are small, they are difficult to detect in underpowered studies that are so common within the psychological sciences. But, because of sampling error, small effects may *sometimes* be detected (i.e., judged to be statistically significant) even by underpowered studies; and, when this happens, the observed effect size is almost certain to be larger than the actual effect size in the underlying population (Schmidt, 1996). These overestimated effect size estimates may find their way into the scientific literature but subsequent studies will typically produce smaller—often substantially smaller—effect sizes. Unless these subsequent studies use substantially larger samples than those within which the effects were originally detected, those subsequent studies are unlikely to obtain an effect that is even statistically significant.

Erroneously Broad Effects (Unfounded Assumptions About Generalizability)

Many psychological phenomena are context-contingent. An effect may really occur within some populations, but not within others. An effect may really exist under some circumstances, but not under others. As an enormous body of social psychological research has revealed, an enormous number of ostensibly unremarkable contextual characteristics (e.g., physical dimensions of the space within which an experiment is conducted; Cesario, Plaks, Hagiwara, Navarrete, & Higgins, 2010) can have nontrivial psychological consequences. Effects that are especially "fragile" (i.e., especially context-contingent) are especially difficult to empirically detect, because the empirical circumstances have to be just right in order for the

effects to emerge. But sometimes (either by chance or as a result of some intuitively brilliant stage-managing of experimental procedures) researchers create just the right empirical circumstances that allow these fragile effects to be detected; and these effects find their way into the literature. These effects aren't false positives, and they may not overestimate effect sizes either, but they may still be difficult to replicate. In the absence of evidence or explicit statements bearing on the fragility of these effects, researchers who conduct subsequent studies are likely to tacitly assume that the effects are broader and more generalizable than they actually are. Unless these subsequent studies employ methods that exactly replicate the idiosyncratic context in which the effect was originally detected, these studies are unlikely to replicate the effect. Indeed, because many psychologically important contextual variables may lie outside the awareness of researchers, even ostensibly "exact" replications may fail to create the conditions necessary for a fragile effect to emerge (Stroebe & Strack, 2014).[2]

Psychological Roots of Non-Replicable Effects

Whereas the first two kinds of non-replicable effects (erroneous effects and erroneously big effects) are linked to the empirical tools that scientists use to produce data, the third kind (erroneously broad effects) is an error in the inferences that scientists draw from those data. Also, while the first two errors might be considered errors of commission (the reporting of empirical evidence that is erroneous), the third is more of an error of omission (failure to report either evidence or theory bearing on the fragility of an effect). What all three forms of non-replicability have in common is this: They can be traced, in part, to researchers' unrealistic expectations about the phenomenon under inquiry, and the effects that these expectations have on decisions that researchers make when planning a study, interpreting its results, and communicating those results to the scientific community. Erroneous effects are especially likely to emerge when researchers expect an effect to exist (when it really doesn't), and consequently fail to employ methods that guard sufficiently against the documentation of false-positive inference errors. Erroneously big effects are especially likely to emerge when researchers expect an effect to be bigger than it actually is, and consequently employ underpowered research designs. Erroneously broad effects are especially likely to emerge when researchers expect an effect to be more generalizable than it actually is, and consequently fail to either employ methods that might reveal its context-specificity or to otherwise draw others' attention to its fragility. In essence, all three species of non-replicable findings have their roots in researchers' optimistic tendency to believe than an effect is "better" (truer, bigger, broader) than it really is.

The implication is this: The accuracy and replicability of empirical results depends not merely on the empirical strategies that researchers use when testing hypotheses, but on the conceptual strategies that researchers use when thinking about those hypotheses in the first place.

Informal Approaches to Hypotheses, and Their Problematic Consequences

Like most human beings, researchers have hunches, opinions and beliefs about what they think might be true about the world. These hunches and opinions and beliefs are often expressed as "hypotheses" to be tested by empirical evidence. The means of doing so—in private deliberations and in articles written for public dissemination—is often informal and idiosyncratic, and characterized by (a) the articulation of a research hypothesis in the form of *personal* prediction, and (b) an attempt to *justify* the prediction as plausible.

Personalization and justification of research hypotheses may seem non-problematic, and perhaps even be perceived as exemplary scientific practices. (After all, a hypothesis does require some rationale in order to be perceived as plausible. And if a hypothesis really is nothing more than a scientist's personal prediction, then it would be disingenuous to pretend otherwise.) But both personalization and justification can increase the risk of producing non-replicable effects.

Equating Hypotheses with Personal Predictions

As human beings, we personalize attitudes and beliefs (Abelson & Prentice, 1989). So, as scientists, there is a natural tendency to personalize hypotheses too, treating them as personal expectations, personal possessions, and personal creations. The personalization of hypotheses is tacitly, and sometimes explicitly, encouraged in the training of scientists. Students in research methods courses may be instructed to express scientific hypotheses in the form of personal predictions. Editors may ask authors to more clearly identity what they—the authors—personally predicted. And when pre-registering methodological procedures and analytic strategies, researchers may be encouraged to identify their own personal predictions about how the results might turn out. The personalization of hypotheses is so common, so ingrained within the academic culture, it is easy to overlook its potentially problematic consequences.

One obvious implication arises from the tacit association between the hypothesis and the self. An enormous literature on self-serving biases and motivated reasoning (e.g., Kunda, 1990) suggests that any such personal association may enhance the likelihood for confirmatory bias in the analysis and interpretation of empirical results (MacCoun, 1998; Nickerson, 1998). The problematic consequences of confirmatory bias—particularly consequences for the production of false positives—have been amply discussed (Greenwald, Pratkanis, Leippe, Baumgardner, 1986; Pashler & Harris, 2012; Simmons et al., 2011).

Less amply discussed are the further consequences of treating hypotheses not merely as personal expectations but also as personal *possessions* ("my hypothesis is...") and personal *creations* ("we hypothesized that..."). Research on the psychology of possession (e.g., the endowment effect; Morewedge, Shu, Gilbert, & Wilson, 2009) implies that when researchers perceive hypotheses as possessions they are likely to overvalue it—to perceive that hypothesized effect to be "better" (truer, bigger, broader) than it actually is. These psychological consequences may be amplified by the additional tendency to perceive hypotheses as entities created by researchers themselves. By doing so, researchers may subjectively perceive "their" hypotheses to be a kind of intellectual offspring (Chamberlain, 1890/1965). Because people adopt parental attitudes toward things that merely mimic offspring (Buckels, Beall, Hofer, Lin, Zhou, & Schaller, 2015), the implication is that when researchers personalize hypotheses, they are likely to experience a nepotistic inclination to protect and support them, and are also likely to overestimate the value of hypothesized effects (i.e., to perceive them to be truer, bigger, and broader than they really are).[3]

The further implications are straightforward: When researchers overestimate the veracity of hypothesized effects, they are less likely to make the kind of decisions (in data analytic strategies and subsequent reporting of empirical results) that guard against the documentation of false-positive inference. When researchers overestimate the size of hypothesized effects, they are more likely to employ underpowered research designs—increasing the likelihood that, whenever effects are detected, they are likely to be erroneously big. And when researchers

overestimate the generalizability of hypothesized effects, they are less likely to empirically test its context-specificity or to otherwise draw attention to its potential fragility.[4]

Justifying the Plausibility of a Hypothesis

Rarely do researchers test a hypothesis without first engaging in some sort of appraisal of its plausibility. And rarely do researchers report results bearing on a hypothesis without also providing some conceptual context establishing why that hypothesis was plausible in the first place. It is entirely sensible to appraise the conceptual plausibility of any conceptual hypothesis. But there are different ways of doing so, and not all ways lead to accurate diagnoses.

Because hypotheses are often experienced subjectively as personal beliefs, it is no surprise that the kinds of conceptual appraisals that researchers perform privately (and report publicly) tend to be idiosyncratic and non-systematic. These idiosyncratic appraisals often amount simply to a set of arguments that—based on prior theory and/or prior empirical findings—are designed to *justify* a hypothesis as plausible.

Justification can have benefits, both practical and rhetorical. Unless researchers perceive hypotheses to be at least somewhat plausible, they are unlikely to conduct the sorts of empirical tests necessary for a progressive empirical science. And when conceptual justifications are articulated, they help to provide coherent conceptual context within which empirical results can be more readily understood and interpreted. But there are also potential costs associated with any appraisal of plausibility that focuses primarily on justification.

These costs arise from the psychological consequences of justification, especially when hypotheses are experienced as personal beliefs. When people actively attempt to justify their beliefs—to provide compelling rationales for those beliefs—they come to hold those beliefs even more strongly (Abelson, 1995; Tesser, 1978). The implication is that when researchers articulate a compelling rational designed to justify "their" hypotheses, they may further persuade themselves that the hypothesized effects are "better" (truer, bigger, broader) than they actually are. Consequently, they may be all the more inclined to make the kinds of decisions that increase the likelihood of producing non-replicable results.

These problematic psychological consequences might be avoided if, in addition to addressing the justification-oriented question "Why is this hypothesis likely to be true?" researchers also addressed the more skeptical question "Why is this hypothesis likely to false?" Or if they systematically addressed additional questions such as "How big (or small) is the hypothesized effect realistically likely to be?" and "Under what conditions is the hypothesized effect likely (or unlikely) to emerge?" But those questions aren't compelled by informal plausibility analyses that focus just on justification. In order for researchers to ask—and answer—those additional diagnostic questions, it can be helpful to adopt more rigorous and systematic approaches to the conceptualization of research hypotheses.

Rigorous and Systematic Approaches to Conceptual Hypotheses

Although scientists may subjectively experience hypotheses as their own personal predictions, a more formal approach to scientific inquiry treats scientific hypotheses as statements that have an independent logical status—independent of the scientists who articulate them and test them, and independent of the extent to which scientists personally believe them to

be true or false. Scientific hypotheses are like the ideas and the angels in the Gabriel Garcia Marquez quote that opens this article: They do not belong to anyone.

As Karl Popper suggested (in the other quote that opens this article), a rigorous science distinguishes between personal convictions—"which can never justify any statement"—and the more carefully articulated logical bases of scientific statements. If we take that perspective seriously, then a hypothesis *cannot* attain its status as an actual scientific hypothesis from the mere fact that some scientist stipulates it to be one. Nor can its plausibility be meaningfully diagnosed simply on the basis of idiosyncratic appraisal. Any truly rigorous approach to psychological science requires that scientific hypotheses cannot be equated to personal predictions; hypotheses must instead be articulated as de-personalized products of some systematic analysis, and appraised accordingly.

This isn't a foreign concept. Virtually every aspect of the scientific research enterprise is characterized by the deliberate use of systematic methods to supplement and supersede personal hunches and subjective beliefs. Rather than merely trusting personal hunches about methods that might effectively manipulate or measure psychological constructs, scientists are advised to systematically survey existing empirical literatures in order to inform their methodological decision-making, and to systematically employ additional means (e.g., pilot tests, psychometric analyses) of ascertaining whether seemingly promising methods do, in fact, work. Rather than merely eyeballing raw data and offering subjective impressions about what the numbers might mean, scientists are instructed to employ rigorous statistical analyzes on those data before drawing conclusions. And so forth. The same scientific mindset—prioritizing depersonalized systematic strategies over personal beliefs—can be profitably applied to the articulation of hypotheses too. One way of doing so is to explicitly employ basic principles of logical analysis.

Systematic Use of "If-Then" Logic

Rather than tacitly assuming that "hypotheses" are nothing more than subjectively plausible personal predictions, hypotheses can instead be explicitly articulated as depersonalized statements that follow from the systematic application of "if-then" logic. The basic principles are simple, and intuitively appreciated by most psychological scientists: Some set of underlying assumptions or assertions are specified; and then some set of further implications (e.g., in the form of "if-then" statements) are logically derived. These logically derived implications have the logical status of hypotheses.

Festinger (1954) provides examples of this systematic approach in his classic article on social comparison theory, in which the theory is described in the form of a set of conceptual assertions (which Festinger referred to as "hypotheses" and "corollaries") along with a set of more nuanced hypotheses ("derivations") deduced from those initial assertions. For example, based explicitly on the initial assertions that "There exists, in the human organism, a drive to evaluate his opinions and abilities" and "If the only comparison available is a very divergent one, the person will not be able to make a subjectively precise evaluation of his opinion or ability," Festinger identified the following derivation: "A person will be less attracted to situations where others are very divergent from him than to situations where others are close to him for both abilities and opinions." More recent examples of this approach can be found in an article on the psychological interface of time and motion and its implications (Kruglanski, Pierro, & Higgins, 2015). Kruglanski et al. explicitly identify a set of initial "postulates" and, on the basis of those postulates, they derive a set of additional statements that function as conceptual hypotheses.

By articulating hypotheses as products of an if-then logical analysis, both Festinger (1954) and Kruglanski et al. (2015) avoid the easy temptation to express hypotheses in the form of personal predictions, and instead present them in the form of depersonalized statements about possible relations between psychological variables. These statements do not attain their status as plausible scientific hypotheses simply because some scientists might subjectively believe that they might be true. (The personal beliefs of Festinger and of Kruglanski et al. are irrelevant.) Instead, these statements are plausible scientific hypotheses to the extent that they follow logically from a set of plausible premises.

Just about any psychological hypothesis can be systematically articulated in a similar fashion. Consider, for example, the well-known "negative state relief" hypothesis linking negative mood states to increased helping behavior (Cialdini, Darby, & Vincent, 1973). This hypothesis can be articulated as the product of a logical analysis that proceeds from three underlying assumptions. Two initial assumptions are that people generally prefer to be in good moods rather than bad moods, and that when people are in an undesired psychological state they will be inclined to engage in behavior that has the potential to eliminate that psychological state. Given those assumptions, one may logically derive the statement that if people are in a negative mood, then they will be inclined to engage in behavior that has the potential to eliminate that negative mood state. Here's where a third assumption must be made: engaging in helping behavior has to potential to enhance people's mood state. Given this further assumption, one may logically derive the statement that if people are more likely to engage in behavior that has the potential to eliminate that negative mood state, then they will be more likely to engage in helping behavior. When considered in tandem, the two logical derivations imply the further derivation that if people are in a negative mood, then they will be more likely to engage in helping behavior. This derivation has the logical status as a scientific hypothesis regardless of any scientist's hunches, hopes, or personal investment in its veracity.

This may seem like an excessively formal method of articulating hypotheses that appear intuitively straightforward. But, if one is to avoid the methodological mistakes and inferential errors that can follow from the tacit tendency to equate hypotheses with personal predictions, then some sort of more formal and systematic approach is probably necessary.

If scientists can deliberately avoid thinking about hypotheses as personal predictions, then they are less likely to optimistically assume hypothesized effects to be truer, bigger, and broader than they actually are. Plus, not only does the rigorous deployment of if-then logic help to limit these biases and their problematic consequences, it may also provide a means through which researchers can more realistically appraise the plausibility, size, and generalizability of hypothesized effects. By doing so, researchers will be in a position to make more fully-informed decisions when designing studies, analyzing the results of studies, and drawing conclusions from those results. The exact details of these appraisal strategies would necessarily vary depending upon the research context, but here is a general overview of how this systematic approach to *a priori* hypothesis appraisal might proceed.

Appraisals of Plausibility

When researchers systematically employ the principles of if-then logic to articulate a hypothesis, they are compelled to identify a set of necessary underlying assumptions as well as all the subsequent logical derivations that are required to produce the final hypothesis. By explicitly identifying these assumptions and derivations, researchers are in a position to

systematically consider the plausibility of each and every assumption and logical derivation underlying the hypothesis. Assumptions about psychological phenomena aren't mathematical givens; different assumptions may be held with varying degrees of confidence. (Some assumptions may be assumed to have a high probability of being true because they have been buttressed by vast amounts of prior observation and/or empirical evidence; other assumptions may be somewhat more speculative.) Similarly, the logical implications of those assumptions also tend to be probabilistic. ("If X, then Y" typically means something along the lines of "If X occurs, then there is some non-zero probability that it will lead to Y." There is a lot of variability within that "non-zero" range.) Once they are explicitly estimated, the probability values associated with each assumption and underlying derivation can be considered conjunctively (e.g., by multiplying across probability estimates) to arrive at a systematic appraisal of the logical plausibility of the hypothesis.

On the basis of this kind of systematic appraisal, some hypotheses may be judged to have a high plausibility, whereas others—including many that might otherwise have seemed highly plausible based on the intuitive optimism that accompanies researchers' hopes, hunches, and personal predictions—may be revealed to be less plausible. Researchers should still be encouraged to empirically test hypotheses of uncertain plausibility (science would be dull if only the most highly plausible hypotheses were ever tested, and professional rewards accrue to scientists whose research transcends the obvious); but in doing so they may be more readily compelled to employ empirical methods and analytic techniques that help protect against the production and publication of false positives.

This systematic appraisal of hypotheses need not be merely a matter of private ratiocination. If researchers transparently articulate hypotheses as logical implications of underlying assumptions, then other scientists—collaborators, colleagues, reviewers, editors—are more readily invited to conduct their own independent appraisals of the conceptual plausibility (or implausibility) of hypotheses, and to offer empirical guidance accordingly. This too can help protect against the production and publication of erroneous effects.[5]

Appraisals of Effect Size

The deployment of if-then logic not only provides a systematic means of establishing (and questioning) the possibility that a hypothesized effect might occur, sometimes it may also provide a means of systematically identifying a specific sequence of psychological events that must transpire in order for it to occur. In such cases, each intermediate psychological step can be appraised according to its likely effect size; and this in turn can help researchers to more realistically estimate the effect size of the hypothesized effect itself.

To illustrate, consider the well-known priming effect documented by Bargh, Chen, and Burrows (1996), in which incidental exposure to words such as "Florida," "old," and "bingo" led young adults to subsequently walk more slowly down a corridor. This result has been interpreted as offering support for a hypothesis that, when deconstructed systematically, implies that three underlying psychological events must occur in sequence: (a) Incidental exposure to stimulus words must activate some cognitive representation of a particular social category (elderly people) into working memory; (b) the activation of this social category must trigger the activation of a specific stereotypical expectation associated with that social category (elderly people generally walk slowly); and (c) this stereotypical behavioral expectation must consequently manifest in perceivers' own motor behavior. Each of these intermediate psychological steps can be appraised

according to its plausibility, of course, and this has logical implications for the plausibility of the conceptual hypothesis itself, as described above. (Based on theory and evidence available at the time that this finding was published, the first two steps might be judged to be highly plausible; in contrast, the third step might be judged to a somewhat iffier proposition.) Perhaps even more usefully, each underlying step can be further appraised according to its likely effect size. If indeed incidental exposure to stimulus words does activate some cognition representation of elderly people, how realistically big might that effect be? If indeed activation of this social category also activates the specific stereotypical expectation regarding the slow motor movements of elderly people, how realistically big might that effect be? And if indeed the specific stereotypic expectation manifests in perceivers own walking speed, how realistically big might that effect be? Answers to these questions can be represented numerically in the form of correlation coefficients ($r_{effect\ size}$; Rosnow, Rosenthal, & Rubin, 2000). And—consistent with basic principles of statistical mediation—by multiplying across these values, one may arrive systematically at an estimate of size of the effect predicted by the conceptual hypothesis itself.

It may not be simple or straightforward to arrive at confident estimates of the effect sizes associated with those intermediate steps. It almost certainly requires careful consideration of prior research results bearing on each intermediate step. (Even when there is abundant research evidence available to inform effect size estimates, those prior results may overestimate true effect sizes; Schmidt, 1996.) Sometimes there may be no directly relevant empirical evidence to draw upon, in which case other means might have to be used to arrive at some cruder estimate. (E.g., estimates may be informed by meta-analyses of conceptually related phenomena.) Of course, these difficulties attend any attempt to arrive at an *a priori* estimate of the size of any hypothesized effect—which is one reason why researchers often fail to do so. By explicitly deconstructing a conceptual hypotheses into a set of logically necessary constituent steps—for which there may indeed to relevant prior empirical evidence—the approach identified here provides a means of overcoming this problem, and of doing so in a manner that limits researchers' natural tendency to intuitively assume that effects are bigger than they actually are.

The most obvious benefit of doing so pertains to statistical power. If researchers can make more realistic *a priori* effect size estimates, then they are more likely to consequently make the kinds of methodological decisions (regarding sample sizes, development and use of reliable measures, optical experimental designs, etc.) that provide the power necessary to detect those effects, with the further consequence that when those effects are detected they provide reasonably accurate estimates (rather than over-estimates) of those effect sizes.

Other benefits may also accrue. For instance, this approach may help encourage psychological scientists to use Bayesian statistical analyses. Bayesian analyses are most informative under conditions in which researchers are able to specify exactly what a hypothesis actually predicts—not just whether there is a hypothesized relation between variables but also how big that relation is hypothesized to be (Dienes, 2014). Any conceptual tool that allows researchers to more realistically estimate the size of a hypothesized effect therefore provides researchers greater opportunity to avail themselves of the unique inferential information offered by Bayesian analyses.

Appraisals of Generalizability

Just as one can appraise underlying assumptions and logical derivations according to their plausibility and/or effect size, one can also appraise them according to their likelihood of

generalizing across circumstances and populations. By doing so—and doing so in a rigorous and systematic manner—researchers are more likely to arrive at realistic appraisals of the context-dependence of hypothesized effects, and are less likely to communicate unfounded assumptions about an effect's generalizability.

Consider some of the examples identified above. The hypotheses articulated by Festinger (1954) and Cialdini et al. (1973) are predicated upon assumptions about specific motivational systems and their associated goal states. These assumptions may well be true for many people in many contexts, and so seem intuitively appealing. But they may not be true for all people in all contexts (e.g., some people may be chronically disinclined to evaluate their opinions and abilities; and under some circumstances people may be disinclined to engage in behavior that eliminates an undesired psychological state), with the implications that the hypotheses that follow from these assumptions are also likely to be predictably context-dependent.

Or consider the sequence of psychological events that, hypothetically, explains the effect documented by Bargh et al. (1996). Is that first step (activation of a cognition representation of elderly people) likely to occur under all circumstances and within all populations? Is that second step (activation of the specific stereotypical expectation that elderly people walk slowly) likely to occur under all circumstances and within all populations? Is that third step (manifestation of the stereotypical expectation in perceivers' own walking speed) likely to occur under all circumstances and within all populations? Based on many relevant bodies of psychological theory and research (including research on category activation, multiple sub-types of "elderly" stereotypes, and context-specific effects of temporarily-activated cognitive structures; e.g., Bargh, 1994; Brewer, Dull, & Lui, 1981), realistic answers to those three questions are likely to be, in order, "Yes," "Probably not," and "Almost certainly not." The logical implication of that last answer is that the hypothesized effect is itself almost certainly *not* highly generalizable across contexts. Thus, on *a priori* grounds, the hypothesized effect can be logically diagnosed to be fairly fragile and highly sensitive to specific contextual conditions—a diagnosis that, years later, has been borne out by failures to replicate (Doyen, Klein, Pichon, Cleeremans, 2012) and by empirical identification of moderating variables (Cesario, Plaks, & Higgins, 2006).

These are just illustrative examples. The same approach—systematic appraisal of the extent to which each underlying assumption and each logical implication might be limited to specific populations and specific contexts—can be applied to just about any hypothesis. Doing so can help researchers realistically appraise the fragility of hypothesized effects. These anticipatory insights may motivate empirical studies designed to explicitly document the effects of moderating variables and limiting conditions. And even if those kinds of empirical studies are not immediately forthcoming, any initial evidence supporting the hypothesized effect may be more readily accompanied by a conceptual analysis that draws attention to its context-specificity. Either way, others researchers are less likely to erroneously assume that a fragile effect is more generalizable than it really is.

Additional Considerations

Does the employment of a logically rigorous approach to research hypotheses exclude hunches and hopes and personal predictions from the development of scientific hypotheses? Certainly not! Intuitions and personal beliefs can still play important roles in the process of conceptual discovery, just as they also can be useful when scientists design experiments to test hypotheses. But, just as idiosyncratic inclinations to use particular methods are best

supplemented by more rigorous and systematic approaches to experimental design, personal predictions about psychological phenomena serve science best if they too are supplemented by—and, ideally, superseded by—more systematic means of articulating hypotheses.

Are there limitations associated with a more rigorous and systematic approach to hypotheses? Of course. It isn't a panacea. It's a tool—or a set of tools—and its benefits accrue only when employed in conjunction with the many other tools (which are primarily empirical rather than conceptual) that scientists use when testing hypotheses and drawing conclusions from the results. Also, these conceptual tools are most useful within a hypothesis-testing framework of scientific inquiry; they are less applicable to exploratory or descriptive inquiries, or to purely inductive methods of answering research questions (e.g., Glass & Hall, 2008). This is a modest limitation. Hypothesis-testing remains a prevailing mode of inquiry within the psychological sciences; accordingly, these conceptual tools have a wide range of application.

Are there costs associated with the actual use of these conceptual tools? Perhaps. Hunches, beliefs, and other personal predictions arise relatively effortlessly, whereas the logical articulation of scientific hypotheses requires greater exercise of executive control, greater expenditure of cognitive effort, and more care in describing hypotheses for public consumption. (And it requires more than mere linguistic tinkering. If researchers simply substituted "*the* hypothesis" in place of "*my* hypothesis" in manuscripts, this might amount to little more than a masquerade—dressing up personal predictions in superficially formal attire—that is unlikely to solve the deeper issues identified above.) But these costs are modest. Most psychological scientists are already familiar with, and have an intuitive feel for, the logical principles that underlie the approach described here. All that is required is to use these logical principles in a more thoughtful, deliberative, and systematic way. It may help to follow the examples set by others. I have already identified the useful examples offered by Festinger (1954) and Kruglanski et al. (2015). In rather different context, Wallach and Wallach (1994) provide rigorously detailed logical deconstructions of more than dozen hypotheses on a wide range of social psychological topics. Although Wallach and Wallach's (1994) reason for engaging in this analytic exercise was absurd,[6] their article still usefully illustrates systematic articulations of psychological hypotheses.

Might some researchers be disappointed when their rigorous appraisals reveal that hypothesized effects (even those that are real) are likely to be smaller or more fragile than they—based on their personal hunches and unrealistically optimistic hopes—had tacitly assumed? Perhaps. And if so, might these researchers be discouraged by the practical implications (the need to obtain larger samples, or to more effortfully document the effects of moderating variables, etc.), abandon their plans to test these hypotheses, and seek out other "easier" topics to study instead, thus inhibiting empirical documentation of small or fragile effects? Perhaps some researchers might respond this way, but it seems unlikely to become a widespread problem—or, at a least, no more of a problem than already exists. (For years psychological scientists have known that effect sizes are generally smaller—often much smaller—than we tacitly assume; and there is already renewed awareness of the need for larger samples, more stringent methodologies, and more cautious conclusions. And within social psychology especially, there is widespread recognition that virtually all phenomena of interest vary across populations and contexts.) Of course, in order to ensure that researchers maintain motivation to pursue research on topics such as these, it may be useful for researchers to be reminded regularly that there is an important distinction between the size of an effect and its scholarly value (Prentice & Miller, 1992), that there is enormous scholarly utility in the documentation of moderating variables and limiting

conditions (Greenwald et al., 1986), and that professional rewards tend to accrue to scientists whose research reveals phenomena that are subtle and non-obvious.

In fact, I suspect that widespread adoption of this approach will actually *facilitate* the empirical discovery of subtle and non-obvious phenomena, as a downstream consequence of its capacity to facilitate conceptual discovery of novel hypotheses about these phenomena. For instance, by drawing researchers' attention to specific sequences of psychological events that might transpire in order for a hypothesized effect to occur, it can help researchers discover new hypotheses about mediating mechanisms. Klatzky and Creswell (2014) offer one example of this kind of conceptual discovery in an article that addresses the replicability of the walking-speed effect documented by Bargh et al. (1996). Klatzky and Creswell explicitly identify a series of psychological events that might transpire in order to produce the phenomenon. In doing so, they articulate a novel model of mediating mechanisms (a model distinct from that implied by Bargh et al., and described above), which has implications for priming effects more generally, and which logically implies novel hypotheses that can—and surely will—be tested in future empirical research.

Furthermore, by more explicitly drawing researchers' attention to the potential fragility of hypothesized effects, a systematic approach to the articulation and appraisal of hypotheses can stimulate researchers to formulate additional, more nuanced hypotheses about limiting conditions and moderating variables. These kinds of additional conceptual refinements may emerge over time within any sustained program of research—and, in fact, may be stimulated by failures to replicate previously documented findings (Cartwright, 1973; Dijksterhuis, 2014). But that balky and haphazard route toward conceptual progress is non-optimal. Ideally, that conceptual progress can transpire more swiftly (and without the costly stimulant of a replicability crisis). The approach outlined here may help.

When psychological scientists rigorously apply the principles of if-then logic to the specification of mediating mechanisms and the identification of meaningful moderating variables, they have at their disposal the basic ingredients for articulating more ambitious knowledge structures of the sort that might legitimately be considered theories. Good theories—those that are rigorously constrained by logic while still generating a large number of novel testable hypotheses—are invaluable engines of scientific progress. In contrast to the copious amounts of formal instruction that students receive on empirical and statistical methods, they typically receive very little instruction on *conceptual* methods that can used to discover, develop, and articulate conceptually coherent theories. This pedagogical omission may contribute to a chronically underdeveloped state of theory within the psychological sciences (Klein, 2014; Kruglanski, 2001). This deficit may be overcome through explicit encouragement to employ more logically rigorous and systematic conceptual practices.

Gentle Suggestions for Putting These Principles Into Practice

If there is any merit to the preceding analysis, and if psychological scientists are truly committed to the twofold goal of maximizing the production of novel empirical discoveries while minimizing the production of erroneous and non-replicable results, then some suggestions for scholarly practices logically follow.[7]

As researchers, we would be wise to be wary of the problematic consequences that follow from the natural tendency to treat hypotheses as personal possessions and personal creations. It is best to avoid doing so (not only in professional discourse, but also in our private thoughts). To

aid in that endeavor, it will be helpful to systematically employ the principles of if-then logic—to actually specify underlying assumptions and to rigorously derive the logical implications of those assumptions—as a means of articulating hypotheses. To further assist scientific decision-making, we would be wise to critically appraise each assumption and each logical derivation that underlies any hypothesis. By doing so, we are likely to more realistically diagnose the plausibility of hypothesized effects, as well their effect size and generalizability—all of which usefully inform decisions about empirical methodology, data analysis, and reporting of results.

In order to adopt new habits, researchers typically require the support—and sometimes explicit nudging—from reviewers and editors (Maner, 2014). Therefore, as reviewers and editors we might be wise to attend just as carefully to conceptual rigor as we do to methodological and inferential rigor. We might strive to be more vigilant for "hypotheses" that are simply personal predictions, and for "hypotheses" that lack a transparent logical foundation—and make publication recommendations accordingly.

Undergraduate instruction in the psychological sciences might sensibly be revised so that it more effectively disabuses students of the natural (but problematic) tendency to think that a scientific hypothesis is no different than a personal prediction. When we teach research methods to undergraduates, it would be helpful to instruct students on the logical distinction between a formal scientific hypothesis and an idiosyncratic personal belief about the veracity of hypothesis, and to explain how a failure to make that distinction can lead to problematic consequences. It would also be beneficial to explicitly introduce students to the basic principles of if-then logic, and to show them how these principles can be used to transform intuitively appealing personal predictions into actual scientific hypotheses. Research methods textbooks might sensibly be revised accordingly.

The training of graduate students might also be enhanced through greater attention to these issues. In the context of providing formal methodological training to graduate students, it will be useful to provide dedicated instruction not only on empirical tools that can be used to collect and analyze data, but also on conceptual tools that can be used to discover, articulate, and appraise research hypotheses in the first place. Relevant graduate courses might include readings that attend explicitly to the logical principles underlying theory development and theory-testing in the psychological sciences (e.g., Gawronski & Bodenhausen, 2015), that provide examples of these logical principles put to practical use in the systematic articulation of testable hypotheses (e.g., Kruglanski et al., 2015), and that highlight the various benefits of doing so.

Of course, in order for rigorous conceptual practices to become habitual, graduate students need to actually practice them. As teachers and mentors, we can help them to be more vigilant of the problematic tendency to equate personal predictions with scientific hypotheses, and to overcome that tendency with a more depersonalized, deliberative and logically rigorous approach. (For example: "Yes, I appreciate the appeal of that prediction. But it still seems like it's a kind of informal hunch, and I'm struggling to connect all the underlying logical dots. How about if you pretend to be Leon Festinger for few minutes: Show me exactly the assumptions that underlie this hypothesis, and show me exactly the logical derivations that follow from those assumptions, and we'll see if, in fact, the hypothesis logically emerges. And then we'll use that set of assumptions and derivations to more cynically appraise the hypothesis and see where that takes us..."). We might also be wise to provide guidance on how to respond artfully when well-meaning others tacitly invite them to backslide. (For example: "When someone asks you what *your* hypothesis was, do you respond by telling them what your personal prediction was? I hope not. Ideally, you'll say something far more impressive. Something like: 'You asked me what *my*

hypothesis was. But, of course, as a scientist, I'm not designing studies to test my personal predictions. I'm designing studies to test logically plausible hypotheses that have their status as hypotheses independent of what I—or you, or any of us in this room—might believe to be true. And so, to answer your question, I'm going to tell you about not just one, but two hypotheses, both of which are tested by the results I'm about to show you. Both hypotheses seem plausible—although, for reasons I'll describe in a moment, one of them seems somewhat more logically plausible than the other—but neither is so obvious as to be uninteresting. Also, these two different hypotheses are predicated upon two distinct sets of underlying assumptions and, as a consequence, they logically imply different patterns of results; and it's for that reason that I think the results of this study are especially informative. Anyway, let me tell about these two rival hypotheses...'")

Whether one is temporarily inhabiting the role of a researcher, reviewer, editor, teacher, or mentor, it will be useful to remember that, regardless of what researchers are individually inclined to do, the peer-reviewed discourse of the discipline provides a mechanism through which researchers are regularly reminded of their responsibilities to other researchers. Just as there are costly cumulative consequences of skimpy and selective reporting on empirical methods, so too there are cumulative costs associated with idiosyncratic presentation of conceptual hypotheses. So, just as we are advised to describe empirical methods in ways that are fulsome and transparent (Brown, Furrow, Hill, Gable, Porter, & Jacobs, 2014; Eich, 2014), we also would be well advised to articulate conceptual hypotheses in ways that transparently reveal their underlying logic, so that our audience can independently appraise the plausibility, size, and fragility of the hypothesized effects—and make their own decisions accordingly.

Above all, as psychological scientists, we would be wise to be mindful of the fact that what distinguishes scientific from non-scientific inquiry is the use of methodologies that minimize the problematic intrusion of subjective perceptions and personal beliefs. In order to minimize the production of erroneous results and to maximize the discovery of the many fascinating psychological phenomena that really exist, our empirical methods need to be rigorous and systematic. Our conceptual methods probably should be too.

Author Note

Preparation of this article was supported by Insight Grant #435- 2012-0519 from the Social Sciences and Humanities Research Council of Canada. The article also benefited from generous comments on an earlier draft from Andre Beukers, Jeremy Biesanz, Elizabeth Dunn, Bertram Gawronski, Klaus Fiedler, Marlise Hofer, Arie Kruglanski, and Jessica Tracy.

Footnotes

1. These scientific goals correspond to the motivational distinction between an “eager” and a “vigilant” orientation toward any ongoing activity (Higgins, 2005). Just as there is a psychological tension between eagerness and vigilance, so too there is a tension between the promotion of novel scientific discoveries and the prevention of erroneous results. Systematic initiatives designed to facilitate publication of bold new findings often have the collateral consequence of increasing the likelihood that erroneous—and non-replicable—results find their way into print (an increase in “false positive” errors). Conversely, systematic initiatives designed to inhibit the publication of false positive errors often have the collateral consequence

of increasing the likelihood that novel new phenomena—which actually do exist—go undetected and unreported (i.e., an increase in "false negative" errors).

2. For example, in recent years there has been much hand-wringing over the fact that some well-known priming effects have proven difficult to replicate. Surely one reason for this is that the original effects overestimated the actual effect size. But another reason is that priming effects (even the big ones) are fragile. Because of the psychological processes through which priming effects actually occur, any particular effect is contingent upon specific cognitions that are already active in working memory—cognitions that are likely to differ depending the population from which a study sample was drawn and on the many incidental elements of the perceptual environment within which a study is conducted (Cesario, 2014; Gawronski & Cesario, 2013; Klatzky & Creswell, 2014; Loersch & Payne, 2011). Ironically, it is exactly because of the ubiquity of priming effects that any *specific* priming effect is difficult to replicate.

3. Here is how Chamberlin (1890/1965, p. 755) described the parental affection that scientists feel for favored hypotheses: "The moment one has offered an original explanation for a phenomenon which seems satisfactory, that moment affection for his intellectual child springs into existence; and as the explanation grows into a definite theory, his parental affections cluster around his intellectual offspring, and it grows more and more dear to him... Instinctively there is a special searching-out of phenomena that support it, for the mind is led by its desires. There springs up, also, an unconscious pressing of the theory to make it fit the facts, and a pressing of the facts to make them fit the theory. When these biasing tendencies set in, the mind rapidly degenerates into the partiality of paternalism."

4. To a limited extent, these problems might be addressed by the deliberate use of the research strategies advocated as useful correctives to conformation biases that creep into a hypothesis-testing approach to science. One example is a "condition-seeking" strategy (Greenwald, Pratkanis, Leippe, Baumgardner, 1986) in which, rather than testing the veracity of a hypothesized effect, researchers instead try to systematically identify the conditions under which the effect does and doesn't occur. Another corrective strategy is offered by a "strong inference" approach (Platt, 1964), in which researchers specify competing conceptual hypotheses and design "crucial experiments" that might support one while disconfirming the other. But these corrective strategies only work if researchers use them; and researchers may be especially unlikely to employ these strategies when testing personal predictions. When researchers are personally invested in the veracity and generalizability of a hypothesis, they are disinclined to systematically identify the conditions under which it's wrong. And if a researcher overestimates the veracity, size, and generalizability of a hypothesized effect, this overvaluation may pose a psychological barrier to the researcher's inclination or ability to identify additional hypotheses that contradict it.

5. In addition, some alleged hypotheses—most likely those that are based on personal hunches, hopes, guesses, or flights of fancy—may be revealed not to have the status as scientific hypotheses at all, because they cannot be logically deduced from any remotely plausible set of assumptions. Consequently, even if these "hypotheses" were empirically tested (and supported), the results might be perceived to be of dubious scientific merit. If so, it would be even more difficult than it already is to publish transparently preposterous findings (of the sort highlighted by Simmons et al., 2011) and provocatively inexplicable results (of the sort reported by Bem, 2011) in prominent scientific journals.

6. Wallach and Wallach (1994) did *not* perform this exercise as a means of diagnosing the hypothesized effects' potential implausibility, small size, or fragility. Instead, they did so as a

means of buttressing the thesis that many social psychological hypotheses are so unassailably true that it serves no purpose to actually test them—an assertion that, in this time of collective concern over the perceived prevalence of false positives, seems even more hilariously misguided now than it did then (Schaller, Crandall, Stangor, & Neuberg, 1995; Schaller & Crandall, 1998).

7. I admit to some squeamishness in summarizing these suggestions. Not because the suggestions are silly (they aren't!) but because—in the "anything goes" spirit of Feyerabend (1975)—I personally believe that a diversity of methods is important to any healthy and progressive science, and so I am leery of methodological recommendations that have an overtly prescriptive or proscriptive tone. But that's just me. It would be professionally irresponsible to let my own personal half-baked philosophical beliefs prevent me from conveying a set of practical suggestions that appear to follow straightforwardly from the preceding analysis and that, if implemented, will almost certainly be beneficial.

References

Abelson, R. P. (1985). A variance explained paradox: When a little is a lot. *Psychological Bulletin, 97,* 129-133.

Abelson, R. P. (1995). Attitude extremity. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 25-42). Mahwah NJ: Erlbaum.

Abelson, R. P., & Prentice, D. A (1989). Beliefs as possessions: A functional perspective. In A. R. Pratkanis, S. J. Breckler, & A. G. Greenwald (Eds.), *Attitude structure and function* (pp. 361-381). Hillsdale NJ: Erlbaum.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108–119.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition* (2nd ed., pp. 1-40). Hillsdale, NJ: Erlbaum.

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*, 230-244.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407-425.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., et al. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224.

Brewer, M. B., Dull, V., & Lui, L. (1981) Perceptions of the elderly: Stereotypes as prototypes. *Journal of Personality and Social Psychology,* 41, 656-670.

Brown, S. D., Furrow, D., Hill, D. F., Gable, J. C., Porter, L. P., & Jacobs, W.J. (2014). The duty to describe: Better the devil you know than the devil you don't. *Perspectives on Psychological Science, 9*, 626-640.

Buckels, E. E., Beall, A. T., Hofer, M. K., Lin, E., Zhou, Z., & Schaller, M. (2015). Individual differences in activation of the parental care motivational system: Assessment, prediction, and implications. *Journal of Personality and Social Psychology, 108*, 497-514.

Cartwright, D. (1973). Determinants of scientific progress: The case of research on the risky shift. *American Psychologist, 28*, 222-231.

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science, 9*, 40-48.

Cesario, J., Plaks, J. E., Hagiwara, N., Navarrete, C. D., & Higgins, E. T. (2010). The ecology of automaticity: How situational contingencies shape action semantics and social behavior. *Psychological Science, 21*, 1311-1317.

Cesario, J., Plaks, J. E., & Higgins. E. T. (2006). Automatic social behavior as motivated preparation to interact. *Journal of Personality and Social Psychology, 90*, 893-910.

Chamberlin, T. C. (1890/1965). The method of multiple working hypotheses. *Science, 148*, 754-759. (Originally published in *Science* in 1890).

Cialdini, R. B., Darby, B. L., & Vincent, J. E. (1973). Transgression and altruism: A case for hedonism. *Journal of Experimental Social Psychology, 9*, 502-516.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. doi:10.3389/fpsyg.2014.00781

Dijksterhuis, A. (2014). Welcome back theory! *Perspectives on Psychological Science, 9*, 72-75.

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE, 7*, e29081. doi: 10.1371/journal.pone.0029081

Eich, E. (2014). Business not as usual. *Psychological Science, 25*, 3-6.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*, 117-140.

Feyerabend, P. (1975). *Against method*. London UK: New Left.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science, 7*, 661-669.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology, 108,* 275-297.

Funder, D. C., Levine, J. M., Mackie, D. M., Morf, C. C., Vazire, S., & West, S. G. (2014). Improving the dependability of research in personality and social psychology: Recommendations for research and educational practice. *Personality and Social Psychology Review, 18*, 3-12.

García Márquez, G. (1995). *Of love and other demons*. New York: Knopf. (Originally published as *Del amor y otros demonios*. Barcelona, Spain: Mondadori. Translated by Edith Grossman.)

Gawronski, B., & Bodenhausen, G. V. (2015). *Theory and explanation in social psychology*. New York: Guilford Press.

Gawronski, B., & Cesario, J. (2013). Of mice and men: What animal research can tell us about context effects on automatic responses in humans. *Personality and Social Psychology Review, 17*, 187-215.

Glass, D. J., & Hall, N. (2008). A brief history of the hypothesis. *Cell, 134*, 378-381.

Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review, 93*, 216-229.

Higgins, E. T. (2005). Value from regulatory fit. *Current Directions in Psychological Science, 14*, 209-213.

Klatzky, R. L., & Creswell, J. D. (2014). An intersensory interaction account of priming effects—and their absence. *Perspectives on Psychological Science, 9*, 49-58.

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory and Psychology, 24*, 326-338

Kruglanski, A. W. (2001). That 'vision thing': The state of theory in social and personality psychology at the edge of the new millennium. *Journal of Personality and Social Psychology, 80*, 871-875.

Kruglanski, A. W, Pierro, A., & Higgins, E. T. (2015). Experience of time by people on the go: A theory of the locomotion-temporality interface. *Personality and Social Psychology Review, 19*, ___-___.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480-498

Loersch, C., & Payne, B. K. (2011). The situated inference model: An integrative account of the effects of primes on perception, behavior, and motivation. *Perspectives on Psychological Science, 6*, 234-252.

MacCoun, R. J. (1998). Biases in the intrepretation and use of research results. *Annual Review of Psychology, 49,* 259-287.

Maner, J. K. (2014). Let's put our money where our mouth is: If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science, 9*, 343-351.

Maner, J. K. (this issue). Into the wild: Field research can increase both replicability and real-world impact. *Journal of Experimental Social Psychology*.

Morewedge, C. K., Shu, L. L., Gilbert, D. T., & Wilson, T. D. (2009). Bad riddance or good rubbish: Ownership and not loss aversion cause the endowment effect. *Journal of Experimental Social Psychology, 45*, 947-951.

Murayama, K, Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false positive rates. *Personality and Social Psychology Review, 18*, 107-118.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175-220.

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

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). DOI: 10.1126/science.aac4716

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*, 531-536.

Popper, K. (1959 / 2005). *The logic of scientific discovery*. New York: Routledge. (Originally published in 1934 as *Logik der Forschung*. Vienna, Austria: Verlag von Julius Springer.)

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112*, 160-164.

Platt, J. R. (1964). Strong inference. *Science, 146*, 347-353.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11*, 446-453.

Sakaluk, J. K. (this issue). Exploring small, confirming big: An alternative system to The New Statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology*.

Schaller, M., & Crandall, C. S. (1998). On the purposes served by psychological research and its critics. *Theory and Psychology, 8*, 205-212.

Schaller, M., Crandall, C. S., Stangor, C., & Neuberg, S. L. (1995). "What kinds of social psychology experiments are of value to perform?" Comment on Wallach and Wallach (1994). *Journal of Personality and Social Psychology, 69*, 611-618.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115-129.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., et al. (this issue). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*.

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

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychologial Science, 9*, 59-71.

Tesser, A. (1978) Self-generated attitude change. *Advances in Experimental Social Psychology, 11,* 289–338.

Wallach, L., & Wallach, M. A. (1994). Gergen versus the mainstream: Are hypotheses in social psychology subject to empirical test? *Journal of Personality and Social Psychology, 67*, 233-242.