Perspectives on Psychological Science
2021, Vol. 16(4) 816–826

DOI: 10.1177/1745691620970602
www.psychologicalscience.org/PPS

# Quo Vadis, Methodology? The Key Role of Manipulation Checks for Validity Control and Quality of Science

**Klaus Fiedler, Linda McCaughey, and Johannes Prager**
Psychology Department, Heidelberg University

**Abstract**

The current debate about how to improve the quality of psychological science revolves, almost exclusively, around the subordinate level of statistical significance testing. In contrast, research design and strict theorizing, which are superordinate to statistics in the methods hierarchy, are sorely neglected. The present article is devoted to the key role assigned to manipulation checks (MCs) for scientific quality control. MCs not only afford a critical test of the premises of hypothesis testing but also (a) prompt clever research design and validity control, (b) carry over to refined theorizing, and (c) have important implications for other facets of methodology, such as replication science. On the basis of an analysis of the reality of MCs reported in current issues of the *Journal of Personality and Social Psychology*, we propose a future methodology for the post–$p < .05$ era that replaces scrutiny in significance testing with refined validity control and diagnostic research designs.

**Keywords**

manipulation check, significance testing, validity, scientific scrutiny, attention check, demand effect, diagnostic design

What crucial difference makes scientific inquiry superior to lay wisdom and reasoning? A widely accepted answer is that the advantage of science stems from its reliance on a repertoire of methods that enable scientists to go beyond "common sense." Scientific methodology encompasses standardized tests, scaling and measurement techniques, technical instruments, and analytical procedures that make it possible to quantify and analyze empirical findings while filtering out noise and misleading confounds. At the heart of this methodology, however, lies strict *theorizing*—forming, testing, and developing theories, ideally in a cumulative fashion, in the spirit of the seminal writings of Paul Meehl (1978).[1] Methodology makes it possible for theories to explain carefully assessed behavior and performance in terms of empirical laws and causal principles.

## Hierarchical Ordering of Levels of Scientific Methodology

Theories occupy a double role in this framework. They constitute the substance and goal of scientific endeavors but, at the same time, they also entail the most important means to their own end, given that they are the most fundamental ingredient in theorizing. Without a theory, within which concepts can be defined and related to each other and from which hypotheses can be derived, there is no way of planning and conducting meaningful and methodologically sound research.

Subordinate to theorizing is the issue of constructing a *research design* that affords a cogent way of realizing the theorizing goals, using appropriate materials, procedures, and operational measures in a sufficiently complex task setting that is aligned with the theoretical hypothesis. The research design must be determined by the theory that researchers set out to test in order to be a *valid* representation thereof (Cronbach & Meehl, 1955). Theorizing is superordinate to research design in the sense that the theory defines what an appropriate, or valid, design is in the first place. Moreover, the more distinct and richer in implications a theory is, the more will it typically be able to guide a valid research design, increasing chances for informative results and advancement of theory-driven science.

**Corresponding Author:**
Klaus Fiedler, Psychology Department, Heidelberg University
E-mail: kf@psychologie.uni-heidelberg.de

Finally, on the bottom layer of the hierarchy, subordinate to the level of research design, is a set of methods used for *statistical analysis*. Theorizing and research designing constrain the choice of statistical analyses, not the other way around. Theory and design not only determine which statistical analysis is appropriate but also restrict the insights and inferences gained from statistics. The most elaborate statistical procedures are meaningless when empirical results emerge from inadequate research designs or are obscured through sampling and measurement error (Schmidt, 2010), demand effects, attrition, or inefficient manipulations. Despite the hierarchical ordering of theorizing over research design over statistical analysis, the ongoing debate about how to improve the quality of psychological research revolves almost exclusively around issues of statistical significance testing, at the lowest hierarchy level.

### *Cult of null-hypothesis significance testing*

"Is there an effect?" has become a most frequently asked question. Typically, the answer is "yes" if "$p < .05$"; otherwise, "probably not." Null-hypothesis significance testing (NHST) has been interpreted as "the survival of a flawed method" (Krueger, 2001), the logical insufficiency of which has been articulated many times by many scholars (Fiedler, 2020; Lykken, 1968; Meehl, 1978; Trafimow, 2003). Cohen (1994) deplored that NHST relies on the conditional probability $p(D|H_0)$ of obtaining data pattern $D$ given $H_0$, but what researchers want to infer is a reverse conditional $p(H_0|D)$ or $p(H_1|D)$ of $H_0$ or $H_1$ given $D$. There is no viable way of inferring the status of the hypothesis $H$ from the observed data $D$ in significance testing, and yet it continues to be misunderstood as affording exactly such an inference. This critique uttered by Cohen was never refuted but still seems poorly understood. Cassidy and colleagues (2019) showed that, in a review of 28 popular undergraduate textbooks, the frequentist-probability definition of NHST was entirely correct in only three textbooks.

Although Bayesian statistics promise a mathematical solution to this fundamental problem of scientific inference (Lavine & Schervish, 1999), any Bayesian appraisal of $p(H_0|D)$ versus $p(H_1|D)$ depends on issues that go beyond statistics, as we shall see in a later inspection of Bayes' theorem.

### *Neglect of research design and theorizing*

In stark contrast to the clear hierarchy of the levels of methodology, hardly any attention is given to superordinate issues of research design and theorizing. Methods courses, Internet blogs, and the journal review process largely ignore the intricate problems of auxiliary assumptions built into the operational means of testing theoretical hypotheses (Lakatos, 1978; Trafimow, 2019b), the attrition-rate problem in online research (Zhou & Fishbach, 2016), or the lack of diagnosticity of designs that juxtapose exactly two possible outcomes (Fiedler, 2017). Yet virtually all ongoing methodological debates revolve around preconditions of proper significance testing: questionable research practices (John et al., 2012), exploitations of researchers' degrees of freedom (Simmons et al., 2011), "*p*-hacking" and hypothesizing after the results are known (HARKing; Simonsohn et al., 2014), publication bias, or the alleged preponderance of false positives (Lilienfeld & Waldman, 2017).

## From Significance Testing Cult to More Mature Methodology

How can this conspicuous imbalance between the prominence of statistical significance testing and its logical inadequacy and insufficiency be explained? Why do most of us—including the present authors—continue to conduct and publish significance tests, even though we feel, or fully understand, that a precise $p$ below or above .05 or a Bayes factor higher or lower than 5 tell us hardly more than do descriptive statistics (Trafimow, 2019b)?

### *Striving for formal precision*

A plausible answer could be that significance testing serves an important function for the identity of behavioral scientists' striving for scrutiny and formal precision. In the absence of an alternative set of formal and objective principles of research designing and theorizing, significance testing offers a crutch to boost our identity as scientists who value scrutiny and methodological rigor. However, this part of our identity may soon fade anyway, given that *The American Statistician* has declared significance testing worthless (Wasserstein et al., 2019). The question is what alternative canon of rules can be employed in the post–$p < .05$ era (Trafimow, 2019a; Wasserstein et al., 2019). Is there a chance to rid ourselves of the crutch and to replace significance testing with more appropriate quality control? While the need for stricter theorizing is often emphasized (Fiedler, 2017; Fiedler et al., 2012; Meehl, 1967, 1990; Platt, 1964), many researchers, editors, and reviewers find it hard to imagine a nonstatistical alternative for scientific quality control.

Here we suggest a moderately optimistic solution. We believe that the recipe for a sound methodology in the post–$p < .05$ era is already well articulated in Campbell's (1957) seminal lessons on internal and external validity

and in Cronbach and Meehl's (1955) work on construct validity. Validity indeed affords a stricter and logically more coherent quality criterion than statistical significance. Validation of a construct ensures convergence and exchangeability with other measures and differentiation against other unrelated factors. Validity of applied measures is the key to proper research design and sound theorizing.

## *Manipulation checks as a catalyst of refined validity control and scientific scrutiny*

A highly useful but underused tool for validity control, and a major catalyst for improvement of the quality of science, is a proper manipulation check (MC). MCs are critical for the viability of the logical premise of a theoretical hypothesis H: $\Delta x \rightarrow \Delta y$, which predicts a shift in the dependent variable (DV) $y$ ($\Delta y$) given a shift in independent variable (IV) $x$ ($\Delta x$). This prediction is logically contingent on the premise that an experimental treatment actually succeeds in inducing the intended $\Delta x$ shift. Without that premise, predicting an effect $\Delta y$ is unwarranted.

To illustrate this point, consider the central assumption of construal-level theory (CLT; Liberman & Trope, 1998; Trope & Liberman, 2010; Soderberg, 2014): Abstractness and psychological distance are intrinsically related to each other. When shift in distance (Δdistance) is manipulated experimentally (e.g., by asking for judgments of close in-group vs. remote out-group), CLT assumes that the more distant out-group judgments tend to be more abstract than the less distant in-group judgments. Operationalizing distance in this way is contingent on the premise that out-groups are judged from a more distant perspective than are in-groups. Rather than taking this seemingly trivial assumption for granted, scientific scrutiny calls for a proper MC that (a) ensures the intended purpose of the manipulation and (b) is operationally independent of the DV in the experiment. Thus, participants might be asked to identify the group they have in mind and to indicate or estimate their location, contact frequency, or familiarity. An alternative MC could be some unobtrusive uncertainty measure, such as constructing confidence intervals for different knowledge questions (about in-group and out-group members' attributes or behaviors; see Krüger et al., 2014).

Support for the auxiliary assumption that the treatment indeed affords a suitable way of manipulating distance helps to validate the finding of more abstract out-group judgments relative to in-group judgments. From here, it is but one step further to realize that the validity of the DV (abstractness) is equally important. Attentive experimenters striving for optimal MCs will be sensitized to the validity of all variables, engaging in convergent validation (Garner et al., 1956)—an analogue of MC in the validation of dependent measures.

Now if researchers want to test the bidirectional hypothesis, or the correlation Δdistance ↔ Δabstractness, that Δdistance may not only reflect but also affect Δabstractness, CLT still assumes, as in the experimental case, that abstractness and distance are related systematically, reflective of a valid principle, rather than being incidental or spurious. Again, once a researcher is sensitive to proper MCs, he or she will apply the same scrutiny to correlational research and go to greater effort to improve the validity of all theoretical variables. The spirit of MC-based experimentation will thus carry over to improve methodology and theorizing more generally.

Conversely, experiments without MCs suffer from serious deficits of convergent and discriminant validity (Campbell & Fiske, 1959) because no manipulation can be expected to affect only a single IV. Thus, when choices for others, for example, do not produce the same choice-overload effect that is found in choices for oneself, the manipulation may have affected (a) social distance in the sense of CLT, (b) a shift from prevention focus to promotion focus in choices for others, or (c) a change in individual payoffs (Polman, 2012). How could a theoretical explanation of this finding make do without a refined MC?[2]

Discriminant validity becomes a particularly treacherous problem when a very concrete, specific manipulation mimics a clearly circumscribed influence, which may, however, reflect a much more general, superordinate construct. As explained in Wason's (1960) seminal article, a numerical sequence 2, 4, 8, 16, 32 does not provide unequivocal evidence for specific rule $2^N$, despite the perfect fit. It is also compatible with many other less specific, more general rules, such as superlinearly increasing integer numbers, increasing integer numbers, any numbers, or even alphanumerical symbols. Thus, if a distinct manipulation of mortality exposure (e.g., thinking of a funeral, the 9/11 attacks, or one's own mortality) produces a politically conservative shift (Pyszczynski et al., 2015), the causally effective manipulation may have nothing to do with mortality salience. The effect may be due to a more general variable, such as incompleteness, of which mortality is but a special case. Reminding early-semester student participants of their incompleteness (related to graduation) or simply interrupting participants on a task (Zeigarnik effect) may produce a similar conservative shift (Wicklund & Braun, 1987), independently of mortality (Fiedler, 2012).

Thus, refined MCs can be the most ingenious achievement of the most excellent pieces of psychological

science. MCs are indeed indispensable to validate empirical findings. We therefore do not fully share the conclusion implied by Hauser et al. (2018) that MCs should be viewed critically and omitted if they might influence the results. Although we agree that a blatant and demand-prone MC can be counterproductive and interfere with the natural influence of the IV on the DV, we do not think that we may be better off without one. We pose rather that MCs are a logical precondition of validation and that one always should and always can find a way to implement one. So the question is not about when demand-prone MCs should be omitted but about how to construct and implement clever and creative MCs that evade such side effects.

## Scope of Present Article and Preview

In the remainder of this article, we try to substantiate the notion that systematic validity control triggered by MC as a catalyst affords a canon of strict and clearly spelled-out methodological rules that can contribute to replacing significance-testing in a post–$p < .05$ era. The strength and scrutiny of a validity-based methodology originates in the primary value given to theoretical inferences, beyond the uncritical analysis of statistical data, cognizant of the hierarchical ordering of theorizing over research design over statistics. We argue that cultivating a single pivotal tool—the MC—can serve as a catalyst or heuristic that, once established, carries over to enhanced validity control and scrutiny of science, way beyond the immediate purpose of MCs. Although the idea that a single methodological tool can trigger the renewal of an entire methodology with a distinct focus on theoretical validation may appear overly optimistic, one need only to bring to mind the analogously engulfing effect that significance testing has exerted on methodology to appreciate that it is not entirely unrealistic.

To substantiate this notion, we begin with a review of the reality of MCs in a full year of research published in the *Journal of Personality and Social Psychology: Attitude and Social Cognition* (*JPSP:ASC*). The unsurprising outcome of the review—a conspicuous neglect of an essential instrument of validity control—motivates the subsequent discussion of advantages of MCs and its potential to trigger high-quality research. The focus of this discussion is on the strength and simplicity of a revised methodology, how easily it can be implemented, and how it strengthens the weight of cogent theorizing and theory-driven research design in empirical research process. Indeed, we are well prepared for a shift from significance testing to validation as a major quality criterion. In the final section, we present a tentative list of recommendations for how to implement these ideas in teaching, peer review, and scientific discourse.

## Status of Manipulation Checks in *JPSP* 2018

Our aim was not just to assess the prevalence of MCs, but also to differentiate between types of MCs or, if none was conducted, to assess why MCs were considered unnecessary.

### *Method*

***Journal selection and study sample.*** Hauser and colleagues (2018) found that, out of five journals included in a broader review, *JPSP:ASC* had the highest rate of MCs conducted; thus, the present sample can be expected to represent the upper tail of a distribution of scientific quality.[3] We diverge from, and go beyond, Hauser et al. (2018) in several ways. Our review emphasizes assets and diagnostic chances of MCs and its potential for improving the quality of research rather than problems of inadequate MCs. We distinguish between different MC levels that vary in adequacy, and we discuss reasons for not conducting MCs. We included all empirical articles published in 2018, comprising a total of 175 studies in 33 articles. One article was excluded because it did not involve empirical research. The list of articles and the coding rating procedure are documented at https://doi.org/10.5281/zenodo.4384127

***Coders.*** The three authors of the present article were randomly assigned to 5 months of the issues of 2018 and coded all articles contained in their assigned issues. This resulted in an overlap of 1 month for each pair of raters. A fourth rater coded all 12 months of articles.[4]

***Coding scheme.*** Variables of particular relevance were whether or not an MC was conducted; if not, why; and if so, what type of MC it was. We specified five reasons for not conducting an MC:

- *Correlational*: a study was merely correlational without any manipulation;
- *Not mentioned*: MC was simply ignored;
- *Dismissed*: Authors discussed the possibility, but decided deliberately not to include an MC;
- *Inherent in paradigm*: No MC was conducted because the manipulation's effectiveness was deemed to be necessitated by the paradigm;
- *Referred to other experiment/pretest*: MC was conducted in a pretest or accompanying experiment.

**Table 1.** Ratings Averaged Across Raters and All Studies/Articles

| MC classification | Studies in this category | Articles with at least one study in this category |
|---|---|---|
| Included MC (of some kind) | 40% | 50% |
| Included an MC subject to a *demand-effect* | 10% | 9% |
| Included *attention* check or instruction *recollection* | 5% | 6% |
| Included a *nondiagnostic* MC | 19% | 22% |
| Tried to realize a *diagnostic* MC | 5% | 9% |
| Did not include MC | 60% | |
| Because of *correlational* design | 15% | |
| MC was *not mentioned* | 38% | |
| Inclusion of MC explicitly *dismissed* in the study | 0% | |
| Apparently deemed MC to be *inherent in paradigm* | 6% | |
| Study *referred* to MC *in other experiment or pretest* | 1% | |

MCs that were conducted were coded into four categories:

- *Demand effect*: MC obviously revealed research intention or desired response;
- *Attention/recollection*: Participants were merely asked to recollect an instruction part or an attention-demanding task property;
- *Nondiagnostic*: MC affords actual test of the intended manipulation of the IV but without any attempt to rule out unintended effects on alternative causal factors.
- *Diagnostic*: The ideal case; the intended effect on a focal IV was assessed and unwanted effects on alternative variables were ruled out.

Positive MC-coding did not require authors to explicitly state the presence of an MC, we also coded MCs that were labeled or motivated differently.

## Results

Of 175 coded studies, 143 (82%) were experiments; the rest were correlational.

***Coder agreement.*** The primary variable (whether an MC was conducted or not) was coded at a high consensus rate of 87%, pooling across all pairs of coders who agreed on the main IV of a given experiment. Coding for the different types and quality levels of MCs and different reasons for not conducting an MC was more subjective, leaving more room for disagreement. Although the coding of demand-proneness proved to be difficult, the distinction between mere attention checks and all genuine MCs was accomplished at a very high consensus rate of 95% (excluding nine pairs or raters because of nonmatching IV identification).

***Prevalence of MC.*** The percentages reported in Table 1 are based on average ratings of either two or three coders, pooling across all articles and studies, respectively. Of all 143 studies, 71 (50%) included a validity check (attention check or MC) of some kind. In one experiment (1%), an MC in a related study was referred to instead of being included in the experiment itself. Regarding MC quality, only nine (6%) experiments were rated as containing a genuine MC, meaning that the MC tried to exceed demand-prone applications and mere attention or recollection-checks.

Roughly one half of the reviewed articles were judged as including some type of MC in at least one study. However, most of those were unsatisfactory—or even misleading—surrogates for MCs: They were either highly prone to demand effects or required mere attention to or recollection of the instructions. Only about a third of the articles included an MC that went beyond that superficial level. Very few were rated as diagnostic, making a deliberate attempt to also rule out some fundamental contenders for the manipulation's effect.

## Discussion

***Experimental versus correlational designs.*** Our review demonstrates that a great part of current psychological science relies on correlational (or quasiexperimental) designs. If studies do not involve a manipulated IV, MCs seem not applicable, and validity control in general seems to be no longer necessary. However, this inference is premature and unwarranted. Even for nonmanipulated variables such as gender, climate, or belongingness to ethical groups, it is essential to ensure that the variables of theoretical interest were operationalized in a sound and valid manner. The need to operationalize theoretical variables proper, rather than their spurious confounds, holds for IVs and DVs and for experimental and correlational

$$\Omega_{prior} \quad \times \quad \mathbf{LR} \quad = \quad \Omega_{posterior}$$

$$\frac{p(H_{true})}{p(H_{false})} \quad \times \quad \frac{p(D \mid H_{true})}{p(D \mid H_{false})} \quad = \quad \frac{p(H_{true} \mid D)}{p(H_{false} \mid D)}$$

**Fig. 1.** The logic of scientific inference in Bayesian odds notation. The prior odds ratio $\Omega_{prior}$ on the left is the ratio of the probability, $p(H_{true})$, that the focal theoretical hypothesis is true, divided by the complementary probability, $p(H_{false})$, that the hypothesis is false. $\Omega_{prior}$ highlights the need for a priori theorizing as it reflects the theoretical expectations before the assessment of empirical data $D$. The posterior odds $\Omega_{posterior} = p(H_{true} \mid D)/p(H_{false} \mid D)$ on the right indicate the updated ratio in the light of new data. The updating factor or likelihood ratio (LR) reflects diagnosticity – the ratio of $p(D \mid H_{true})$, the likelihood of $D$ given $H_{true}$ divided by the likelihood $p(D \mid H_{false})$ that another hypothesis accounts for the data $D$.

research. The latter distinction is often overemphasized (Fiedler, 2020) because no experimental treatment comes with the guarantee of a pure manipulation of the focal IV. As a rule, treatments can always induce variance in different variables, blurring the boundaries to correlational research.

***Attention check versus MC proper.*** An increasing portion of so-called MCs do not exceed the level of a simple attention check or instruction-recollection check. Such control devices merely assess whether participants have read and paid minimal attention to the instructions. As long as they do not ignore the instructions, they will be able to answer such an "MC." This must not be confused with a real check on the success of the experimental manipulation. Although a failure to meet an attention check certainly disqualifies a participant, a positive attention check tells us nothing about the manipulation's effectiveness. This can have dire consequences and cause serious misinterpretations of experimental results.

***Overcoming demand effects.*** A similarly radical conclusion pertains to demand-prone MCs, although these at least constitute a genuine attempt to assess the intended variation in the IV rather than assessing only participants' minimal cooperation. Demand characteristics in MCs may affect the participants' motivation and sensitize them to the research hypothesis, thus obscuring the evidence of the whole investigation (Hauser et al., 2018). However, recognizing the uselessness and dangerous side effects of demand-prone MCs does not mean, conversely, that MCs per se are dangerous and can be used only in exceptions. We argue, on the contrary, that unobtrusive MCs are generally possible and that the construction of clever and creative MCs that do not interfere with the study purposes is a major competence in good experimentation. When interference is unavoidable, MC can still be run in extra conditions that only serve to diagnose the treatment effects, without considering the contaminated DV. Or the effectiveness of manipulations can be tested in pilot studies or in only one of several experiments of a series. In any case, rather than discarding MCs as not feasible, researchers should tackle the problem and develop better MCs.

***Discriminant validity.*** Although a sizeable percentage (roughly 40%) of studies went beyond superficial attention checks and demand-prone compliance checks, there were hardly any ideal cases of truly diagnostic MCs. Quite a few experiments did include a clever check on the assumption that the manipulation did affect the focal IV, but only a very small portion of the best MCs reported attempted to rule out the possibility that a manipulation also exerted unintended effects on other variables, suggesting alternative accounts. Truly diagnostic MCs are particularly needed when complex manipulations simultaneously affect many variables.

## Immediate and Mediate Implications

The necessity of valid manipulation in the experimental context becomes evident when we illustrate inferences from experimental research by the Bayesian odds notation (Fig. 1). This formal exercise also helps in understanding the graded differences between MC types.

An ideal experiment provides new insights and theoretical evidence to the extent that it achieves a maximal change from prior odds to posterior odds. This increase is quantified by the likelihood ratio (LR) in the center of the Bayesian formula—the more extreme the LR, the stronger is the impact of a conducted study on the state of the art in the respective field of research. The ratio underlying LR highlights the contingency of the data on the theoretical origin, that is, the degree to which the focal hypothesis renders the obtained data more

likely than alternative hypotheses. Thus, with regard to the CLT hypothesis Δdistance → Δabstractness, an experimental design enables a diagnostic LR if the experimentally induced Δabstractness more likely reflects a manipulation in Δdistance than a confounded influence in some competing construct.

This Bayesian perspective on the logic of theoretical inference clarifies the differences between types of alleged MCs. A mere attention check may rule out the worst case (that $D$ is nothing but noise), but it is mute with regard to the theoretical inferences about the relative likelihood of $H_{true}$ versus $H_{false}$. Likewise, a demand-prone "MC" reflects participants' introspective report that $H_{true}$ (rather than $H_{false}$) was at work but is irrelevant for the theoretical assumption that $H_{true}$ actually generates $D$. In contrast, a genuine MC must enhance the likelihood that $D$ occurred when the premise of $H_{true}$ was met. An ideal, diagnostic MC entails the proof that the $H_{true}$ was more likely than the $H_{false}$ to underlie $D$. It should be obvious from this taxonomy of graded MCs that only the latter two devices (i.e., genuine and diagnostic MC) are of inferential value regarding the validity of theoretical hypotheses.

## *MC as catalyst for validity control and theoretical scrutiny*

Yet once researchers start to use MCs to foster LR (diagnosticity), they are inevitably sensitized to the logic of scientific inferences in general. Because any theoretical conclusion, $\Omega_{posterior} = p(H_{true}|D)/p(H_{false}|D)$, is the product of the LR and theoretical priors, $\Omega_{prior}$, it does not make sense to invest time and effort in MCs while neglecting the theoretical priors or other premises of $H_{true}$ and $H_{false}$. A truly useful MC must go beyond the effective manipulation of the intended IV. The validity of DV, or patterns of multiple IVs, are equally relevant. The ultimate purpose of advanced MC techniques is to allow research designers to produce data patterns $D$ in DVs that cannot be plausibly brought about by hypotheses other than $H_{true}$ (Fiedler, 2017; Platt, 1964). Frankly, and more offensively speaking, we hardly see how psychological research could be termed "scientific" if it ignores MCs and the associated validity concerns.

We expect MC to serve as a catalyst that triggers broader interest in validity issues, in research design, and logic of science. An enhanced focus on MC may be sufficient to cause a methodological snowball effect that triggers critical assessment and scrutiny of validation control. To improve the quality of research, it may not be necessary to implement new curricula to teach a plethora of methodological and theoretical lessons: It may be sufficient to focus on MCs, which forces researchers to become clever research designers and theoreticians. A validity-oriented tool that fosters diagnostic research designing and strict theorizing might just be the crucial step in overcoming the misguided focus on significance testing in order to fully arrive in the post–$p < .05$ era.

***Running clever MCs is worthwhile.*** Devising refined and creative MCs can be not only beneficial for science, but also rewarding for scientists, reviewers, editors, and teachers. Laudable investigations that successfully include clever and cogent MCs are likely to be respected and admired as positive models of strong behavioral science that deserve to be imitated. If a psychology teacher wants to provide an example of outstanding research, a stellar moment in empirical research that the scientific community can be proud of, could he or she really point to an experiment that is of questionable validity because it lacks a proper MC? Or, what criterion of excellent peer-reviewing competence is more important than sensitivity to validity issues and MCs? Could one imagine more exciting subject matters for methods seminars at the undergraduate, graduate, and postgraduate levels than validity debates relating, for example, to artifacts and invalid statistical inferences?

***Realistic expectations.*** These somewhat impassioned statements are not supposed to imply that experimental scrutiny is overly hard to accomplish. On the contrary, we argue that the basic MC idea is easy to understand and to acquire at a modest level. It takes only a simple item from a methods toolbox, easier to understand than a maximum-likelihood estimation or a mixed-model analysis of variance. A new developmental task calls only for a focus shift from significance testing to research designing and theorizing. Editors, reviewers, and readers of scientific journals must be sensitized only to the superiority of valid and theoretically sound research, beyond statistics.

Research by Verosky et al. (2018), which was among the coded studies, offers a nice down-to-earth example of a convincing MC. The main finding was that pairing faces with positive or negative behaviors gave rise to higher or lower likeability judgments, respectively, even when faces were not explicitly recognized at the judgment stage because of the extremely brief exposure time of 35 ms. To substantiate the theoretical assumption that faces had been charged with valence, the authors showed that the stimulus faces' acquired valence was also apparent in a later judgment stage involving unlimited presentation. Indeed, useful MCs can be very simple; they can just be an alternative DV as long as it is operationally independent of the focal DV.

An investigation in our own lab (Arslan & Fiedler, 2020) tested the hypothesis, derived from regulatory-focus theory (Higgins, 1997), that promotion focus (vs. prevention

focus) facilitates creative performance across a whole variety of creativity tasks. The ability to detect words related to promotion versus prevention (besides neutral target words) in a crossword puzzle served as an operationally independent MC. Note that the diagnosticity of this MC could be improved by including targets related to rival hypotheses. Future theories and paradigms may be tailored to enable particularly informative, diagnostic MCs.

Almost ideal MC conditions are implemented in so-called sampling approaches to judgment and decision research (Fiedler & Kutzner, 2015). Because the results of information search and environmental sampling processes are assessed independently of the final judgment and decision phenomena to be explained, the experimental task renders the hypothesized IV (i.e., sampling error and bias) amenable to direct empirical assessment, independently of the DVs measuring judgments and decisions (Denrell & Le Mens, 2012; Prager et al., 2018). It may be no coincidence that sampling theories are so illuminating and satisfactory and that the resulting empirical findings are so robust.

***Improving replication research.*** MCs inspire scrutiny in research design and foster enhanced theorizing; in addition, their beneficial side effects extend to further facets of methodology. Consider the essential role of replication research. As long as extensive and expensive replication projects (Camerer et al., 2018; Open Science Collaboration, 2012) are not contingent on MCs, replication failures remain equivocal and of highly questionable scientific value. It is so essential to distinguish between replication failures reflecting invalidity of a hypothesis $\Delta x \rightarrow \Delta y$ and the banal case that an ineffective manipulation failed to establish the premise, $\Delta x$. But granting that MC is a precondition of replications, why should it not be obligatory for all research?

## *Prospective recommendations for how to improve research and peer reviewing*

Our review of MCs suggests a number of ways in which the quality of research can be improved in future. Let us finally summarize our prospective suggestions and recommendations.

***Refining and improving MCs.*** Even when researchers do recognize the necessity to control for the validity of an experimental manipulation, MCs must exceed the minimally required preconditions (such as attention and compliance checks) to actually achieve this and avoid introducing confounds themselves. As our review of articles has shown, there is ample room for MC improvement. Good and excellent scientific practice requires a proactive, creative, and holistic approach to validity instead of stopping at the lowest level of quality checks and avoiding only the worst possible quality of data.

An optimal MC should not ask participants blatantly whether they were high or low in $x$ or exhibited $\Delta x$. Such self-reports are demand-prone and contingent on unwarranted assumptions about introspection (Wilson & Schooler, 1991). Although introspective ratings (e.g., of positive or negative mood states, promotion vs. prevention focus) need not be worthless, they are rarely sufficient. Such obvious measures directly reveal the experimenter's expectations or at least crucial aspects thereof. Asking for ratings of one's mood state reveals that the study focuses on the impact of mood on the experimental task, and this revelation can affect the participants' behavior. Unobtrusive behavioral measures afford much better MCs than verbal self-reports. For instance, "blind" raters may assess the affective appeal of participants' associations to keywords (e.g., "future," "youth," "window," etc.). Or participants' mood may be assessed from their facial expressions or through linguistic analyses of their verbal utterances.

***Beneficial side effects.*** Constructing good MCs can be an inspiring task and a prominent research goal in and of itself, with beneficial side effects for the development of new paradigms and procedural tools. For instance, using evaluative priming as an unobtrusive MC measure of positive versus negative affective states can lead to enlightening novel insights about priming mechanisms (see Unkelbach et al., 2008; Wentura et al., 2000). Ideally, a convincing MC strictly follows explicit theoretical assumptions. Thus, if one conceives of promotion focus versus prevention focus as a liberal versus conservative response strategy (in signal detection terms), respectively, assessing participants' response criteria in a simple detection task affords a compelling MC (Arslan, 2018).

Exploiting construct validity (Cronbach & Meehl, 1955; Westen & Rosenthal, 2003), researchers may base the control of a theoretical variable $x$ on the assessment of theoretically related variables, if $x$ itself is too difficult to assess. Construct validity relates positive mood to such variables as enhanced response speed, creativity, susceptibility to stereotyping, and false memories (Fiedler & Hütter, 2013). A profile of these related measures affords a highly informative check on the successful induction of good mood.

***Realistic expectations once more.*** Forcing each and every published study to contain any form of MC would certainly not help to improve but possibly hinder the quality of psychological research. Not always can an MC afford a one-to-one measure of relevant $\Delta x$. It is often sufficient to have a related measure ($\Delta x'$) that is imperfectly linked to $\Delta x$. A subtle and unobtrusive MC may be more

useful than an overly reliable but blatant MC that overshadows effects of $\Delta x$ on $\Delta y$.

Realistically, an MC is not required in every single experiment. Once the validity of a treatment is established in an initial experiment, a pilot study, or maybe even in former research conducted with the same paradigm, subsequent experiments can build on this preliminary work. A successful MC should be replicated occasionally but definitely not in every experiment. Alongside creative ways of conducting MCs, this should help avoid the types of MCs that Hauser and colleagues (2018) criticized as causing interference.

It should be admitted frankly that a perfectly diagnostic MC that rules out all possible side effects of a manipulation constitutes an ideal that can be approximated at best. One cannot jointly examine the entirety of all possible influences of an intervention. One should, however, effectively rule out theoretically relevant rival interpretations. In testing strong hypotheses against other theoretical approaches, it often becomes apparent which factors are crucial to carefully exclude and which other possible influences can be controlled by simpler procedures, such as randomness in the design. The ideal of a diagnostic design is most likely achievable in a *paradigm* (Fiedler, 2011), conceived as an auspicious research environment for testing a specific theory. Still, good research always entails the critical ability to be aware of the limitations inherent in any experiment, or even paradigm. Only then can research be conducted and appraised in a fashion that truly strives for steadily increasing quality and knowledge.

## Transparency

*Action Editors:* Travis Proulx and Richard Morey
*Advisory Editor:* Richard Lucas
*Editor:* Laura A. King
*Declaration of Conflicting Interests*
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
*Funding*
The work underlying the present article was supported by Deutsche Forschungsgemeinschaft Grant FI 294/29-1 (to K. Fiedler).

## ORCID iD

Klaus Fiedler https://orcid.org/0000-0002-3475-0868

## Notes

1. We set aside the details and depth of the philosophical debate concerning the method as a so-called demarcation criterion, of which a good overview is given by Andersen and Hepburn (2016; for a historical overview and a new perspective, see also Hoyningen-Huene, 2013).

2. Note that for an MC to be diagnostic, it is not sufficient to demonstrate, say, an effective shift in a regulatory-focus test (Polman, 2012); one must also ensure that other variables were not affected.

3. We initially chose *Psychological Science* for MC coding but decided to replace it with the *Journal of Personality and Social Psychology: Attitudes and Social Cognition* because *Psychological Science* published very few experiments and consisted almost exclusively of correlational research in the January issue of 2018.

4. We express our sincere gratitude to Janne Krippl, who completed this task.

## References

Andersen, H., & Hepburn, B. (2016). Scientific method. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Summer 2016 edition)*. https://plato.stanford.edu/archives/sum2016/entries/scientific-method

Arslan, P. S. (2018). *A dynamic perspective on self-regulation and adaptive strategy: The advantage of a regulatory shift* [Doctoral dissertation, Heidelberg University].

Arslan, P. S., & Fiedler, K. (2020). *Creativity depends on regulatory focus and, more strongly, on regulatory-focus shift* [Unpublished manuscript]. Heidelberg University.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*, *2*(3), 233–239. https://doi.org/10.1177/2515245919858072

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Denrell, J., & Le Mens, G. (2012). Social judgments from adaptive samples. In J. I. Krueger (Ed.), *Frontiers of social psychology. Social judgment and decision making* (pp. 151–169). Psychology Press.

Fiedler, K. (2011). Voodoo correlations are everywhere—Not only in neuroscience. *Perspectives on Psychological Science*, *6*, 163–171. https://doi.org/10.1177/1745691611400237

Fiedler, K. (2017). What constitutes strong psychological science? The (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, *12*(1), 46–61. https://doi.org/10.1177/1745691616654458

Fiedler, K. (2020). Elusive alpha and beta control in a multicausal world. *Basic and Applied Social Psychology*, *42*(2), 79–87.

Fiedler, K., & Hütter, M. (2013). Memory and emotion. In T. J. Perfect & D. S. Lindsay (Eds.), *The SAGE handbook of applied memory* (pp. 145–161). SAGE.

Fiedler, K., & Kutzner, F. (2015). Information sampling and reasoning biases: Implications for research in judgment and decision making. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (pp. 380–403). Wiley.

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*(6), 661–669. https://doi.org/10.1177/1745691612462587

Garner, W. R., Hake, H. W., & Eriksen, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, *63*(3), 149–159.

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology*, *9*, Article 998. https://doi.org/10.3389/fpsyg.2018.00998

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*, 1280–1300.

Hoyningen-Huene, P. (2013). *Systematicity: The nature of science*. Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*(1), 16–26. https://doi.org/10.1037/0003-066X.56.1.16

Krüger, T., Fiedler, K., Koch, A. S., & Alves, H. (2014). Response category width as a psychophysical manifestation of construal level and distance. *Personality and Social Psychology Bulletin*, *40*(4), 501–512.

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge University Press.

Lavine, M., & Schervish, M. J. (1999). Bayes factors: What they are and what they are not. *American Statistician*, *53*(2), 119–122. https://doi.org/10.1080/00031305.1999.10474443

Liberman, N., & Trope, Y. (1998). The role of feasibility and desirability considerations in near and distant future decisions: A test of temporal construal theory. *Journal of Personality and Social Psychology*, *75*(1), 5–18. https://doi.org/10.1037/0022-3514.75.1.5

Lilienfeld, S. O., & Waldman, I. D. (Eds.). (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions*. John Wiley & Sons.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3), 151–159. https://doi.org/10.1037/h0026141

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660.

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

Polman, E. (2012). Effects of self–other decision making on regulatory focus and choice overload. *Journal of Personality and Social Psychology*, *102*(5), 980–993.

Prager, J., Krueger, J. I., & Fiedler, K. (2018). Towards a deeper understanding of impression formation-New insights gained from a cognitive-ecological perspective. *Journal of Personality and Social Psychology*, *115*(3), 379–397.

Pyszczynski, T., Solomon, S., & Greenberg, J. (2015). Thirty years of terror management theory: From genesis to revelation. In M. P. Zanna & J. Olson (Eds.), *Advances in experimental social psychology* (Vol. 52, pp. 1–70). Academic Press.

Schmidt, F. (2010). Detecting and correcting the lies that data tell. *Perspectives on Psychological Science*, *5*(3), 233–242. https://doi.org/10.1177/1745691610369339

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*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). *p*-Curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681. https://doi.org/10.1177/1745691614553988

Soderberg, C. K. (2014). *The effect of psychological distance on abstraction: A meta-analysis of construal level theory*. University of California, Davis.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review*, *110*(3), 526–535. https://doi.org/10.1037/0033-295X.110.3.526

Trafimow, D. (2019a). Five nonobvious changes in editorial practice for editors and reviewers to consider when evaluating submissions in a post $p < 0.05$ universe. *The American Statistician*, *73*(Suppl. 1), 340–345.

Trafimow, D. (2019b). A taxonomy of model assumptions on which P is based and implications for added benefit in the sciences. *International Journal of Social Research Methodology*, *22*(6), 571–583.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463. https://doi.org/10.1037/a0020319

Unkelbach, C., Fiedler, K., Bayer, M., Stegmueller, M., & Danner, D. (2008). Why positive information is processed faster: The density hypothesis. *Journal of Personality and Social Psychology*, *95*, 36–49. https://doi.org/10.1037/0022-3514.95.1.36

Verosky, S. C., Porter, J., Martinez, J. E., & Todorov, A. (2018). Robust effects of affective person learning on evaluation of faces. *Journal of Personality and Social Psychology*, *114*(4), 516–528. https://doi.org/10.1037/pspa0000109

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$." *The American Statistician*, *73*(1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Wentura, D., Rothermund, K., & Bak, P. (2000). Automatic vigilance: The attention-grabbing power of approach-and avoidance-related social information. *Journal of Personality and Social Psychology*, *78*(6), 1024–1037.

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*(3), 608–618.

Wicklund, R. A., & Braun, O. L. (1987). Incompetence and the concern with human categories. *Journal of Personality and Social Psychology*, *53*(2), 373–382.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, *60*(2), 181–192.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056