Voodoo Correlations Are Everywhere—Not Only in Neuroscience

Klaus Fiedler

Perspectives on Psychological Science 2011 6: 163
DOI: 10.1177/1745691611400237

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
6(2) 163–171

DOI: 10.1177/1745691611400237
http://pps.sagepub.com
SAGE

Voodoo Correlations Are Everywhere—Not Only in Neuroscience

Klaus Fiedler
Department of Psychology, University of Heidelberg, Heidelberg, Germany

Abstract

A recent set of articles in *Perspectives on Psychological Science* discussed inflated correlations between brain measures and behavioral criteria when measurement points (voxels) are deliberately selected to maximize criterion correlations (the target article was Vul, Harris, Winkielman, & Pashler, 2009). However, closer inspection reveals that this problem is only a special symptom of a broader methodological problem that characterizes all paradigmatic research, not just neuroscience. Researchers not only select voxels to inflate effect size, they also select stimuli, task settings, favorable boundary conditions, dependent variables and independent variables, treatment levels, moderators, mediators, and multiple parameter settings in such a way that empirical phenomena become maximally visible and stable. In general, paradigms can be understood as conventional setups for producing idealized, inflated effects. Although the feasibility of representative designs is restricted, a viable remedy lies in a reorientation of paradigmatic research from the visibility of strong effect sizes to genuine validity and scientific scrutiny.

Keywords

paradigmatic research, sampling filters, file drawer, representative design

Hardly any methodological article has received as much attention as the recently published article by Vul, Harris, Winkielman, and Pashler (2009) on inflated correlations in social neuroscience. Many months before the article appeared in print, it spread like wildfire, starting a fierce debate about the provocative term "voodoo correlations," used originally to denote the superstitious nature of certain neuroscience results in high-impact journals. The term was dropped from that article.

What specific evidence was challenged by Vul et al. (2009)? In a meta-analysis, the vast majority of correlations between blood oxygenation level dependent activity in the brain and behavioral measures of individual differences in personality, emotionality, social behavior, and related domains turned out to be excessively high. The criterion validity exceeded the upper limit derived from the brain measures' reliability. For example, Eisenberger, Lieberman, and Williams (2003) reported a correlation of $r = .88$ between anterior cingulate activity and social distress in ostracized individuals.

These extremely high criterion correlations were suspected to reflect a severe form of methodological myopia. A typical full-brain analysis encompasses over 100,000 measurement points called voxels, of which only a small subset is included in the final validity test. In most reviewed studies, researchers select only those voxels that bear the strongest correlations to the validity criterion. The use of such nonindependent data to construct the independent variable (i.e., brain activity) from the dependent measure constitutes a severe case of circular inference (see Hahn, 2011, for a differentiated discussion of circularity).

Neuroscientists presented several replies to this critique. Lieberman, Berkman, and Wager (2009), for instance, pointed to neuroscience studies that avoid blatant nonindependence, drawing only voxels from a theoretically predetermined region of interest (ROI). However, this debate of local independence—of the voxels defining a predictor from the criterion to be predicted in the same study—captures only one narrow aspect of a much broader problem. It matters little whether the voxels are selected because they correlate with a criterion identified in the same study or in previous studies. If an ROI is not based on independent theoretical grounds but on the selection of voxels strongly correlated with a criterion in prior studies, using other participants and stimuli, replicating such a selectivity effect in a new study can be hardly considered as

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

independent evidence. It will soon become apparent that these problems are by no means peculiar to neuroscience.

The General Sampling Problem Underlying Voodoo Correlations

Nonindependence has to be understood in a broader sense. The problem is not confined to the selective sampling of predictor measures; it also pertains to many other sampling biases, including the selection effect for published studies. With regard to the sampling of studies conducted, published and cited, Yarkoni (2009) notes that the "... level of inflation may be even worse than Vul et al.'s empirical analysis would suggest" (p. 294). A particularly deceptive source of bias is the low statistical power of the correlation coefficient (Yarkoni, 2009). Given a typical small sample size of 20 participants in fMRI studies, the median size of correlations that happen to be significant in the sample is between $r = .75$ and $r = .80$, regardless of whether the true correlation in the population is .7, .5, or .3. Thus, the expected sample of significant brain–behavior correlations that exceed the significance threshold required for journal publication has a similar distribution as the inflated correlations reviewed by Vul et al. (2009). Lack of statistical power also implies that the published correlations may not be the strongest ones that exist. Other correlations involving different brain functions and potentially supporting different theories may have gone unnoticed. Published fancy correlations, which made it into the highest impact journals, may be little more than outliers.

After some reflection, the issue of local nonindependence is only the tip of an iceberg of sampling effects that can dramatically bias empirical research in all fields, not just neuroscience. Strong criterion correlations in high-impact journals may be deceptive in various ways. Beyond the question of whether predictor data are stochastically independent of a criterion within the local data set, independence can be lost in many other arbitrary sampling decisions in the research process, such as the selection and publication of research questions and the operationalization of variables, tasks, stimuli, and instructions. All these sampling decisions might be made nonindependently of the expected research outcomes; they can therefore all contribute to the selective publication of inflated correlations.

Voodoo Evidence Not Only in Social Neuroscience

What we are dealing with here is a general methodological problem that has intrigued critical scientists under many different labels: the file-drawer bias in publication (Rosenthal, 1979); the tendency to capitalize on chance in model testing; the failure to treat stimuli, tasks, and measures as random factors in experimental designs; the paucity of representative designs; and the notion of circularity in the logic of discovery. They all converge in highlighting the illusive nature of inflated correlations. Visible research outcomes are multiply filtered and subject to various sampling biases: toward auspicious boundary conditions, exceptionally helpful stimuli, powerful manipulations of independent variables, arbitrarily selected factor levels, selective dependent measures, as well as mainstream topics and statistical findings.

Most of these extant sources of bias, which can be found in all areas of research, may appear less serious than the most blatant cases complained about by Vul et al. (2009). However, in their potential to produce inflated results that overstate the true size of a correlation, they are not fundamentally different from the voxel-selection problem in neuroscience.

In the remainder of this article, I present a variety of pertinent sampling biases with reference to recent research findings. I argue that the joint impact of all these biases can greatly inflate the size of reported effects and that paradigmatic research indeed can be understood as set of conventions that warrant invariance of empirical findings. The final discussion concerns possible remedies, the ideal of unbiased methodologies, the criteria of good science, and implications for the reviewing and publication process.

Inflation Due to Biases in Research

Inflated findings can be due to many filters, including those imposed by researchers on the design and analysis of their studies as well as those imposed by the institutions that publish, popularize, and fund research.

Biases from the study design

Sampling stimuli. The success of empirical research depends crucially on the choice of appropriate stimuli. Yet, the selection of stimuli is commonly considered a matter of the researcher's intuition. A "good experimenter" is one who has a good feeling for stimuli that will optimally demonstrate a hypothetical effect. Fortunate stimulus selection is thus respected like an asset or skill rather than treated as a problem. Even the most prestigious journals with the highest standards will not reject a paper when optimally selected stimuli have produced inflated findings. No norms exist to explicate and justify stimulus sampling.

An exemplary study is Tversky and Kahneman's (1973) seminal research on the availability heuristic. Participants were asked to judge the frequency of words with the letter *k* in the first position and in the third position. Although *k* appears more frequently in the third position in the English lexicon, it was judged to be more frequent in the first position, apparently because words that start with a particular letter are more available in memory. This study has entered many textbooks and curricula. However, a systematic replication and extension found that most other letters of the alphabet did not support the original finding (Sedlmeier, Hertwig, & Gigerenzer, 1998).

What mechanism may account for Tversky and Kahneman's (1973) stimulus-selection bias? Doubtlessly, these great scientists did not intentionally select the letter *k* to enforce a desired finding. However, they apparently did select their stimuli intuitively, following the common norm that this is every

experimenter's right. Selecting stimuli intuitively, however, involves a mental simulation process. In thinking about the stimulus materials suitable to demonstrate a phenomenon, researchers mentally simulate the process to be studied. As they are not only researchers, but also ordinary people, they can easily take their participants' role and observe their own reactions to the candidate stimuli. Such an intuitive selection process will typically favor those stimuli that happen to bring about the expected phenomenon, making mental simulation an omnipresent source of bias in behavioral research.

Overconfidence exemplifies a similarly famous stimulus selection bias. The tendency for subjective confidence estimates to be higher than the objective rates of correct responses to knowledge questions is greatly reduced, or almost eliminated, when judgment tasks are randomly drawn from a universe of all tasks (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, Winman, & Olsson, 2000). Although commonly treated as one of psychology's best-established phenomena, overconfidence is largely confined to studies in which judgment items were selected intuitively, presumably with a good feeling of which tricky knowledge questions will produce the desired effect.

Even strong proponents of representative designs are not immune from the tendency to use rewarding stimulus sets repetitively. Numerous studies on Gigerenzer and Goldstein's (1996) take-the-best heuristic use paired comparisons of the population size of towns—a context within which the heuristic was repeatedly shown to work because the best cue actually provides a useful proxy for population size. Researchers hardly ever apply the take-the-best heuristic to other knowledge domains in which the best cues are misleading, such as social stereotyping or lie detection (Vrij, 2007). Thus, sampling problems not only pertain to individual stimuli but also to entire stimulus domains.

Pilot testing of stimuli and tasks. Analogous to the role of mental simulation, pilot testing can serve a more systematic simulation function. Although careful pilot testing has a good reputation in empirical research, it helps to shape a research setting in such a way that the strength of a predicted effect is inflated, overstating the strength of the true effect in the latent universe of all possible ways of testing and operationalizing the same question.

Imagine a researcher who is eager to disconfirm an empirical phenomenon, such as action priming (e.g., that participants spontaneously walk slower when the concept of the elderly is primed; Bargh, Chen, & Burrows, 1996). Running many experiments using different stimuli but only reporting a single study that yields the desired result would be certainly regarded as illegitimate. However, if the same researcher runs and reports only one "main study" with the intended outcome, nobody would care about "pilot testing" used to select the stimuli that bring about that outcome.

This note on pilot testing corroborates that the nonindependence problem is much broader than the voodoo discussion in neuroscience suggests (cf. Lieberman et al., 2009). Predictor X is not nonindependent, or to some degree circular only when it is selected to correlate with Criterion Y in the same study (as suggested by Kriegeskorte, Simmons, Bellgowan, & Baker, 2009). If X has been selected in another study or pilot study using Y or a related measure (Y'), it can be hardly said to be independent of the criterion. It matters little whether the criterion used to select the predictor stems from the same study or from an earlier pilot study. If that distinction was crucial, nonindependence could be easily circumvented by splitting the study sample and calling the first half a pilot test. The first half could then be used (in an "exploratory" analysis) to select X to resemble Y, and the second half would then (in a "confirmatory" analysis) prove that X predicts Y. In either case, X is chosen in an empiricist fashion, to facilitate the desired outcome.

Biases from selecting variables and measures

In experimental research, stimuli serve to elicit responses to be measured. Stimulus sampling thus refers to independent variables whereas the sampling of measurement points (e.g., voxels) refers to dependent variables. However, the reverse is also possible. Stimuli can be part of an instrument used as a dependent measure, and the voxels defining a brain measure can play the role of an independent variable. In any case, there are distinct sampling biases for dependent and independent variables.

Dependent variables. Sampling dependent measures is tantamount to defining a research topic. Studying aggression, for example, means to find measures of aggression, the selection of which entails a potential confirmation bias. To illustrate, countless studies seem to support a causal impact of TV (and other media) on antisocial behavior like aggression (Bushman & Anderson, 2007). There is hardly any evidence for the opposite question, namely, whether TV consumption may facilitate prosocial behavior, although the principles of imitation learning are equally applicable to prosocial and antisocial behavior. Since the good guys are the winners in the vast majority of movies, their prosocial behavior ought to be imitated abundantly. If, however, it is correct that TV facilitates imitation in general, just because it is a rich and influential medium, the take-home message about TV consumption might be quite different if only more dependent measures of prosocial effects were included in media studies. Study outcomes rely heavily on the stimuli used for a test or measurement procedure. In research on the Implicit Association Test (IAT; Greenwald & Farnham, 2000), for instance, it is taken for granted that the obtained measure reflects something about the persons being tested (e.g., their prejudiced attitudes or stereotypes) rather than something about the stimuli used for the test. Thus, when West German participants quickly sort West-German and positive concepts onto one response key and East-German and negative concepts onto another key (relative to a reverse mapping of West-German and negative vs. East-German and positive concepts), this is commonly regarded as evidence for prejudice in West Germans. However this result can be eliminated and even reversed just by replacing the stimulus concepts used to represent West and East Germany and positive and negative valence (Bluemke & Friese, 2006). As this kind of

stimulus influence is generally not controlled in IAT studies (cf. Fiedler, Messner, & Bluemke, 2006), findings may reveal more about the stimuli chosen than the persons being tested. Thus, both internal validity (e.g., of IAT findings) and external validity (e.g., of media studies) can strongly depend on the choice of dependent variables, as we have seen.

Sampling levels of independent variables in fixed-effects designs. In addition to the selection of auspicious predictors and benevolent treatments, the selection of specific levels on an independent variable also affords a powerful means of boosting study results. Although participants are usually treated as a random factor, most study designs treat the independent variables of interest as fixed-effects factors, based on the arbitrary selection of (typically two) levels. To achieve a strong effect in Y, one may contrast highly distant X levels or select extreme groups. To undo or downplay an effect, one may induce smaller X differences or only a median split. The common reliance on such arbitrary designs, and the reluctance to follow Brunswik's (1955) ideal of representative designs with random sampling on all task dimensions (Clark, 1973; Cooksey, 1996; Dhami, Hertwig, & Hoffrage, 2004), has been shown to cause enormous validity problems (Wells & Windschitl, 1999).

The gold standard for the evaluation of scientific findings is still statistical significance. Many journals have also begun asking for effect size. However, both significance and effect size only focus on a difference observed in the dependent variable, independent of how much variation in the independent variable was needed to induce a strong and significant effect. Logically, the causal impact reflected in a study increases to the extent that small changes (Δ_X) in the independent variable cause large changes (Δ_Y) in the dependent variable, as captured in the ratio (Δ_Y: Δ_X). Doubling effect size Δ_Y does not reflect a stronger causal influence when based on double treatment strength Δ_X. Unfortunately, however, it is common practice to ignore the denominator required to normalize effect sizes. The sampling of X levels established in a fixed-effect design is excluded from interpreting effects in Y.

Researchers rarely have to justify what factor levels they compare. As authors or reviewers, we have all witnessed studies not published because Δ_Y was too small, but hardly any manuscript was rejected because the treatment needed for a given outcome was too strong. On the contrary, a maximal treatment (extreme groups, learning to criterion, perfect control of noise) is normally praised as good experimentation. Notably, though, such well-motivated rules of good science add another sampling filter to the illusion of inflated effects. Granting that there is a bias to report and publish strong and significant Δ_Y findings, and assuming that strong Δ_X facilitates the findings of strong Δ_Y effects, it can be inferred that research designs are generally biased toward strong Δ_X differences.

The failure to take Δ_X into account may also shed some light on voodoo correlations in "experiments of nature." For instance, the widely accepted finding that genes account for more variance than environmental influences in many twin studies is conditional on typical designs that allow for the full genetic variation in the population, whereas the environmental conditions to which even twins raised in different families are exposed are greatly restricted through cultural, social, and legal norms (Keller, 2007).

Biases from the analyses

Choices made about analyzing mediating and moderating variables can also lead to inflated findings.

Mediator variables and explanatory constructs. In addition to the obtained effect size, the scientific value attributed to a study also depends on sensible stories of the mediating process. Introducing a sensible mediator Z can greatly contribute to the exaggeration of the impact of an obtained X–Y correlation. One way of selecting a seemingly impressive mediator is to focus on a correlate of the dependent variable Y—let us call it Y'. Given that the correlate Y' is highly redundant with Y, especially when both measures are assessed in close temporal proximity, it can be shown (cf. Fiedler, Schott, & Meiser, 2011) that Y' is also likely to be strongly related to X, hence mimicking a significant result in a statistical test of the mediator model $X \rightarrow Z \rightarrow Y$ (Baron & Kenny, 1986).

For example, when attitude change is explained in terms of the relative number of positive versus negative thoughts generated by receivers of a persuasive communication in a thought-listing task (Tormala, Falces, Briñol, & Petty, 2007; Zuwerink & Devine, 1996), a sensible mediator seems to be found that greatly enhances the apparent consistency of the overall pattern. However, the alleged mediator, spontaneously generated thoughts, may just be a correlate of the dependent measure, another measure of the resulting attitude, rather than a causally antecedent mediator.

Boundary conditions and moderator levels. Moderator variables specify boundary conditions or enabling conditions for X–Y relations. Explicating potential moderators is a highly estimated practice in reviews and meta-analyses. Figuring out the moderator settings that maximize an effect is a most creative aspect of research. There is nothing illegitimate about moderator analysis. Yet, it should also be acknowledged that this well-motivated habit contributes to a bias toward strong effects. The extent of this bias is potentially controllable as long as the moderator influence is analyzed explicitly. However, for any study, there exists a long list of unknown implicit moderators, boundary conditions, parameter settings, and traditional features of approved paradigms that researchers cannot control for. Therefore, moderator settings can strongly contribute to inflated research findings.

The rules of the scientific community almost force their peers to adopt the approved parameter settings of a paradigm. If they do not apply the successful standard method named after a famous paradigm leader, then reviewers and editors may decide not to publish their research, even when the theory being studied is parameter free. A report on priming studies may be rejected if the time interval between stimulus and target onset (SOA) is longer than a few hundred milliseconds, even though the priming construct is meant to explain behavioral phenomena (like attitudes or aggression) that involve longer SOAs.

Because short intervals have been shown to maximize many priming effects, this suggests that paradigms rely on implicit moderators that maximize their effect size.

Biases from selective correction, publication, and funding of research

In addition to sampling filters that affect the internal and external validity of research findings, other filters greatly restrict the visibility of findings and the most fundamental decision to pursue certain research questions but not others. Funding schemes and policies restrict what research is conducted at all. Successful lines of research that warrant strong correlations are certainly more likely to be funded than research leading to weaker findings. Within any funded research project, the principle of selective correction comes into play. Researchers continue to "optimize" their design and procedure as long as findings are weak, but they will stop and freeze their methods as soon as strong and impressive correlations are obtained. Eventually, a file-drawer bias (Rosenthal, 1979) facilitates the selective publication of strong correlations while reducing the visibility of weak research outcomes. Altogether, these visibility-related filters further exaggerate the inflation of strong research findings.

Motives and Reasons for the Manifold Sampling Biases That Underlie Voodoo Correlations

What are the motives and the reasons underlying all these well-established sampling filters that jointly contribute to voodoo effects in behavioral research? What motivational, normative, and structural conditions account for the development of research ecologies that create the bias toward inflated empirical correlations?

Apparently, the depicted sampling biases should not be attributed to researchers' conscious attempts to deceive or to boost their obtained findings. It should also be obvious that many sampling biases reviewed in the preceding section are more subtle than the circularities that motivated Vul et al.'s (2009) critique. Moreover, the sampling biases that create inflated correlations are definitely not peculiar to neuroscience. They rather reflect a ubiquitous sampling phenomenon that can be found in many scientific areas, from genetics to experimental and applied psychology (Benjamini, 2008). To quote from Lazar (2009), "complicated, large data sets used to answer increasingly complex scientific questions ... increase our liability to make errors in the direction of selection bias" (p. 309).

Nevertheless, the selection biases I have listed are neither new nor hard to understand. It is therefore remarkable how widely ignored and even repressed they are, rather than being discussed openly. Such metacognitive myopia (Fiedler, 2000; Fiedler & Wänke, 2004) in sophisticated researchers, who only see the data but overlook the sampling filters behind, may be symptomatic of an industrious period of empirical progress, accompanied by a lack of interest in methodology and logic of science. A comprehensive discussion of this conspicuous insensitivity to selection biases would exceed the scope of this article. However, there can be little doubt that the following factors, in addition to many others science historians may find, are playing a major causal role.

Reinforcement structures

One primary reason for inflation biases certainly lies in the reinforcement structure of the scientific world. Strong effects are what peer researchers, students, journalists, and politicians find fascinating and they're what they want to read about in journals and textbooks—they motivate young scientists and facilitate the career of advanced scientists. As a consequence, the "system" encourages and often actually enforces the depicted sampling filters. Because journal space is expensive, editors are interested in strong and paradigmatic findings. Reviewers, who represent a paradigm, oblige authors to keep within the conditions of a successful paradigm. Supervisors do not recommend their junior partners to offend against the mainstream but to build their theses on well-established empirical laws. If the findings in a well-designed and carefully conducted experiment deviate from such laws, the chances of publication depend heavily on the authors' ability to reconcile their results with earlier, supposedly strong findings.

Thus, even though the system is not guided by the goal to deceive oneself or one's audience, the joint operation of several reinforcement schemes induces a pervasive confirmation bias. Every step of experimental design and scholarly publication is biased toward strong and impressive findings, starting with the selection of a research question; the planning of a design; the selection of stimuli, variables and tasks; the decision to stop and write up an article; the success to publish it; its revision before publication; and the community's inclination to read, cite and adopt the results.

As a consequence, there is a remarkable paucity of debates and a lack of interest in methodological controversies in major journals. An inquiry in the PsycINFO database reveals the following frequencies of papers related to important methodological issues that were published in the *Journal of Experimental Psychology* (all sections) or in the *Journal of Personality and Social Psychology* (between January 1, 1990, and June 21, 2010): "representative design": 2, "circularity": 1, "demand effect": 2, "confirmation bias": 10, "file-drawer bias": 0. In comparison, the corresponding numbers of references related to mainstream topics that promise strong correlations are as follows: "availability": 166, "priming": 939, "self control": 123; "automatic": 489; and of course "mediation": 160. Apparently, incentive structures favor the latter strong effects and discourage the former "methodological disclaimers."

Asymmetry of positive and negative feedback

As demonstrated in Denrell's (2005; Denrell & Le Mens, 2007) ingenious simulation studies, people and organisms tend to continue sampling as long as it is pleasant, but they stop when

it becomes unpleasant. Accordingly, researchers continue to use strong setups but truncate the use of setups that yield weak findings. In the long run, this "law of effect" (Thorndike, 1933) implies a bias toward strong phenomena that are likely to be reproduced and multiplied. A "law of repair" leads researchers to continue reanalyzing data and correcting analyses as long as the evidence does not (yet) support a law, but they truncate data analyses when strong support for a hypothesized law is found. Together, the law of effect and the law of repair contribute jointly to the confirmation of allegedly strong effects.

Shared information bias in science

A well-known phenomenon in group decision-making research is the shared-information effect (Mojzisch & Schulz-Hardt, 2006; Stasser & Titus, 1985). Rather than exchanging novel and independent arguments, group discussions are biased toward old and redundant arguments shared by other group members. Culture, indeed, can be defined as a selective force to communicate some information while omitting or excluding other information (Conway & Schaller, 2007). In scientific culture, too, researchers are inclined to exchange (i.e., publish, discuss, debate, teach) the very findings that they share with peer researchers. One does not need to postulate a motivational bias to explain this natural communication phenomenon. What Clark (1996) and Grice (1975) called common ground and cooperative communication, respectively, is sufficient to account for the bias in scientific discourse toward those strong findings that are shared by the scientific community.

Conclusions

To summarize, there should be agreement about two conclusions. First, the problem of voodoo correlations within social neuroscience is broader than originally shown by Vul et al. (2009). It is particularly not confined to studies lacking a predetermined ROI (Lazar, 2009).

Second, however, and much more broadly, the problem of inflated correlations due to manifold sampling biases is not peculiar to neuroscience. An intrinsic characteristic of all paradigmatic research is that various sampling filters jointly facilitate an illusion of strong effects, due to selective choice of stimuli, task conditions, moderator and parameter settings, and both independent and dependent variables, as well as selectivity in topics of research, debate, and publication.

Throughout this article, I have been concerned with relatively subtle causes of inflated effects rather than blatant mistakes that call the existence of basic effects into question. We have seen that paradigms of behavioral research can be conceived as conventionalized settings that warrant strong and replicable findings while excluding alternative theory tests under less auspicious conditions. Correlations and effect sizes obtained in paradigmatic research should thus be considered as upper limits rather than unbiased estimates of reality. One might object that many sampling biases are of restricted impact and that multiple biases cancel out each other. Yet one should be cautious not to underestimate the problem. The sum of all biases that together constitute a paradigm may well induce a similarly strong illusion as the most blatant cases of criterion-dependent voxel selection.

Remedies and countermeasures

What remedies or countermeasures may avoid or reduce those biases? Several authors suggest Brunswik's (1955) notion of a representative design as an appropriate remedy (Juslin et al., 2000; Sedlmeier et al., 1998). This methodology calls for the natural sampling of multivariate stimulus distributions from reality. Unrestricted sampling in broad-minded meta-analyses may indeed reduce sampling biases and create new insights. For example, evidence by Eisenberger and Cameron (1996) suggests that the overjustification effect derived from Festinger's (1962) dissonance theory (i.e., the reduction of intrinsic motivation after external reward) may not hold in certain work environments. Thus, critical research on the external validity of allegedly universal laws need not be frustrating, or destructive. It can be constructive and encouraging.

Problems with representative design. If all aspects of an empirical study are sampled representatively, then there is by definition no bias. To realize this ideal, it is not only necessary to draw a random sample of participants—one must also construct experimental tasks as a random sample of all possible reality tasks, and stimuli that mirror the universe of all stimuli, and randomly select variable levels that are representative of their natural distribution. Despite some respectable attempts to realize such representative designs (Dhami et al., 2004), they will hardly ever be realized fully, for obvious reasons. Drawing a sample from all possible stimuli, tasks, or contexts is not only hard to achieve technically, financially, and ethically. It is also impossible logically, whenever the universe cannot be defined or does not exist. Thus, what is the universe of all possible emotional stimuli, stressors, utilities, or means of social influence?

Moreover, representative designs are ill suited for the study of rare outcomes and unusual causes, which are often the most interesting ones. In a representative design, researchers would have to wait endless times for the natural occurrence of infrequent incidents and behaviors (cf. Fiedler, 2008). To study minority groups, low base rate diseases like HIV, anomalies, or novel interventions, researchers have to intervene and establish infrequent events at reasonable occurrence rates. Moreover, confining research to representative designs would mean foregoing the power of orthogonal experimental analysis.

Nevertheless, whenever it is possible to treat particular design factors as random factors, one should use the chance to improve the study's external validity. If a theoretical model of, say, basic language comprehension is meant to apply to language in general, it is essential to treat both participants and linguistic stimuli as random factors (cf. Clark, 1973). Any final conclusions about such intensely researched topics as face recognition, semantic and affective priming, irrationality, directed forgetting, or affective forecasting are logically contingent on

representative sampling of faces, primes, rationality tasks, memory contents, and affective episodes, respectively.

Convergent validation. As an alternative to optimizing external validity in a bottom-up fashion, one may try to deliberately maximize method variance in a research strategy called convergent validation (Garner, Hake, & Eriksen, 1956). Here, the goal is to find convergent evidence for an effect across divergent, heterogeneous methods and situations. The aim is not to arrive at an accurate quantitative estimate of a true effect size but to provide a qualitative proof for the existence and invariance of robust effects across diverse conditions. Unlike the benign parameter settings of most paradigmatic science, convergent validation can be expected to underestimate true effect sizes. Contrasting both methodologies could therefore be used to find upper and lower boundaries for an effect or empirical law.

Existence proofs and boundary conditions

One lesson to be gained from reflection on voodoo correlations is that good science does not need to strive for strong effects and seemingly precise models fitted to allegedly representative correlations. It may also consist of careful proofs of basic effects and causal conditions and processes leading to these effects. Even figuring out catalysts or enabling conditions for labile, nonstable findings may be an accomplishment in good science. Behavioral scientists may place more weight on descriptive studies of basic phenomena and their boundary conditions. Being explicit about restrictions and crucial catalysts of a phenomenon can be enlightening and conducive to important theoretical insights and not a concession of weakness and invalidity. For example, realizing that the overly liberal response bias that underlies the high false-alarm rate in eyewitness identification is mostly evident for familiar and likeable faces (Garcia-Marques, Mackie, Claypool, & Garcia-Marques, 2004) can improve our understanding of the phenomenon. Even specific parameter settings can be revealing theoretically. Kareev's (2000) finding that environmental samples are most likely to overestimate an existing effect at sample sizes of 7 ± 2 is fascinating because it matches the span of human working memory that may have evolved to exploit this parameter.

This point should not be reduced to the conventional desideratum to include one or two moderators in a good study design. It rather emphasizes the value of creative search for subtle boundary conditions and easily overlooked moderators that are only implicit in the restricted samples of stimuli, tasks, and experimental contexts. Let me illustrate this with another prominent example. A common law derived from prospect theory (Tversky & Kahneman, 1992) implies that people make risk-aversive decisions for positive outcomes (i.e., gains) but risk-seeking decisions for negative outcomes (losses). For example, when making a choice between two gambles with the same expected value, one offering a large gain at a low probability and one offering a small gain at a higher probability, people will choose the latter. However, as we know from Slovic (1995), empirical support for this prediction is restricted to choice tasks. When a pricing task rather than a choice task is used to assess the preference for the same gambles, people are willing to pay a higher price for the former gamble, which offers a higher outcome at a lower probability. This restriction of prospect theory to choice tasks reflects an implicit moderator that has far-reaching theoretical implications. It cannot be explained by prospect theory's sigmoid subjective-value function or its regressive subjective-probability function.

Closing statement

Eventually, then, Vul et al.'s (2009) provocative paper not only raises a serious threat to empirical research, it may also open a constructive and thoughtful debate about scientific growth and progress. The insights and implications of this debate that I have tried to convey in this article can be summarized as follows. First, paradigmatic research in general, not only in neuroscience, is characterized by conventionalized sampling biases that serve to inflate the strength of empirical findings. Second, these sampling biases reflect both the payoffs of the scientific system and the fact that unbiased methodologies (e.g., representative designs) are hardly feasible. Third, we have to accept the consequence that the size of a correlation is a deceptive index of scientific progress (Roberts & Pashler, 2000). Paradigmatic science cannot be expected to yield accurate estimates of the strength of correlations in the real world. Their purpose is rather to demonstrate the existence of effects and processes under idealized conditions—a crucial precondition for the causal analysis and the controlled application of all behavioral phenomena. Finally, it is important that journal editors, reviewers, funding agencies, academic teachers, and practitioners take these insights to their heart. Rather than ignoring and concealing the various boundary conditions that contribute to inflated correlations, leading theories should explicate and integrate them as essential enabling conditions for the validity of psychological laws.

Exaggerated correlations are published in high-publicity journals because of their visibility, not because of their validity. A healthy side effect of the voodoo correlations debate is to remind scientists of internal and external validity as an ultimate standard of good science.

Declaration of Conflicting Interests

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

Funding

The research underlying this article was supported by a Koselleck Grant awarded by the Deutsche Forschungsgemeinschaft to Klaus Fiedler.

References

Bargh, J., 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.

Baron, R., & Kenny, D. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Benjamini, Y. (2008). Comment: Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, *23*, 23–28.

Bluemke, M., & Friese, M. (2006). Do features of stimuli influence IAT effects? *Journal of Experimental Social Psychology*, *42*, 163–176.

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

Bushman, B., & Anderson, C. (2007). Measuring the strength of the effect of violent media on aggression. *American Psychologist*, *62*, 253–254.

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

Clark, H.H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

Conway, L.G., III, & Schaller, M. (2007). How communication shapes culture. In K. Fiedler (Ed.), *Social communication* (pp. 107–127). New York: Psychology Press.

Cooksey, R. (1996). The methodology of social judgement theory. *Thinking & Reasoning*, *2*, 141–173.

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

Denrell, J., & Le Mens, G. (2007). Interdependent sampling and social influence. *Psychological Review*, *114*, 398–422.

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

Eisenberger, N.I., Lieberman, M.D., & Williams, K.D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*, 290–292.

Eisenberger, R., & Cameron, J. (1996). Detrimental effects of reward: Reality or myth? *American Psychologist*, *51*, 1153–1166.

Festinger, L. (1962). *A theory of cognitive dissonance*. Oxford, England: Stanford University Press.

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 & Cognition*, *34*, 186–203.

Fiedler, K., Messner, C., & Bluemke, M. (2006). Unresolved problems with the "I," the "A," and the "T": A logical and psychometric critique of the Implicit Association Test (IAT). *European Review of Social Psychology*, *17*, 74–147.

Fiedler, K., Schott, M., & Meiser, T. (2011). *What mediation analysis can (not) do.* Manuscript submitted for publication.

Fiedler, K., & Wänke, M. (2004). On the vicissitudes of cultural and evolutionary approaches to social cognition: The case of metacognitive myopia. *Journal of Cultural and Evolutionary Psychology*, *2*, 23–42.

Garcia-Marques, T., Mackie, D., Claypool, H., & Garcia-Marques, L. (2004). Positivity can cue familiarity. *Personality and Social Psychology Bulletin*, *30*, 585–593.

Garner, W., Hake, H., & Eriksen, C. (1956). Operationism and the concept of perception. *Psychological Review*, *63*, 149–159.

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

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506–528.

Greenwald, A.G., & Farnham, S.D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology*, *79*, 1022–1038.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J.L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York: Academic Press.

Hahn, U. (2011). The problem of circularity in evidence, argument and explanation. *Perspectives on Psychological Science*, *6*, 172–182.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review*, *107*, 384–396.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review*, *107*, 397–402.

Keller, J. (2007). When negative stereotypic expectancies turn into challenge or threat: The moderating role of regulatory focus. *Swiss Journal of Psychology/Schweizerische Zeitschrift für Psychologie/ Revue Suisse de Psychologie*, *66*, 163–168.

Kriegeskorte, N., Simmons, W., Bellgowan, P., & Baker, C. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*, 535–540.

Lazar, N.A. (2009). Discussion of "Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition" by Vul et al. (2009). *Perspectives on Psychological Science*, *4*, 308–309.

Lieberman, M.D., Berkman, E.T., & Wager, T.D. (2009). Correlations in social neuroscience aren't voodoo: Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*, 299–307.

Mojzisch, A., & Schulz-Hardt, S. (2006). Information sampling in group decision making: Sampling biases and their consequences. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 299–326). New York: Cambridge University Press.

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

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Sedlmeier, P., Hertwig, R., & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *24*, 754–770.

Slovic, P. (1995). The construction of preference. *American Psychologist*, *50*, 364–371.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling in group discussion. *Journal of Personality and Social Psychology*, *48*, 1467–1478.

Thorndike, E.L. (1933). A proof of the law of effect. *Science*, *77*, 173–175.

Tormala, Z., Falces, C., Briñol, P., & Petty, R. (2007). Ease of retrieval effects in social judgment: The role of unrequested cognitions. *Journal of Personality and Social Psychology*, *93*, 143–157.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

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

Vrij, A. (2007). Deception: A social lubricant and a selfish act. In K. Fiedler (Ed.), *Social communication* (pp. 309–342). New York: Psychology Press.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

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

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*, 294–298.

Zuwerink, J., & Devine, P. (1996). Attitude importance and resistance to persuasion: It's not just the thought that counts. *Journal of Personality and Social Psychology*, *70*, 931–944.