Check for updates

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

What Constitutes Strong Psychological Science? The (Neglected) Role of Diagnosticity and A Priori Theorizing

Perspectives on Psychological Science
2017, Vol. 12(1) 46–61
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691616654458
journals.sagepub.com/home/pps
SAGE

Klaus Fiedler
University of Heidelberg, Germany

Abstract

A Bayesian perspective on Ioannidis's (2005) memorable statement that "Most Published Research Findings Are False" suggests a seemingly inescapable trade-off: It appears as if research hypotheses are based either on safe ground (high prior odds), yielding valid but unsurprising results, or on unexpected and novel ideas (low prior odds), inspiring risky and surprising findings that are inevitably often wrong. Indeed, research of two prominent types, sexy hypothesis testing and model testing, is often characterized by low priors (due to astounding hypotheses and conjunctive models) as well as low-likelihood ratios (due to nondiagnostic predictions of the yin-or-yang type). However, the trade-off is not inescapable: An alternative research approach, theory-driven cumulative science, aims at maximizing both prior odds and diagnostic hypothesis testing. The final discussion emphasizes the value of pluralistic science, within which exploratory phenomenon-driven research can play a similarly strong part as strict theory-testing science.

Keywords

sexy-hypothesis testing, model testing, theory-driven cumulative science, phenomenon-driven research

What constitutes strong psychological science? What are the criteria of compelling future research that those involved in the current discourse about nonreproducible findings, questionable researcher practices, and inappropriate statistical analyses are seeking? What are the ideals or best exemplars of scientific inquiry that deserve to be embraced and imitated?

One need not conduct a representative survey to anticipate the scientific community's consensual answers: Strong research must rely on sufficiently large samples that allow for powerful statistical tests of precisely predicted relationships, minimizing false positives and failures to replicate. In other words, good research findings should be replicable, reflective of a true effect, and based on state-of-the art statistical analyses. Any participant in the current discourse who disagrees with these widely shared positions hardly would be taken seriously.

Philosophers and historians of science might disagree with this consensus, arguing that (a) nonreplication (of an old assumption) is the source of all scientific progress, (b) empirical evidence never reflects the plain truth, and (c) the most highly developed sciences make do without statistics. However, psychologists tend to discard these conjectures as outside perspectives from nonexperts who are not knowledgeable about the reality of psychological science. So, without much contemplation, many psychologists continue to believe that strong science must be built on a strict selection of true and robust findings (Schmidt, 2010; Simmons, Nelson, & Simonsohn, 2011; Verhagen & Wagenmakers, 2014) obtained with new statistics (Cumming, 2014) applied to data samples that must not be underpowered.

In this article, I try to show that philosophers and theoreticians (Earp & Trafimow, 2015) who call for deeper reflection beyond statistical hypothesis testing may not be fully mistaken or too remote from scientific reality. The point here is not to argue that replication, reliability, and statistical analyses are worthless but that these technical issues are subordinate to more fundamental issues of research design and logic of science. Although in this article I am critical and quite in the spirit of the quest for

Corresponding Author:
Klaus Fiedler, Department of Psychology, University of Heidelberg, Hauptstrasse 47-51, 69117 Heidelberg, Germany.
E-mail: kf@psychologie.uni-heidelberg.de

stringency and scrutiny, I arrive at several liberal and broadminded recommendations about how to improve psychological science and how to conserve its current assets. These recommendations clearly deviate from the popular notion that strictly controlled research practices and new statistics afford the major keys to better science (Cumming, 2014; Simmons et al., 2011).

The article is organized as follows. Starting with a discussion of Ioannidis's (2005) memorable statement that "Most Published Research Findings Are False," I first outline a Bayesian analysis of the fundamental dilemma of research that is supposed to be original and surprising but at the same time predictable and replicable. How can science be both surprising and predictable, novel and theoretically clear, ground breaking and replicative of prior evidence? In Bayesian notation, two factors determine the posterior odds that a hypothesis is correct in the light of empirical evidence: first, the prior odds that the hypothesis follows from sound theorizing and prior evidence and, second, the likelihood ratio that indicates the diagnosticity with which the evidence supports the focal hypothesis rather than alternative hypotheses.

Then, with reference to two prominent research approaches—sexy hypothesis testing and model testing—I discuss why Ioannidis's pessimistic conclusions about the modest rate of correct hypotheses are to some degree justified. Because both approaches strive for originality and surprise value, rather than certainty and safety, they focus on unexpected, a priori unlikely, and conjunctive hypotheses tested in simplified designs. As a consequence, both the prior odds and the diagnosticity tend to be low, and so the posterior odds must be inevitably low as well. However, I also outline an alternative research approach, theory-driven cumulative science, which holds the promise of ideally maximizing both the prior odds and the diagnosticity of empirical science. In a plea for pluralistic science, I finally argue that good research need not obsessively optimize posterior odds. Rather, strong and responsible science must sometimes also dare to tackle risky and uncommon ideas that are by definition rarely true but, if they are true, entail the potential for ground-breaking innovations.

Explaining the Success Rate of Research Hypotheses

Ioannidis's (2005) memorable and provocative allusion to false positives in the published research literature highlighted the need to understand the reasons that the empirical results obtained in psychological science must be interpreted with caution. I will show that Ioannidis's pessimistic summary statement is both admittedly true and apparently wrong, depending on the quality of the hypothesis and the empirical evidence.

The predictive value of psychological hypotheses

Figure 1 depicts a telling example adopted from Diekmann (2011). Let the probability (*P*) that a hypothesis H_1, drawn at random from a universe of 10,000 hypotheses, is true to be as low as .04; the complementary probability that H_1 is wrong is 1 − .04 = .96. That is, 400 hypotheses are correct and 9,600 are wrong. Assuming a statistical power of $(1 - \beta) = .80$ and an error probability (α) of .05, this amounts to expecting 320 significant results for true hypotheses (320 = .80 × 400) along with 480 false positives (480 = .05 × 9,600) for wrong hypotheses. Thus, assuming a theory as weak as *P* = .04, the "truth proportion" (*TP*) of all obtained significant effects is indeed less than one half: *TP* = 320/(320 + 480) = .40 (see upper part in Fig. 1);[1] more than half of the observed effects originate in a false H_1.

However, it is also evident from the middle and lower part that *TP* rises quickly when the a priori likelihood of H_1 being true increases to slightly more solid but still modest values of *P* = .20 (*TP* = .80) or *P* = .40 (*TP* = .91). Thus, if the a priori likelihood of correctly predicting an empirical outcome is only .20 or .40 (rather than .04), then *TP* increases to .80 or even .91, reflecting mostly "correct findings." Thus, improving the a priori value of theoretical hypotheses affords an effective means of overcoming Ioannidis's problem.

A Bayesian analysis of the tradeoff between predictable and informative research

Note that this way of deriving *TP* from *P* and $(1 - \beta)/\alpha$ suggests a straightforward application of Bayesian probability calculus, which is commonly presented in an odds format: The posterior odds ($\Omega_{posterior}$) that a research hypothesis H_1, rather than the null hypothesis H_0, is true, given the data obtained in a study are the product of the prior odds (Ω_{prior}) that H_1 versus H_0 is true on a priori grounds times the likelihood ratio (LR)—that is, the likelihood of obtaining the data, given H_1 divided by the likelihood of the data given H_0. Thus,

$$\Omega_{posterior} = LR \times \Omega_{prior,} \quad \text{or}$$

$$\frac{p(H_1 \text{ true} \mid \text{data})}{p(H_0 \text{ true} \mid \text{data})} = \frac{p(\text{data} \mid H_1 \text{ true})}{p(\text{data} \mid H_0 \text{ true})} \times \frac{p(H_1 \text{ true})}{p(H_0 \text{ true})} .$$

In Bayesian notation, the a priori probability $p(H_1 \text{ true})$ and the a posteriori probability $p(H_1 \text{ true} \mid \text{data})$ replace the respective terms *P* and *TP* used in Figure 1. The likelihood ratio LR reflects the diagnosticity of a test, that is, the

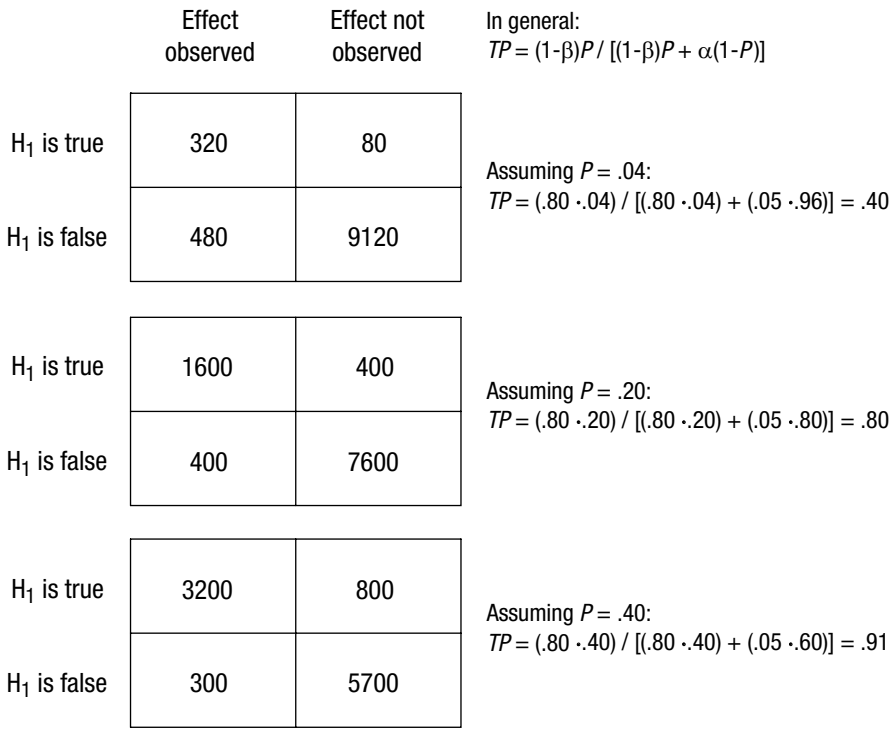

Fig. 1. Numerical illustration of the expected truth proportion (*TP*) of empirical studies as a function of the a priori probability (*P*) that H_1 is true, assuming error probability (α) of .05 and statistical power ($1 - \beta$) of .80.

ability of empirical data to discriminate between H_1 and H_0. Note that in reality LR is not solely a function of statistical power ($1 - \beta$) and type α error in a statistical test (as in Fig. 1); LR also depends on nonstatistical factors, as will soon be apparent. Note also that a Bayesian perspective highlights the fact that the a priori odds in support of the focal hypothesis, $\Omega_{prior} = p(H_1 \text{ true})/p(H_0 \text{ true})$, rely to a large extent on nonempirical arguments, such as strong theorizing and logical derivation.

Obviously, then, the truth proportion *TP* of research hypotheses depends on both Ω_{prior} and LR (estimated as $(1 - \beta)/\alpha$). The critical debate of poor *TP* rates focuses on low statistical power and high error probability, which jointly reduce the statistical likelihood ratio $(1 - \beta)/\alpha$. Critics assume that small samples and unwarranted research practices reduce statistical power ($1 - \beta$) to less than .80 and violations of statistical assumptions lead to effective α levels higher than .05. From a Bayesian analysis, though, it is obvious that theoretical rigor leading to higher Ω_{prior} can easily compensate for reasonable LR decreases in $(1 - \beta)/\alpha$. Doubling *P* (e.g., from .04 to .08) more than doubles Ω_{prior} (from .041 = .04/.96 to .087 = .08/.92) and thus compensates for a decrease in $(1 - \beta)/\alpha$ by more than one half. Reasonable improvements in theoretical rigor (i.e., in Ω_{prior}) can thus be more effective than efforts to control $(1 - \beta)/\alpha$. Increasing statistical rigor is by no means the only way to enhance *TP*.

Most important, a Bayesian perspective reveals a fundamental trade-off, which directly leads to the question of what constitutes strong psychological research. If strong research means maximizing TP, then researchers must refrain from testing risky hypotheses with low Ω_{prior}. As a high Ω_{prior} is a necessary (conjunctive) condition for maximal $\Omega_{posterior}$, TP can only be maximized when a priori theories already support the hypotheses to be tested. Risky hypothesis tests must be avoided, and research must be confined to safe situations, in which expected findings are consistent with high Ω_{prior}. Innovative and ground-breaking research inspired by risky hypotheses inevitably reduce TP; "new discoveries will continue to stem from hypothesis generating research with low or very low pre-study odds" (Ioannidis, 2005, p. 701).

Facing this apparent trade-off between cautious (boring) science warranting high *TP* and courageous (innovative) science leading to low *TP*, one is tempted to adopt the pessimistic conclusion that psychologists have to make a forced choice between either solidity or risk, either conservatism or progress. Scientists would be condemned either to be very cautious and avoid testing exciting hypotheses or, if they dare to test risky hypotheses, to help decrease *TP*. To keep *TP* high, they would have to refrain from studying such exciting issues as the impact of disgust stimuli on immune reactions (Schaller & Park, 2011), the genesis of false confessions (Kassin,

2008) and spectacular false memories (Shaw & Porter, 2015), or the reduction of prejudice via auditory stimulation during sleep (Feld & Born, 2015).

A Bayesian Look at Two Prominent Research Approaches

Before turning to possible solutions of the dilemma, though, let us first elaborate on reasons why both LR and Ω_{prior} are often conspicuously low. Let us particularly examine why, first, the diagnosticity of a hypothesis test (LR) is not solely determined by α and $(1 - \beta)$ and why, second, purely theoretical and logical factors have a strong impact on Ω_{prior}. Both points will become evident from a critical discussion of two prominent research approaches, which often fail to inform strong scientific inferences due to low LR and Ω_{prior}.

Sexy-hypothesis testing

The first of these two research approaches may be called *sexy-hypothesis testing*. It is reflected in many articles published in prominent journals and in common textbooks. Typical of such research is the focus on elementary hypotheses about the impact of a single causal factor on a single dependent measure: Guilt serves to increase risky decisions (Kouchaki, Oveis, & Gino, 2014); unrecognized stimuli trigger implicit learning (Hannula & Ranganath, 2009); sadder people are wiser (Alloy & Abramson, 1979). The predicted unicausal relation is typically not specified quantitatively but confined to a binary outcome; some measure of performance is predicted to increase or decrease (e.g., responses become faster or slower, satisfaction gets higher or lower, people show approach or avoidance responses). Given only two outcomes, the information gained in a study amounts to no more than to one bit; the result is not more informative than the outcome of tossing a coin. The third possibility that the true relation is exactly zero is negligible because "everything is somewhat correlated with everything" (Meehl, 1990, p. 108), suggesting that H_0 is never literally true. However, even if we allow for H_0, one can hardly see why the rate of correct elementary hypotheses should be as low as $P = .04$ that is needed to explain Ioannidis's (2005) pessimistic estimate. By chance alone, the likelihood of binary hypotheses to be true ought to be in the range of $.40 < p < .50$. Any modest theory that is better than chance or $P = .50$ in coin tossing should further enhance the accuracy of scientific predictions to even higher levels.

So why should one assume $P = .04$? How can P derived from empirical replication rates fall markedly below chance? One sensible answer is that the sexy hypotheses being tested in this approach do not constitute a random sample of all possible hypotheses. Such studies focus on unlikely, surprising outcomes rather than on ordinary and commonly expected outcomes. What makes hypotheses "sexy" is their focus on the counterintuitive outcome of a binary question ("sadder but wiser," not "sadder and mistaken"; Alloy & Abramson, 1979). They entail extraordinary and astounding predictions. In Bayesian terms, this preference for surprising and extraordinary hypotheses—which may lead to exciting insights when supported but many negative empirical results otherwise—serves to keep Ω_{prior} at a low level.

At the same time, the limited information value of elementary (binary) hypotheses also restricts their diagnosticity (LR). In a multicausal world, in which virtually all effects can be influenced by several causal factors, a mere upward or downward shift in a dependent measure can hardly provide unequivocal evidence for only one hypothesis focusing on a single causal factor. If a balloon rises up into the sky rather than falling down, this does not invalidate the law of gravity; the balloon's behavior depends on other factors (e.g., specific weight, temperature of gas). One should neither discard the gravitation hypothesis when the balloon rises nor should one interpret downward movement as gravitation proof. The seeming support or nonsupport might reflect the influence of other causes or enabling conditions acting in the same or in opposite direction (Goldvarg & Johnson-Laird, 2001).

The same multicausality problem holds in psychological science. Even in a randomized design, hardly any experimental treatment represents a pure manipulation of the focal independent variable. For instance, in a test of the hypothesis that a cheater-detection motive enhances memory for faces, a subset of faces is presented together with a scenario related to cheating (Nairne, Pandeirada, & Thompson, 2008). However, such a manipulation also may influence a number of other causal factors unrelated to cheating, such as negative affect, affective involvement, depth of processing, or self-reference (Bell & Buchner, 2012; Klein, 2012). Given multiple correlated causes (Fiedler, Kutzner, & Krueger, 2012; Wason, 1960), a successful test of an elementary (binary) hypothesis rarely provides strong diagnostic evidence only for the focal hypothesis. It is rather compatible with two or more hypotheses at the same time. Likewise, a failure to obtain the predicted outcome rarely provides unequivocal evidence against the focal hypothesis. As a rule, empirical hypotheses of the yin-or-yang type, which merely predict one out of two binary outcomes or maybe a single cell in a 2×2 design, rarely yields a strong LR. The diagnosticity of such evidence must remain low.

For the reasons depicted here, the search for extraordinary findings in the sexy-hypothesis testing approach serves to keep both P and Ω_{prior} at a systematically low

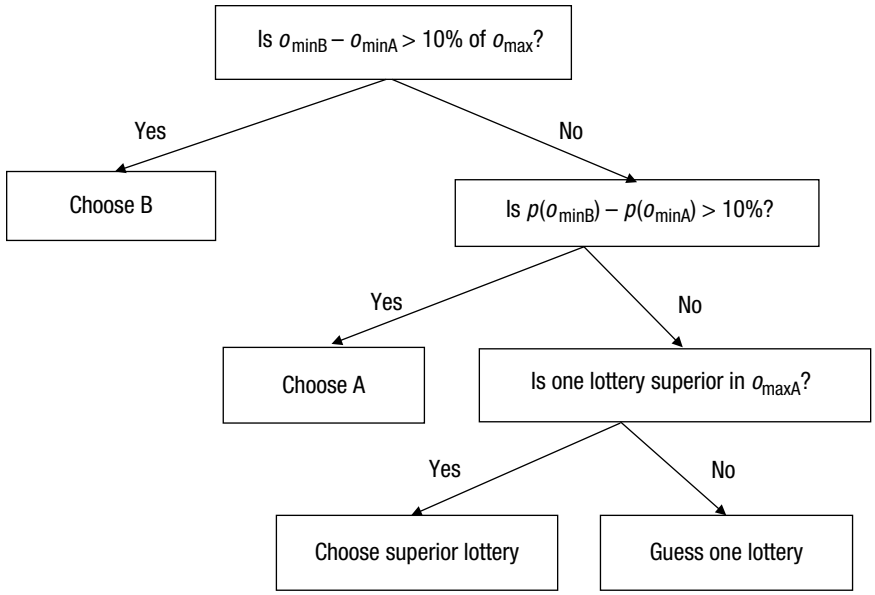

Fig. 2. Illustration of the priority heuristic (Brandstätter, Gigerenzer, & Hertwig, 2006; Fiedler, 2010).

level. An obvious conclusion, then, is that strong and confident scientific inferences (i.e., high posterior odds $\Omega_{posterior}$) are unlikely to be obtained in the sexy-hypothesis testing approach.

Model-testing approach

In an attempt to overcome the simplicity and the crude qualitative level of sexy-hypothesis testing, the goal of another research approach is to attain quantitative precision and commitment to clearly specified models that can be tested strictly. The models can be quite complex and sophisticated, anchored in subsymbolic layers of mental representations, motor functions, or neurological substrates. So the hypotheses of the model-testing approach specify quite refined algorithms or complex functional relationships, in contrast to most sexy hypotheses. The postulated causal mechanisms are refined and complex, calling for high precision and behavioral predictions under clearly spelled-out conditions.

However, despite this apparent contrast, an examination of model-testing research reveals similar restrictions. Again, both Ω_{prior} and LR often remain low, restricting the posterior odds of scientific inferences that can be reached in model-testing studies. On one hand, the assumptions of algorithmic process models are often too strong, complex, and nonparsimonious to render Ω_{prior} high on a-priori grounds. Complexity is inversely related to parsimony (see also Higgins, 1992). On the other hand, if a focal model predicts the behavioral evidence obtained in a study pretty well, this does not rule out that other models (often making fundamentally different assumptions) can also account for the same evidence. As a consequence, the diagnosticity (LR) of model-testing studies is also restricted; most empirical results can be predicted from fundamentally different models.

Algorithmic process models. To illustrate this point, let us consider a prominently published model of decision making under risk[2] that is widely respected for its precision and testability, the priority heuristic (Brandstätter, Gigerenzer, & Hertwig, 2006). It is summarized in Figure 2. The priority heuristic assumes that when making choices between two lotteries or decision options A and B, individuals in a first stage consider only the worst outcomes, o_{minA} and o_{minB}, and select one option only if its worst outcome is superior by at least 10% of the overall best outcome o_{max}. Only if no decision can be reached by this primary criterion will the focus then be on the probabilities of both options' worst outcome, $p(o_{minA})$ and $p(o_{minB})$. A decision will be made in favor of the option with the lower probability of a worst outcome but only if $p(o_{minA})$ and $p(o_{minB})$ differ by at least 10%. Otherwise, a third stage will be sensitive only to o_{max}; individuals will select the alternative with the higher best outcome or, if o_{maxA} and o_{maxB} are indifferent, the choice will be determined by chance (guessing).

Apparently, the cognitive algorithm specified in the priority heuristic is based on a number of distinct assumptions that have to be jointly met for the model to be supported. The cognitive process is supposed to be sensitive to only one attribute at a time, being completely insensitive

to $p(o_{min})$ and o_{max} during the o_{min} stage, fully independent of o_{max} and o_{min} when focusing on $p(o_{min})$, and so forth. Moreover, the three stages always follow the same order, being invoked separately and never working together. A precise quantitative stopping rule (10% of some benchmark) is assumed for the first two stages; the same algorithm is assumed to apply for all choices between options with a similar expected value.[3] No other influences are expected to be at work. The a priori odds of the conjunction of all these specific assumptions must be low. The very precision of the priority heuristic serves to reduce Ω_{prior}.

To be sure, algorithmic models vary in the strength of the assumptions they propose. However, as a rule, refined and sophisticated models inevitably carry the burden of multiple conjunctive assumptions, which reduce the models' prior odds. Modelers might contend that they do not expect their models to be isomorphic but to be only paramorphic representations (Hoffman, 1960) that behave "as if" they were emulating the process to be explained. However, this disclaimer does not increase Ω_{prior}; it is but another way to admit that the model's assumptions are not jointly true. In any case, the a priori odds of model-based hypotheses decrease with increasing constraints imposed on the model.

Moreover, model testing may not only rely on low Ω_{prior} but also on low diagnosticity (LR). The evidence predicted by one specific model is often also compatible with several other models that often make entirely different assumptions. Thus, evidence for binary choice tendencies (i.e., which lottery, A or B, is chosen by a majority of participants) hardly provides unequivocal evidence for the priority heuristic and against other models that may assume fundamentally different processes: noncompensatory and compensatory models (Gigerenzer & Goldstein, 1996), exemplar-based or feature-abstraction algorithms (Juslin & Persson, 2002), symbolic or subsymbolic mechanisms (Fiedler, 1996), sample-based decision algorithms (Gonzalez & Dutt, 2011; Stewart, Chater, & Brown, 2006), or cumulative decision weights (as in prospect theory, Tversky & Kahneman, 1992).

To be sure, a model's diagnostic value can be greatly enhanced if it predicts a highly informative pattern of results that distinctively diverges from predictions of other models (Campbell, 1966; Meehl, 1990). For example, a connected set of lottery choices (A vs. B; C vs. D; E vs. F; A vs. C, and so on) might be deliberately designed to set the priority heuristic apart from other models (Katsikopoulos & Gigerenzer, 2008).[4] However, even in the auspicious case of a study lending distinct support to one model, it is often impossible to explain causally the success of the superior model when it differs in multiple ways from alternative models.

As explained by Roberts and Pashler (2000), there is no logical basis to infer from a model's (absolute or relative) fit that it actually reflects the underlying process. Recent research on illusory correlations highlights this insight (Kutzner & Fiedler, 2015). The illusion that the same high proportion of positive behaviors observed in a large and in a small group leads to more positive impressions of the majority has been explained by fundamentally different models: feedforward (Fiedler, 1996) and recursive models (Van Rooy, Van Overwalle, Vanhoomissen, Labiouse, & French, 2003), exemplar-based models (Dougherty, Gettys, & Ogden, 1999) or prototype formation (Fiedler, 2000), attention shift (Sherman et al., 2009), differential regression (Fiedler & Krueger, 2012), pseudocontingencies (Fiedler, Freytag, & Meiser, 2009), or striving for meaningful distinction (McGarty, Haslam, Turner, & Oakes, 1993). These models are different and in structure and noncomparable in so many aspects (process assumptions, scaling assumptions, number of free parameters, scope, and so on) that no empirical evidence in favor of one particular model implies that all aspects of the model must have been jointly effective.

Last but not least, model testing always focuses on a few selected models drawn from a universe of alternative models, many of which remain untested. It will never be possible to study the full Cartesian product of all combinations of possible model assumptions. The priority heuristic alone—disregarding all other models—allows for hundreds of ways in which choice algorithms can utilize o_{minA}, o_{minB}, $p(o_{minA})$, $p(o_{minB})$, o_{maxA}, and o_{maxB} or seemingly irrelevant variables like $p(o_{maxA})$ or $p(o_{maxB})$, which can be combined in different orders, strictly sequentially or in compensatory ways with different weightings, moderated by countless interaction terms. As it is impossible to investigate all models or instantiations of the same class of models, a plethora of ignored models must delimit the prior odds of selected models and the diagnosticity of selective empirical evidence.

Functional-level models. Functional-level models entail similar problems as the mechanistic process models discussed so far. A prominent example can be found in tests of mediation models that have become a gold standard for research to be published in leading journals. Otherwise non compelling correlations between an independent variable *X* and a dependent variable *Y* are augmented by testing a mediation model ($X \rightarrow Z \rightarrow Y$) suggesting that some third variable *Z* mediates the impact of *X* on *Y*. Thus, Kouchaki et al. (2014) assumed that an enhanced sense of control (*Z*) mediates the impact of guilt (*X*) on increased risk taking (*Y*). The argument relies on a statistical test showing that the correlation between *X* and *Y* decreases when the proposed mediator, *Z*, is controlled

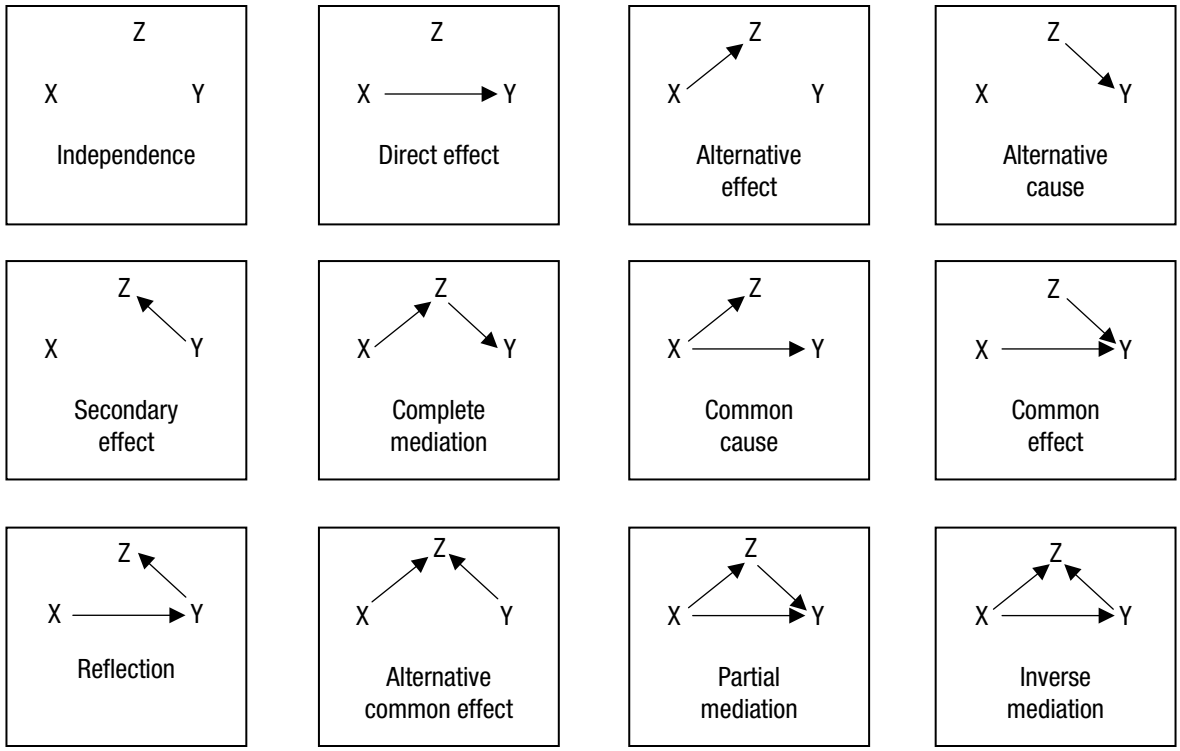

Fig. 3. Illustration of causal models. From only three variables, *X, Y,* and *Z*, a variety of 12 different causal models can be constructed (Danner, Hagemann, & Fiedler, 2015). The selective focus on mediation models and the neglect of alternative models restrict the diagnosticity of model testing.

statistically. Inferring from a significant mediation-model test that *Z* is indeed the causal mediator has become common practice, even though multiple alternative mediators *Z′*, *Z″*, *Z‴*, and so on and many other causal models involving three variables *X, Y,* and *Z* (see Fig. 3) remain untested.

This sort of causal inference is unwarranted. A significant test favoring one mediation model does not provide cogent evidence for the only tested mediator *Z*. Moreover, statistical mediation tests cannot discriminate between the mediation model $X \rightarrow Y \rightarrow Z$ and the other causal models in Figure 3 (Danner, Hagemann, & Fiedler, 2015; Fiedler, Schott, & Meiser, 2011). Mediation tests may be significant when *Z* is in fact not a mediator but a secondary measure of the dependent variable (as in the common-cause model $X \rightarrow Y, Z$). For instance, a statistical test may support the assumption that sense of control (*Z*) mediates the impact of guilt (*X*) on risk seeking (*Y*) even when sense of control is but another index of risk taking. Or the correlation between a virus (*X*) and a disease (*Y*) may be significantly reduced when fever (*Z*) is included in a regression model, mimicking mediation, even though fever is not a mediator but merely a symptom of the disease. Thus, given multiple mediators and multiple causal models, the diagnostic value of mediation-model tests must remain modest.

Toward a More Optimistic Appraisal of Psychological Science

Thus, our discussion of two prominent research approaches—sexy-hypothesis testing and model testing—indeed shows why Ioannidis's (2005) pessimistic conclusion is not off the point. On one hand, the prior odds Ω_{prior} remain low because sexy hypotheses are selected to be unlikely and unexpected and refined models entail multiple conjunctive assumptions. On the other hand, a low likelihood ratio (LR) reflects the limited diagnosticity of sexy hypotheses and model tests. If, however, both Ω_{prior} and LR are low, then the posterior odds $\Omega_{posterior}$ of research hypotheses in the light of empirical findings must also remain low.

Fortunately, however, solid research can also lead to valid findings. Why should the a priori likelihood of valid research hypotheses be only $P = .04$? Why should researchers be so strongly biased to embrace wrong hypotheses? Indeed, psychological science has generated many valid hypotheses leading to firmly established results: Self-generated information has a memory advantage, not a disadvantage (Bjork, 1994). Distributed learning is superior, not inferior, to massed learning (Hintzman, 1974). Partial reinforcement increases (rather than decreasing) resistance to extinction (Sheffield, 1949).

Creativity is facilitated by positive mood, not negative mood (Rowe, Hirsh, Anderson, & Smith, 2007). To be sure, not every study leads to the discovery of a stable law. However, to view the fruits of creative science, one has to separate the wheat from the chaff and to focus on the best findings rather than counting the noisy findings arising as byproducts of the discovery process. Such a reframed perspective alone results in a much more optimistic appraisal.

A more optimistic point could also be made for model-testing. Impressive convergent evidence exists for the validity and the usefulness of even very sophisticated models. For example, a family of sampling models can explain a multitude of judgment and decision biases in terms of restricted samples of information that happen to be provided by the information environment (Denrell, 2005; Fiedler, 2000; Stewart et al., 2006). Signal detection or related models are extremely useful to optimize legal or medical decisions (Swets, Dawes, & Monahan, 2000). Or, connectionist models describe cognitive and ecological structures that account for judgment and decision biases (Roe, Busemeyer, & Townsend, 2001).

Both Ω_{prior} and LR can be jointly enhanced in theory-driven cumulative science

Apart from such impressive examples of strong hypothesis testing and model testing research, the remainder of this article is devoted to outlining a largely neglected research approach that aims at jointly maximizing Ω_{prior} and LR. This alternative approach—call it theory-driven cumulative science—consists of the systematic derivation of diagnostic hypotheses from incontestable laws and logical constraints. To keep Ω_{prior} as high as possible, theory-driven cumulative science relies on undisputable logical rules or only well-established empirical laws. To maximize LR, the predicted patterns are so refined and informative that neither chance nor alternative theories can provide a reasonable account (Campbell, 1966; Shadish, Cook, & Campbell, 2002). Thus, the aim is to predict distinct patterns of functional relations (rather than merely binary trends) to maximize *LR* and to rely on strong theoretical ground to raise Ω_{prior} to the highest possible level. Note that Ω_{prior} in this approach is to a large extent based on logical and theoretical arguments rather than empirical data. Let us illustrate the idea of theory-driven cumulative science with two concrete examples, one representing theorizing based on analytical arguments and one relying on well-established empirical laws.

Strong theorizing based on analytical arguments. Regression to the mean constitutes a universal and incontestable law of the probabilistic world (Campbell & Kenny, 1999; Rulon, 1941). If the correlation r_{xy} of two variables X and Y is less than perfect ($|r_{xy}| < 1$), the expected $\hat{Y}$ values that can be predicted from given values of X must be regressive (Furby, 1973; Galton, 1886). Assuming scales of equal variance and ruling out other influences on Y, the expected individual $\hat{Y}_i$ scores must be less extreme (i.e., deviating less from the mean) than the corresponding individual predictor scores X_i scores. High (above-average) X_i values predict relatively lower $\hat{Y}_i$ values, and low (below-average) X_i values predict relatively higher $\hat{Y}_i$ values. More precisely, the deviations of $\hat{Y}_i$ from the mean, $\hat{y}_i = \hat{Y}_i - \text{Mean}(Y)$, can be expected to be r_{xy} times the X deviations, $x_i = X_i - \text{Mean}(X)$. Thus, if $r_{xy} = .5$, $\hat{Y}_i$ scores can be expected to be only half as extreme as the corresponding predictor values X_i.

To be sure, this rule describes the regressive shrinkage of expected $\hat{Y}$ scores that can be explained by predictor X. It does not determine obtained measures of Y that may be influenced by other factors (besides X) that may counteract the regression of $\hat{Y}$ on X. However, in any case, regression conceived as a theoretical construct (Fiedler & Krueger, 2012; Fiedler & Unkelbach, 2014) allows for strict theorizing. As r_{xy} decreases, $\hat{Y}$ can be expected to exhibit regressive shrinkage relative to X, but if obtained measures of Y do not show regression, one is on safe logical ground inferring the existence of an extraneous causal factor.

In psychophysics, for example, frequency judgments cannot be expected to match objective stimulus frequencies unless judgments are perfectly accurate ($r_{xy} = 1$). Because this condition is never met in reality, imperfect judgments ($r_{xy} < 1$) can be expected to exhibit regressive shrinkage (on comparable scales of equal variance). Large frequencies should be underestimated, whereas small frequencies should be overestimated; the larger (smaller) the objective frequencies, the stronger the underestimation (overestimation). In other words, regressive shrinkage is a (multiplicative) function of r_{xy} and the extremity of the stimulus quantities. Assuming a modest correlation of $r_{xy} = .5$, expected judgments shrink to half the objective values. If $r_{xy} = .75$, judgments shrink to three quarters of the objective quantities.

Regression is "as inevitable as death and taxes" (Campbell & Kenny, 1999, p. ix). Just as expected retest scores are less extreme than original test scores, or replication effect sizes cannot be expected to match original effect sizes, subjective judgments can be expected to regress on objective quantities. The incontestable law of regressive shrinkage can be used for strict theorizing. For example, Fiedler, Unkelbach, and Freytag (2009) tested the following refined set of predictions derived from the regression law. High frequencies should be underestimated and low frequencies should be overestimated, as already noted, and the degree of regressive under- and overestimation should increase with extremity. Moreover, because noise or cognitive load should reduce the judgment performance

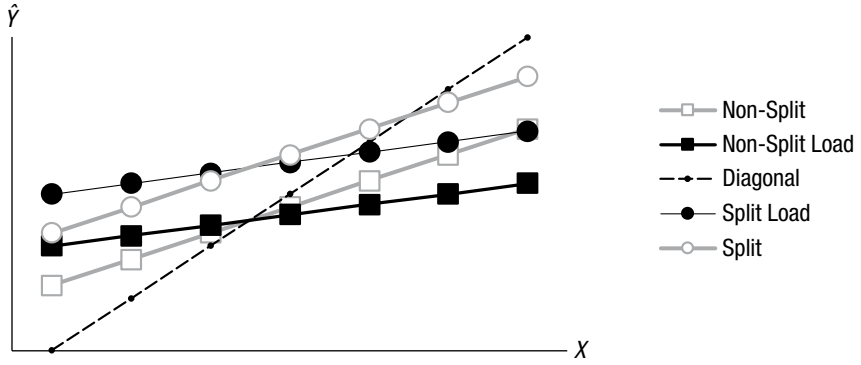

Fig. 4. Example of a refined hypothesis predicting a systematic pattern of expected frequency estimates $\hat{Y}$ (across seven levels) as a function of corresponding objective stimulus frequencies X and two moderating conditions (split and cognitive load).

r_{xy}, the manipulation of cognitive load should amplify the regressive pattern. It also was predicted that unpacking overall frequencies of stimulus categories (butterfly types) into smaller frequencies of subcategories (different color mutants of the same butterfly types) would cause extra regression. Splitting a small frequency into two extremely small subfrequencies should produce enhanced overestimation effects on both estimates, the sum of which must therefore exceed the estimate of the nonsplit category. Similarly, splitting a highly frequent category into two medium-size subcategories should undo regressive underestimation; summed subcategory ratings should no longer underestimate actual frequencies. Note in passing that the regression approach offers a natural explanation for unpacking or category split effects (Fiedler & Armbruster, 1994; Tversky & Koehler, 1994): Summed estimates of split categories should exceed estimates of nonsplit categories. It can be shown that splitting of small, medium, and high frequencies produces a constant overestimation effect (cf. Fiedler & Krueger, 2012).

Thus, logically sound theorizing using an incontestable law predicts the refined pattern in Figure 4, which plots expected subjective frequency judgments on the vertical axis as a function of objective stimulus frequencies on the horizontal axis. Several experiments, indeed, support this differentiated set of predictions (Fiedler, Unkelbach, & Freytag, 2009): High (low) frequencies are underestimated (overestimated); regression increases with extremity; cognitive load (induced by secondary task) amplifies the regression effect (see flatter slope of black than gray curves); and splitting (unpacking) category frequencies into smaller subcategory frequencies causes a constant increase in the summed estimates (see constant elevation of circles over squares).

Supportive evidence for such a complex pattern could not be simply due to chance. Obtaining the full pattern, derived from an incontestable law, can hardly reflect an α error. Significance testing becomes obsolete when strong theorizing predicts such informative, diagnostic patterns that cannot be explained by chance or by apparent alternative hypotheses. The epistemological status of the theory is not built on empirical data but on certitude; regression is always at work when the correlation between two variables (subjective and objective frequencies) is less than perfect (Campbell & Kenny, 1999).

Regression as a theoretical construct can not only generate novel hypotheses that would be overlooked otherwise. It also offers a parsimonious theoretical account of many established empirical phenomena. Unequal regression can explain biases in conditional reasoning (Fiedler, 2008), overconfidence (Erev, Wallsten, & Budescu, 1994; Moore & Healy, 2008; Oskamp, 1965), unrealistic optimism (Krueger & Mueller, 2002), or illusory correlations (Denrell & Le Mens, 2011; Fiedler, 1991; Kutzner & Fiedler, 2015). Thus, the regression construct provides a nice illustration of strictly theory-driven cumulative science. Innovative and highly diagnostic hypotheses can be derived logically. It matters little what an author believes or declares. What matters in theory-driven science is whether a pattern follows from the theory, which speaks for itself, independent of individual authors' motives and beliefs.

The predicted pattern in Figure 4 follows logically from the universal, nonfalsifiable law of regression. This is not to say that empirical research is obsolete because frequency judgments are predetermined anyway. In a multicausal world, it is always possible that other causal factors counteract and override the regression effect. For instance, a participant in a psychophysical experiment might deliberately correct for regressive shrinkage and thus produce nonregressive, polarized frequency judgments. Note, however, that such an empirical outcome—which is rarely found in frequency estimation studies—would not logically falsify the predictions derived from the regression construct. It

could rather be interpreted as cogent evidence that another causal influence has overridden the regression effect. Thus, in theory-driven research, the failure to obtain a predicted pattern is getting a completely new meaning. Rather than falsifying a solid theory, it can be exploited as a benchmark for the discovery of formerly unrecognized causal influences. In any case, the regression example highlights the fact that Ω_{prior} can be high for purely analytical reasons, independently of prior data, reflecting the a priori core of a theory (Lakatos, 1970).

Theorizing based on inductive empirical inferences. One might object that undoubted rules like regression are too scarce to inform many studies, but such a disclaimer strikes me as unwarranted. A whole variety of analytical principles is waiting to be exploited for theory-driven science: the principle of aggregation (i.e., cancellation of error variance), the law of the large number, the principle of relativity (i.e., context dependency), the increase in resolution level with decreasing psychological distance, or the impact of information density on judgment and learning to list but a few examples.

Still, in addition to science resting on a priori rules, there is also the possibility of basing theory-driven research on firmly established empirical laws. Negatively accelerated (concave) psychophysical functions (Stevens, 1957); the positive–negative asymmetry in associative learning (Unkelbach, Fiedler, Bayer, Stegmüller, & Danner, 2008); rules of conditioning (De Houwer, 2007); or the notion of white, pink, and brown noise in biological systems (Gilden, 2001) reflect empirical laws that can be used as benchmarks for cumulative science.

Decision-affect theory (Mellers, Schwartz, & Ritov, 1999) affords an illustrative case of such theorizing informed by solid empirical evidence. Numerous experiments highlight the relativity of judgments and decisions (Parducci, 1968; Stewart, Chater, Stott, & Reimers, 2003). The attractiveness of a decision option not only depends on the utility and probability of the chosen outcome but also on the utility and probability of alternative options. A positive (negative) outcome from one option may be disappointing (satisfying) if another option had resulted in an even better (worse) outcome. Relativity effects are particularly strong if the obtained outcome is unlikely or unusual, relative to the forgone outcome.

In formal notation, decision-affect theory specifies the reward value of an obtained option ($R_{obtained}$) as a weighted additive function of the utility of the obtained outcome ($u_{obtained}$) plus the disappointment (d) reflecting the difference $u_{obtained} - u_{forgone}$ of the obtained and forgone utility weighted by one minus the subjective probability of the obtained outcome ($s_{obtained}$):

$$R_{obtained} \sim u_{obtained} + d(u_{obtained} - u_{forgone}) \times (1 - s_{obtained})$$

Let "~" denote "is linearly related to." Thus, reward $R_{obtained}$ depends not only on the utility of the obtained outcome $u_{obtained}$ but also on a disappointment function d that is sensitive to relative utility ($u_{obtained} - u_{forgone}$). If d is negative because the forgone outcome exceeds the obtained outcome, disappointment will make $R_{obtained}$ smaller than $u_{obtained}$, especially if the obtained outcome is rare; that is, if $(1 - s_{obtained})$ is large. Figure 5 exhibits a possible pattern of predicted reward values $R_{obtained}$, assuming $u_{obtained}$ values of −12, −9, −6, −3, 0, 3, 6, 9, and 12, and $u_{forgone}$ values of −8, 0, and 8, for $s_{obtained}$ = .2 and .8. Disappointment d is assumed to be a power function of utility differences ($d = u^{0.5}$ and $d = -|u|^{0.5}$ for positive and negative u values, respectively). Note that the specific exponent of 0.5 does not belong to the theory's core assumptions but is only an auxiliary assumption used for scaling purposes (according to the distinction used by Lakatos, 1970; Meehl, 1990; and others).

Again, the theory predicts an entire set of curves that have been tested and supported in a series of experiments (McGraw, Mellers, & Ritov, 2004; Mellers, Schwartz, Ho, & Ritov, 1997). Unlike the analytical type of strong theorizing, decision affect theory contains falsifiable assumptions. One could falsify that disappointment (or regret) is a power function of utility differences that can be amplified by a weighting factor and by the improbability of the obtained outcomes. Such assumptions may be rooted in empirical findings that are not necessarily true. Still, even when the theory is not correct on a priori grounds, once the pattern of Figure 5 has been obtained regularly, this could hardly reflect a false-positive error by chance. It would be difficult to find alternative accounts for such a diagnostic pattern.

Theory-driven and phenomenon-driven research

The deductive approaches depicted in the last section represent one extreme on a continuum from theory-driven to phenomenon-driven research. Whether highly diagnostic and distinct hypotheses are derived from a priori rules or from approved empirical laws, the situation is quite different from the lottery-like hypothesis-testing game we have used to explain the dilemma of research that must be either innovative or solid but not both. Patterns like the ones in Figures 4 and 5 cannot be expected by chance, at an error rate of .05, nor can such a hypothesis be drawn by chance from an urn containing 10,000 hypotheses, of which 400 happen to be correct. The criterion of statistical reliability (α error) has to be replaced by a superordinate criterion of construct validity, conceived as an isomorphic match between predicted and obtained patterns (Westen & Rosenthal, 2003). Ω_{prior}

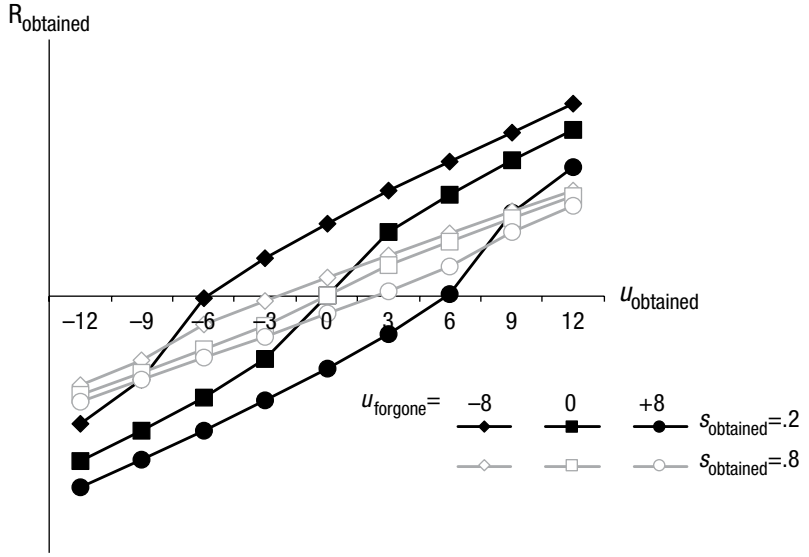

Fig. 5. Reward value, according to decision affect theory (Mellers, Schwartz, & Ritov, 1999), as a function of the utility of the obtained outcome ($u_{obtained}$) and the utility of the forgone outcome ($u_{forgone}$), assuming two different subjective probabilities of the obtained outcome ($s_{obtained}$). Because of an arbitrary scaling factor, $R_{obtained}$ is not stated numerically.

can be high and LR can be diagnostic when hypotheses rely on solid theorizing rather than only on noisy empirical data.

However, it is important to note that it is neither justified nor desirable to restrict behavioral science to strictly theory-driven research of the latter type. One should not underestimate the value of phenomenon-driven research on the opposite pole of the dimension, regardless of its modest levels of Ω_{prior} and LR. It is neither unscientific nor useless when social scientists, who live outside an ivory tower, dare to tackle risky research topics for which no strictly theory-driven account is available (yet). On the contrary, such phenomenon-driven research, which is often of the sexy-hypothesis testing type, can be as creative, proficient, and fascinating, deserving the same respect as theory-driven research. Once an originally unlikely, risky hypothesis has received unexpected support, it brings about real scientific progress (cf. Van Lange, 2013).[5]

Let us illustrate this point with a few up-to-date examples. Given the key function of the immune system for well-being and health, the finding that psychological interventions like humor (Lefcourt, Davidson-Katz, & Kueneman, 1990) or self-disclosure (Pennebaker, 1989) lead to measurable improvement in immune-system indices must be embraced as exciting and challenging. Recent evidence shows that disgusting stimuli serve as catalysts that can amplify psychologically triggered immune reactions (Schaller & Park, 2011). How could psychologists refrain from tackling such exciting research topics just because the precise causal mechanism and hence the boundary conditions for replication are unknown? Given the potential benefits of insights on the immune system, should the quality of such bold, risk-abiding research depend on the rate of correct hypotheses tests or successful replications?

The answer to both questions is certainly no. It is not reflective of inferior research but actually an obligation and a necessary part of responsible science when psychologists dare to study such important issues. Other examples of bold phenomenon-driven research—without a well-established theory—can be easily found in the literature. The phenomenon of stereotype threat refers to academic-performance reduction of members of a stereotyped group that are merely reminded of the stereotype (Steele, 1997). False confessions due to memory illusions and interrogation techniques have been shown to be more prevalent than expected (Shaw & Porter, 2015). Unconscious stimulation during sleep can systematically affect attitudes (Feld & Born, 2015). Multiple-choice formats may not provide appropriate measures of academic performance (Roediger & Marsh, 2005). Self-control may affect health, professional success, well-being, and achievement (Tangney, Baumeister, & Boone, 2004).

The psychological value of such enlightening findings does not necessarily depend on their generality or external validity (Campbell, 1957). Thus, even when the ability of disgust to trigger immune reactions can be established as a robust finding, this need not mean that the same

disgust effect generalizes across all individuals, settings, and cultures. Assuming the perfect validity of a single causal hypothesis (i.e., that disgust strengthens immune reactions), it is nevertheless possible that other causal influences on the immune system can overshadow the disgust effect. Yet, at a functional level of analysis, disgust effects can be enlightening and intriguing, fostering further research and interventions—well before the underlying causal mechanism is fully understood.

Concluding Remarks

Science is a pluralistic endeavor that should not be forced into the corset of one specific format. If science is to flourish and to achieve progress, there must be room for competing theories, methods, and different conceptions of what science is about. Symbiotic collaboration must be possible between theory-driven and phenomenon-driven research. There is no reason to disqualify or downgrade properly conducted research of any particular type.

However, for science to grow and to unfold its potential in the future, it is essential to recognize the chances and limitations of distinct types of research and to deal with many challenges in theorizing and logic of science—beyond superficial issues of data analysis. No statistical analysis can be better than the design of a study, and no research design can be better than the rationale of the underlying theory. The logic and the purpose of scientific inquiry often are neglected in the current discourse on quality of science (cf. Higgins, 2004; Kruglanski, 2001), which is almost totally centered on statistics and compliance rules.

To summarize the arguments presented in this article, I started with a discussion of Ioannidis's (2005) metascientific count of "wrong findings." A critical analysis revealed that minimal hypotheses suggesting an increase or decrease in Y as a function of a single causal condition X can hardly inform strong scientific inferences. In Bayesian terms, such studies suffer from low Ω_{prior} (when sexy hypotheses are unlikely on a priori grounds) as well as from low LR (because alternative causes also account for increases or decreases in Y). As a consequence, the results of singular studies, relying on arbitrarily selected operationalizations of $Y = f(X)$, neither justify the inference that the hypothesis is true (because a predicted outcome might reflect another, correlated cause) nor the inference that the hypothesis is false (because a negative outcome may reflect the overshadowing impact of an uncontrolled cause).

Increasing the number of participants to achieve higher statistical power or running a Bayesian rather than a Fisherian t test cannot solve this fundamental problem. To improve the validity of scientific inferences, investigators must take the first and foremost step of improving research designs, treating not only participants but also stimuli, tasks, and measures as random variables. However, although approximating a representative design (Brunswik, 1955; Dhami, Hertwig, & Hoffrage, 2004; Wells & Windschitl, 1999; Westfall, Kenny, & Judd, 2014) serves to foster the external validity or generality of scientific findings, it cannot solve theoretical problems of discriminant validity. Even the most sophisticated statistics and designs cannot rule out the possibility that other hypotheses provide more adequate accounts than the focal hypothesis.

This is not to say that strong scientific inference is impossible; it is just something ambitious and difficult to achieve, occurring in only a few extraordinary moments when a deeper understanding of an entire causal structure is obtained after a long series of careful and clever studies in a well-understood paradigm. One should not expect the outcome of each and every singular study to provide unequivocal evidence for the validity of a hypothesis and to discover ultimate solutions of major theoretical problems. In addition, one should not underestimate the exploratory insights gained from many ordinary studies that do not involve final theory tests. Phenomenon-driven research can be enlightening and practically important, especially when studies are content valid (e.g., measuring real people's immune system) and when they help researchers to develop clever research designs or new methods. Nevertheless, the distinction between such laudable exploratory research in applied and natural domains and strong theoretical inferences in cumulative science must not be blurred. Fully different criteria are needed to evaluate the quality of both kinds of research.

With respect to a second research approach, model-testing, I have shown that more complex and sophisticated hypotheses that are wired into formal models need not overcome the limitations in Ω_{prior} and LR. Regardless of the respect and admiration that I have for model builders' formal skills, model tests involve very strong conjunctive assumptions that reduce Ω_{prior}, and LR is often restricted when different models (often with qualitatively different structures and architectures) can account for the same empirical findings. As a consequence, model testing per se need not warrant unequivocal scientific inferences, even though it may inspire precise theoretical and causal reasoning. It seems fair to conclude that modeling mainly contributes to the logic of discovery (Reichenbach, 1938/1952) and only rarely leads to confirmatory diagnostic inferences. Really compelling models that have greatly improved behavioral science (such as Swets et al., 2000) are precious and rare.

The main section of this article was devoted to delineating an alternative research approach, which suggests a straightforward but somewhat neglected way to improve

scientific inferences, in terms of both Ω_{prior} and LR. This alternative involves the derivation of distinct and refined patterns of predictions, supposed to be as diagnostic and informative as possible, from firmly established principles. Such research suggests how both LR (making predicted patterns too refined to be expected from chance or alternative hypotheses) and Ω_{prior} (deriving hypotheses from incontestable or firmly established assumptions) could be maximized. This approach emphasizes strong theorizing, logic of science, and a priori reasoning. Although several examples testify to the viability of this alternative approach and although its domain may be larger than apparent at first sight, one might contest that this sort of science is hard to realize. This may be true. However, should this prevent scientists from trying to pursue both ideals at the same time, improving the a priori odds of hypotheses as well as designing diagnostic hypothesis tests? Are both ideals not worthwhile of being pursued anyway?

For the potential of pluralistic science to be explored, there has to be a competitive collaboration between methodologies and meta-theories. My aim here was not to denigrate sexy-hypothesis testing or model testing while idealizing the assets of cumulative theory-driven science as the only viable alternative. In fact, the boundaries of these research modes are blurred anyway; most real studies represent blends of more than one of these approaches. The high a priori odds of a good theory often take advantage of the empirical knowledge accumulated in previous research, including phenomenon-driven studies. Examples of fascinating and admirable research that deserves to be imitated can be found in all three (overlapping) camps.

Yet, while sexy-hypotheses and model testing flourish in current psychological science, it is amazing to see to what extent the a priori value of theories and the diagnosticity of research designs continue to be neglected. When it comes to evaluating quality of science, awarding the work of individual researchers, funding of research projects, making publication decisions for major journals, or selecting topics to be included in training curricula, the scientific community is giving almost all the weight to visibility indices and citation frequencies of surprising results, formal skills and precision cues associated with model fitting, and proficiencies in using statistical tools. Reviewers and editors of even the best publication outlets praise contributions that focus on mainstream positions (dual-process theories, rational choice, and so on), fast and sexy findings (automatic cognition and embodiment), fashionable methods (new statistics and mediation tests), and compliance with research practices (large sample sizes and transparent data repositories).

However, the "system" that governs publication, funding, and awards is hardly sensitive to slow and rigorous research or clever designs and even less sensitive to strong and logically sound theorizing. It does not sufficiently appreciate the Salmon principle as formulated by Meehl (1990, p. 115): "The main way a theory gets money in the bank is by predicting facts that, absent the theory, would be antecedently improbable." Paul Meehl is also explicit in stating that the diagnosticity or "intolerance" of a hypothesis "is not best judged by traditional significance testing" (p. 139) but only by comparing competing theories' logical constraints (i.e., their "Spielraum").

The future growth of psychological science calls for a change in the value hierarchy from statistics to research design and theorizing. For research to flourish and to enable strong scientific inferences, in addition to surprising and inspiring discoveries and reputable methods and models, it is essential to take the diagnosticity of empirical hypothesis tests and the a priori likelihood of underlying theories into account.

Declaration of Conflicting Interests

The author declared no conflicts of interest with respect to the authorship or the publication of this article.

Funding

The research for this article was supported by a Koselleck grant (Fi 294/23-1) from the Deutsche Forschungsgemeinschaft.

Notes

1. Ioannidis (2005) originally referred to positive predictive value (PPV), an index slightly different from TP.
2. In decision making under risk, the outcome probabilities are known, unlike decision making under uncertainty.
3. Thus, another strong assumption is that the required expected value is known or estimated to select a heuristic.
4. Such a diagnostic pattern should be observed within individual participants and not only at group level in majority choice rates.
5. Note that the confirmation of unlikely hypotheses is asymmetrically more diagnostic than its disconfirmation (Trope & Thompson, 1997).

References

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General*, *108*, 441–485.

Bell, R., & Buchner, A. (2012). How adaptive is memory for cheaters? *Current Directions in Psychological Science*, *21*, 403–408.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, A. P. Shimamura, J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: The MIT Press.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Making choices without tradeoffs. *Psychological Review*, *113*, 409–432.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*, 193–217.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*, 297–312.

Campbell, D. T. (1966). Pattern matching as an essential in distal knowing. In K. R. Hammond (Ed.), *The psychology of Egon Brunswik* (pp. 81–106). New York, NY: Holt, Rinehart and Winston.

Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York, NY: Guilford Press.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Danner, D., Hagemann, D., & Fiedler, K. (2015). Mediation analysis with structural equation models: Combining theory, design, and statistics. *European Journal of Social Psychology*, *45*, 460–481.

De Houwer, J. (2007). A conceptual and theoretical analysis of evaluative conditioning. *The Spanish Journal of Psychology*, *10*, 230–241.

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, *112*, 951–978.

Denrell, J., & Le Mens, G. (2011). Seeking positive experiences can produce illusory correlations. *Cognition*, *119*, 313–324.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*, 959–988.

Diekmann, A. (2011). Are most published research findings false? *Jahrbücher für Nationalökonomie und Statistik*, *321*, 628–635.

Dougherty, M. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180–209.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, Article 621. Retrieved from http://journal.frontiersin.org/article/10.3389[/fpsyg.2015.00621/

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*, 519–527.

Feld, G. B., & Born, J. (2015, May 29). Exploiting sleep to modify bad attitudes. *Science*, *348*, 971–972.

Fiedler, K. (1991). The tricky nature of skewed frequency tables: An information loss account of distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology*, *60*, 24–36.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. *Psychological Review*, *103*, 193–214.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*, 659–676.

Fiedler, K. (2008). The ultimate sampling dilemma in experience-based decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 186–203.

Fiedler, K. (2010). How to study cognitive decision algorithms: The case of the priority heuristic. *Judgment and Decision Making*, *5*, 21–32.

Fiedler, K., & Armbruster, T. (1994). Two halfs may be more than one whole: Category-split effects on frequency illusions. *Journal of Personality and Social Psychology*, *66*, 633–645.

Fiedler, K., Freytag, P., & Meiser, T. (2009). Pseudocontingencies: An integrative account of an intriguing cognitive illusion. *Psychological Review*, *116*, 187–206.

Fiedler, K., & Krueger, J. I. (2012). More than an artifact: Regression as a theoretical construct. In J. I. Krueger & J. I. Krueger (Eds.), *Social judgment and decision making* (pp. 171–189). New York, NY: Psychology Press.

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.

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, *47*, 1231–1236.

Fiedler, K., & Unkelbach, C. (2014). Regressive judgment: Implications of a universal property of the empirical world. *Current Directions in Psychological Science*, *23*, 361–367.

Fiedler, K., Unkelbach, C., & Freytag, P. (2009). On splitting and merging categories: A regression account of subadditivity. *Memory & Cognition*, *37*, 383–393.

Furby, L. (1973). Interpreting regression toward the mean in developmental research. *Developmental Psychology*, *8*, 172–179.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669.

Gilden, D. L. (2001). Cognitive emissions of 1/f noise. *Psychological Review*, *108*, 33–56.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*, 565–610.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, *118*, 523–551.

Hannula, D. E., & Ranganath, C. (2009). The eyes have it: Hippocampal activity predicts expression of memory in eye movements. *Neuron*, *63*, 592–599.

Higgins, E. T. (1992). Increasingly complex but less interesting articles: Scientific progress or regulatory problem? *Personality and Social Psychology Bulletin*, *18*, 489–492.

Higgins, E. T. (2004). Making a theory useful: Lessons handed down. *Personality and Social Psychology Review*, *8*, 138–145.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso & R. L. Solso (Eds.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 78–99). Oxford, England: Erlbaum.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin*, *57*, 116–131.

Ioannidis, J. P. (2005). Why most published research findings are false. *Chance*, *18*(4), 40–47.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*, 563–607.

Kassin, S. M. (2008). False confessions: Causes, consequences, and implications for reform. *Current Directions in Psychological Science*, *17*, 249–253.

Katsikopoulos, K. V., & Gigerenzer, G. (2008). One-reason decision-making: Modeling violations of expected utility theory. *Journal of Risk and Uncertainty*, *37*, 35–56.

Klein, S. B. (2012). A role for self-referential processing in tasks requiring participants to imagine survival on the savannah. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1234–1242.

Kouchaki, M., Oveis, C., & Gino, F. (2014). Guilt enhances the sense of control and drives risky judgments. *Journal of Experimental Psychology: General*, *143*, 2103–2110.

Krueger, J., & Mueller, R. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*, 180–188.

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.

Kutzner, F. L., & Fiedler, K. (2015). No correlation, no evidence for attention shift in category learning: Different mechanisms behind illusory correlations and the inverse base-rate effect. *Journal of Experimental Psychology: General*, *144*, 58–75.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–195). London, England: Cambridge University Press.

Lefcourt, H. M., Davidson-Katz, K., & Kueneman, K. (1990). Humor and immune-system functioning. *Humor: International Journal of Humor Research*, *3*, 305–321.

McGarty, C., Haslam, S. A., Turner, J. C., & Oakes, P. J. (1993). Illusory correlation as accentuation of actual intercategory difference: Evidence for the effect with minimal stimulus information. *European Journal of Social Psychology*, *23*, 391–410.

McGraw, A. P., Mellers, B. A., & Ritov, I. (2004). The affective costs of overconfidence. *Journal of Behavioral Decision Making*, *17*, 281–295.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*, 108–141.

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science*, *8*, 423–429.

Mellers, B. A., Schwartz, A., & Ritov, I. (1999). Emotion-based choice. *Journal of Experimental Psychology: General*, *128*, 332–345.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*, 502–517.

Nairne, J. S., Pandeirada, J. S., & Thompson, S. R. (2008). Adaptive memory: The comparative value of survival processing. *Psychological Science*, *19*, 176–180.

Oskamp, S. (1965). Overconfidence in case study judgments. *Journal of Consulting Psychology*, *29*, 261–265.

Parducci, A. (1968, December). The relativism of absolute judgments. *Scientific American*, *219*(6), 84–90.

Pennebaker, J. W. (1989). Confession, inhibition, and disease. In L. Berkowitz & L. Berkowitz (Eds.), *Advances in experimental social psychology* (Vol. 22, pp. 211–244). San Diego, CA: Academic Press.

Reichenbach, H. (1952). *Experience and prediction*. Chicago, IL: University of Chicago Press. (Original work published 1938)

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review*, *108*, 370–392.

Roediger, H. I., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1155–1159.

Rowe, G., Hirsh, J. B., Anderson, A. K., & Smith, E. E. (2007). Positive affect increases the breadth of attentional selection. *PNAS: Proceedings of the National Academy of Sciences of the United States of America*, *104*, 383–388.

Rulon, P. J. (1941). Problems of regression. *Harvard Educational Review*, *11*, 213–223.

Schaller, M., & Park, J. H. (2011). The behavioral immune system (and why it matters). *Current Directions in Psychological Science*, *20*, 99–103.

Schmidt, F. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, *5*, 233–242.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shaw, J., & Porter, S. (2015). Constructing rich false memories of committing crime. *Psychological Science*, *26*, 291–301.

Sheffield, V. F. (1949). Extinction as a function of partial reinforcement and distribution of practice. *Journal of Experimental Psychology*, *39*, 511–526.

Sherman, J., Kruschke, J., Sherman, S., Percy, E., Petrocelli, J., & Conrey, F. (2009). Attentional processes in stereotype formation: A common model for category accentuation and illusory correlation. *Journal of Personality and Social Psychology*, *96*, 305–323.

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.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613–629.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*, 153–181.

Stewart, N., Chater, N., & Brown, G. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*, 1–26.

Stewart, N., Chater, N., Stott, H. P., & Reimers, S. (2003). Prospect relativity: How choice options influence decision under risk. *Journal of Experimental Psychology: General*, *132*, 23–46.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better

grades, and interpersonal success. *Journal of Personality*, *72*, 271–322.

Trope, Y., & Thompson, E. P. (1997). Looking for truth in all the wrong places? Asymmetric search of individuating information about stereotyped group members. *Journal of Personality and Social Psychology*, *73*, 229–241.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Tversky, A., & Koehler, D. J. (1994). Support theory: A non-extensional representation of subjective probability. *Psychological Review*, *101*, 547–567.

Unkelbach, C., Fiedler, K., Bayer, M., Stegmüller, M., & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology*, *95*, 36–49.

Van Lange, P. M. (2013). What we should expect from theories in social psychology: Truth, abstraction, progress, and applicability as standards (TAPAS). *Personality and Social Psychology Review*, *17*, 40–55.

Van Rooy, D., Van Overwalle, F., Vanhoomissen, T., Labiouse, C., & French, R. (2003). A recurrent connectionist model of group biases. *Psychological Review*, *110*, 536–563.

Verhagen, J., & Wagenmakers, E. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *The Quarterly Journal of Experimental Psychology*, *12*, 129–140.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*, 1115–1125.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*, 608–618.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*, 2020–2045.