Check for updates

Journal of Management
Vol. 52 No. 6, July 2026 2133–2165
DOI: 10.1177/01492063261424849
© The Author(s) 2026



Article reuse guidelines:
sagepub.com/journals-permissions

Controlling the Control Condition: A Critical Methodological Review of Control Conditions in Experimental Management Research

Johannes Stark
Christian Tröster
Kühne Logistics University (KLU)
Niels Van Quaquebeke
Kühne Logistics University (KLU)
University of Exeter Business School

Control conditions are essential to establishing causal relationships in experimental management research, yet they receive little attention compared to treatments. This study thus examines the current state of control-condition selection and design in top-tier management journals, reviewing 958 experiments from 421 study papers published from 2021 to 2023. Our review shows that researchers use true and pseudo-control conditions. True control conditions—such as no-treatment, all-but-treatment, and treatment-as-usual controls—provide a baseline for interpreting the effect of the treatment condition. In contrast, pseudo-control conditions (e.g., opposite-treatment-level or alternative-treatment designs) allow relative comparisons across conditions without providing a baseline. Notably, 20% of the studies we examined presented causal claims that were not supported by their designs, opening the risk of their results being misinterpreted and their effect sizes being exaggerated. These issues were further exacerbated by a lack of method transparency and construct validity. In response, we offer guidelines not only for primary study researchers to support the selection and design of control conditions, thereby enhancing transparency and yielding valid interpretations of causal claims, but also for research synthesists, reviewers, and editors to evaluate the same.

Keywords: *research methods; experiments; research design; control conditions*

Acknowledgments: The authors would like to acknowledge the help of two research assistants, Alexa Claes and Vinda Mohamed, and the participants of the management research seminars at UCL (UK) and the University of Groningen (NL).

Supplemental material for this article is available with the manuscript on the JOM *website.*

Corresponding author: Johannes Stark, Department of Management and Marketing, The Hong Kong Polytechnic University, 11 Yuk Choi Road, Hung Hom, Hong Kong.

E-mail: johannes.stark@polyu.edu.hk

Experiments in which participants are randomly assigned to either treatment or control conditions are considered the gold standard for identifying causal relationships in management research (Highhouse, 2009; Hill, Johnson, Greco, O'Boyle, & Walter, 2021; Shadish, Cook, & Campbell, 2002). As such, these experiments are often used to triangulate the effects of field studies and scrutinize their internal validity (Wellman, Tröster, Grimes, Roberson, Rink, & Gruber, 2023). Additionally, they allow researchers to study sometimes difficult-to-measure organizational processes (Bolinger, Josefy, Stevenson, & Hitt, 2022). Previous literature reviews have highlighted that experiments are widely used in various management disciplines, from organizational behavior (Thau, Pitesa, & Pillutla, 2014) and operations management (Lonati, Quiroga, Zehnder, & Antonakis, 2018) to strategy (Bolinger et al., 2022), entrepreneurship (Grégoire, Binder, & Rauch, 2019), and organization theory (Levine, Schilke, Kacperczyk, & Zucker, 2023).

In experiments, the effect of an independent variable (IV) on a dependent variable (DV) is determined by comparing a treatment condition that manipulates the IV (at the level or category of interest) with a control condition that serves as a baseline or comparison treatment to which participants are randomly assigned. This comparison allows researchers (a) to assess whether there is an effect of the IV on the DV and (b) to interpret the direction and size of this effect (Bailey, 2008; Lonati et al., 2018). However, widely cited guidelines for experimental research in management focus on treatment conditions far more than control conditions (e.g., Aguinis & Bradley, 2014; Bitektine, Lucas, Schilke, & Aeon, 2022; Bolinger et al., 2022; Highhouse, 2009; Podsakoff & Podsakoff, 2019; Schabram, Myers, & Hardin, 2024). As a result, researchers employ a wide variety of control-condition approaches, but this is problematic when the same treatment can exert very different effects depending on the type of control (Mullen & Monin, 2016; Schaerer, Du Plessis, Yap, & Thau, 2018). Additionally, if control conditions are not chosen with theoretical intent, studies risk making invalid causal inferences, reporting inflated effect sizes, and advancing fragile theoretical claims. These concerns not only complicate the evaluation of results but also risk undermining the credibility of experimental evidence and the resulting management policy advice.

In short, the control condition is not a secondary detail, but a core design choice that requires the same theoretical and methodological scrutiny as the treatment itself. With this in mind, we conducted a critical methodological review (Aguinis, Ramani, & Alabduljader, 2023) of how control conditions are selected and designed in experimental management research. Our goal is not only to synthesize existing practices but also to critically evaluate them against conceptual and methodological standards in related fields. We present a typology of true and pseudo-control conditions to guide this evaluation, which identified three recurring challenges in current research: overstated causal claims, limited transparency in reporting, and insufficient attention to construct validity. Our review of 958 experiments from 421 study papers published in top-tier management journals is as informative as it is worrying: In 20% of all experiments, authors make invalid or ambiguous causal claims that their experimental conditions do not allow for. Furthermore, our within-study meta-analysis shows that differences in the control conditions used can substantially bias the effect of a treatment condition by a factor of two. The potential for misinterpretations and biases creates the risk that management scholars are propagating misinformed theories and policies.

In light of this issue, our paper offers several key contributions. Firstly, our review of control conditions in experimental management research compiles existing knowledge and practices, as well as identifies frequent problematic selection and design choices. As stated, these choices can lead to erroneous causal claims that hinder theoretical advancement and

misinform managerial practice. Secondly, we develop a taxonomy of control-condition types to distinguish true control conditions from pseudo-controls, weighing their respective advantages and limitations. Notably, a true control condition is not merely the inverse of a given treatment; it must establish a theoretically reasonable counterfactual of what would have occurred in the absence of the treatment. Against that background, we explore when and how to implement true control conditions, as well as the circumstances in which it is appropriate to use pseudo-control conditions. Thirdly, we recommend best practices that enhance the construct validity of the control condition while ensuring transparency in its selection and design. Together, our guidance enables primary study researchers to improve the selection and design of control conditions in experimental management research, as well as support research synthesists, reviewers, and editors in the evaluation of the same.

Control Conditions

Among the landmark articles on how to conduct experimental studies in management (Aguinis & Bradley, 2014; Bitektine et al., 2022; Bolinger et al., 2022; Chatterji, Findley, Jensen, & Nielson, 2016; Croson, Anand, & Agarwal, 2007; Di Stefano & Gutierrez, 2019; Grégoire et al., 2019; Highhouse, 2009; Levine et al., 2023; Schabram et al., 2024; Stevenson & Josefy, 2019), authors have devoted little, if any, time to key aspects of control-condition selection and design (Lonati et al., 2018; Podsakoff & Podsakoff, 2019), such as construct validity or types of control conditions. While all these articles make important contributions to our understanding of experiments in management research, they tend to be heavily focused on treatments. For instance, Schabram et al. (2024) addressed how to manipulate (or prime) the IV in the treatment condition, but they ignored the role of the control condition. This omission is unfortunate because the control condition is not merely the mirror image of the treatment condition, but a carefully designed baseline or counterfactual against it. Schaerer et al. (2018) offered a notable exception to this status quo, explicitly engaging with the limitations of experimental design choices and the importance of control conditions for interpretability in studies that manipulate high vs. low power. Beyond that, papers on control-condition design in other disciplines, such as in clinical psychology and medicine (e.g., Gold et al., 2017; Michopoulos et al., 2021; Mohr et al., 2009), are informative but provide limited practical guidance for organizational research due to inherent conceptual and methodological differences.

Building on this foundation, we aim to extend the conversation by drawing attention to the un-/under-examined aspects of control conditions (i.e., variation in condition types, interpretation validity, evidence of construct validity, or transparency in design choices) across the diverse domains of management research. Our goal is to guide researchers in improving the rationale and transparency behind their control-condition decisions. To do so, we first synthesize and outline best practices to control-condition design and then use them as benchmarks to assess the articles included in our methodological review.

Types of Control Conditions

After gauging the methodological literature in other research fields (e.g., clinical psychology and medicine: Gold et al., 2017; Michopoulos et al., 2021; Mohr et al., 2009; statistics: Schickore, 2024), we identified two different approaches to designing control conditions: true and pseudo-control conditions (Table 1). True controls establish a counterfactual

Table 1
Definition of Control-Condition Types

		Definition	Micro Example *(IV: Power)*	Macro Example *(IV: CSR)*
	Treatment condition	Participants receive a stimulus that manipulates the treatment at the level or category of interest.	Participants write about a situation when they felt high power.	Participants read an acquisition proposal indicating a firm's high levels of CSR.
True control conditions	All-but-treatment	Participants receive a stimulus that omits the treatment but keeps everything else equal.	Participants write about their daily routine.	Participants read an acquisition proposal with a CSR section but no information on the firm's CSR.
	No-treatment	Participants receive no stimulus.	Participants skip the writing task and proceed to the DV measure.	Participants skip the reading and proceed to the DV measure.
	Treatment-as-usual	Participants receive a stimulus that manipulates the treatment at the real-world baseline level that is typical or expected within the study context.	Participants write about their felt power in a standard situation.	Participants read an acquisition proposal indicating a firm's average levels of CSR.
Pseudo-control conditions	Opposite-treatment-level	Participants receive a stimulus that manipulates the opposite end of the theoretically relevant spectrum (e.g., high vs. low; male vs. female) of the IV.	Participants write about a situation when they felt low power.	Participants read an acquisition proposal indicating a firm's low levels of CSR.
	Alternative-treatment	Participants receive a different treatment, i.e., a manipulation of a variable that is conceptually distinct from the focal treatment.	Participants write about a situation when they felt competent.	Participants read an acquisition proposal indicating a firm's digital transformation efforts.

Note. IV = independent variable; DV = dependent variable.

baseline against the treatment condition—what the DV would be without the treatment (Shadish et al., 2002). In contrast, pseudo-controls compare the treatment condition to another condition that is theoretically meaningful but does not provide a baseline for what the DV would have been without the treatment. The term pseudo-control is commonly used in adjacent disciplines such as medicine (Tenekedjiev et al., 2025); we adopt it here to classify comparable designs in management research. Again, drawing on the methodological literature in other disciplines (e.g., in clinical psychology; Mohr et al., 2009), we create a typology of true and pseudo-control conditions.

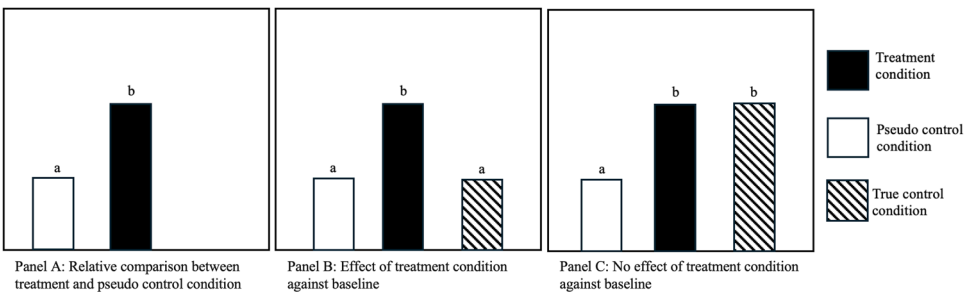
Among ***true controls***, we see three different types: *No-treatment controls* involve participants receiving no stimulus at all, such as no intervention in a field experiment or no prime in a lab study. *All-but-treatment controls* provide participants with a stimulus identical to the treatment except for the treatment component(s), thereby holding other aspects constant, such as time or effort. *Treatment-as-usual controls* provide participants with the real-world baseline level of the treatment variable typical in the study context (e.g., the amount of feedback typically given in an organization). Together, these three approaches provide a counterfactual against the treatment condition—what would have happened without it. In contrast, ***pseudo-controls*** enable a comparison of relative differences or alternative treatments. For instance, *opposite-treatment-level controls* expose participants to the same variable, but at different ends of a theoretically relevant spectrum (e.g., high vs. low feedback frequency; male vs. female). Likewise, *alternative-treatment controls* expose participants to a different treatment altogether, manipulating a variable that is conceptually distinct from the focal treatment (e.g., mentoring sessions instead of structured feedback). Notably, when researchers simply vary the dosage of a treatment, it is important to distinguish whether the reduced dosage represents simply a lower intensity of the same construct (e.g., high vs. low feedback frequency, which we classify as opposite-treatment-level) or whether the reduction alters the construct itself (e.g., a shortened leadership program that drops core elements and shifts focus to different skills, which we classify as alternative-treatment). Importantly, a control-condition's status as a meaningful counterfactual depends on the theoretical question and study context.

Interpretation Validity

The combination of a treatment condition and a particular type of control condition fundamentally determines the inferences that researchers can draw and the types of effects that a design can validly test.

Pseudo-control designs—those that pair a treatment condition with a pseudo-control condition but do not include a true control—allow researchers to test meaningful relative differences when the research question focuses on whether conditions differ, rather than on which condition drives the effect. For example, comparing high versus low levels of abusive supervision can illuminate their relative impact on stress (Podsakoff & Podsakoff, 2019). Such opposite-treatment-level designs are sometimes referred to as "donut designs," as they omit a baseline condition (Mullen & Monin, 2016). Similarly, comparing two active treatments (i.e., an alternative-treatment control), such as a mindfulness intervention versus a stress-relief intervention, can reveal which approach is more effective in reducing stress. In both cases, pseudo-control designs provide theoretically meaningful and often pragmatic contrasts (Figure 1a).

Figure 1
Interpreting Effects of a Treatment Condition With and Without a True Control Condition



Note. Example patterns of experimental results are displayed. Different subscripts indicate statistically significant differences. The true control condition establishes a baseline or as-usual counterfactual against the treatment condition, defined as the treatment level or category of interest. The pseudo-control condition does not establish a baseline for what the DV would have been without the treatment, but manipulates a different level or category of the IV or another theoretical variable, allowing for relative comparisons.

True control conditions are particularly well aligned with research questions that center on whether observed differences reflect harmful versus beneficial effects ("poison" vs. "cure"; Lonati et al., 2018). Returning to the abusive supervision example, only a well-designed true control condition can reveal whether high abusive supervision increases stress (Figure 1b), low abusive supervision reduces stress (Figure 1c), or both. Without such a baseline, these possibilities cannot be logically distinguished. Importantly, this limitation persists even under a strictly linear underlying effect: Two data points can establish that conditions differ, but they cannot reveal how each level deviates from the baseline. Linearity would resolve this ambiguity only if the low-level condition were positioned exactly midway between the baseline and the high condition. However, a high–low design alone cannot determine whether the low level is close to that baseline, far from it, or positioned symmetrically to the high condition and the baseline. Returning to our example, the low abusive supervision condition should be closer to the baseline than the high abusive supervision, assuming that the baseline of abusive supervision is low. A true control provides the reference point needed to estimate how each condition compares to its counterfactual; without it, comparisons speak only to relative differences rather than the effect of each condition.

This distinction is especially important when high and low levels represent conceptually distinct constructs rather than endpoints of the same continuum. Consider trust versus distrust: Distrust is not merely a smaller amount of trust but a qualitatively different psychological state (Lewicki, McAllister, & Bies, 1998). Identifying which level drives an effect can be theoretically and practically consequential because avoiding harm may require different actions than ensuring benefits. In the case of abusive supervision, for example, it matters whether stress arises because high abusive supervision is actively harmful, because low abusive supervision removes a harmful influence, or because both processes contribute. These scenarios imply different intervention priorities because the underlying causal patterns

differ—a distinction analogous to treating a poison versus administering a cure. Policies aimed at preventing high abusive supervision (e.g., sanctioning, monitoring, or disciplining harmful behaviors) are fundamentally different from policies aimed at promoting low abusive supervision (e.g., rewarding respectful day-to-day conduct or reinforcing supportive leadership behaviors). If stress is reduced only when supervisors operate at the low-abuse end, then preventing or sanctioning high-abuse episodes has little practical consequence. Conversely, if stress arises only from highly abusive supervision, then rewarding nonabusive or mildly supportive behaviors adds no meaningful benefit. Without a baseline condition to reveal the causal pattern, organizations cannot know which action category will be effective, risking an inefficient allocation of efforts and resources.

Taken together, both types of control conditions are valuable: Pseudo-controls support testing relative differences or theoretically motivated contrasts (Figure 1a), whereas true controls help clarify how each condition diverges from a baseline (Figure 1b–c). The central task for researchers is to select the design that best aligns with their research question and to articulate the scope of interpretation that their design supports.

Construct Validity of the Control Condition

When designing experimental and control conditions, researchers should take great care to achieve a high degree of construct validity—that is, the extent to which the operationalizations adequately reflect the constructs they are meant to represent (Grégoire et al., 2019; Shadish et al., 2002). To achieve a high level of construct validity, researchers need to not only carefully design a treatment but also create a control condition that helps isolate the treatment's effect on an outcome while ruling out confounding influences. To this end, researchers need to identify the unique properties of the treatment variable—or "things that only exemplars of the concept possess" (Podsakoff, MacKenzie, & Podsakoff, 2016: 18). These unique properties should be omitted in the control condition or manipulated at a different level. In contrast, features of the manipulation that do not reflect the treatment's core construct (e.g., expectations, cognitive load, positive/negative affect) should be held constant across treatment and control conditions to avoid introducing confounds. Moreover, researchers should design their treatment and control conditions to avoid evoking an "unfair comparison" (Werner, 1994: 105; Whiting, Podsakoff, & Pierce, 2008: 124), such as by keeping the strength of the manipulation (e.g., engagement) constant across conditions. By a similar token, researchers should avoid demand effects that favor one experimental condition above another one (Lonati et al., 2018).

After designing the experimental and control conditions, researchers need to empirically validate the manipulation of both, ideally by using pretests or pilot studies. These should examine whether the manipulations affect the intended construct (convergent validity) without unintentionally altering unrelated constructs (discriminant validity; Shadish et al., 2002). To establish the construct validity of their manipulation, researchers typically employ manipulation checks to test whether the treatment has a stronger effect on the independent variable than the control condition (Hauser, Ellsworth, & Gonzalez, 2018; Sigall & Mills, 1998). While such manipulation checks may support convergent validity, they do not safeguard against construct confounding (e.g., manipulating more than the intended construct; i.e., discriminant validity; Sigall & Mills, 1998). Since no manipulation can be expected to affect

only a single variable (Fiedler, McCaughey, & Prager, 2021), researchers should take great care to rule out potential confounding influences (Lonati et al., 2018). Otherwise, a causal claim cannot be made because it is unclear whether the independent variable or a confound caused the difference between the treatment and the control condition. For instance, researchers may test whether participants in the control condition have similar expectations, affective experiences, or levels of cognitive load compared to those in the treatment condition. This can help rule out potential confounds, but only to the extent that these aspects are not unique properties of the treatment itself. Thus, construct validation, whether for the treatment or the control condition, requires substantive theorizing and thorough quantitative specification (e.g., predictions about convergent and discriminant variables; Ejelöv & Luke, 2020).

Transparency in Control-Condition Selection and Design

Given the various selection and design choices for control conditions, authors should justify these choices by outlining the specific steps they take and the decisions they make. This is important because it enhances method transparency, which allows other authors to replicate studies and draw similar conclusions (Aguinis, Ramani, & Alabduljader, 2018). Specifically, there are three aspects of control-condition design that authors can address in the methods section: First, authors should provide a *diagnostic rationale*—that is, explain how a selected control condition helps to isolate the effect of the treatment on the outcome. The authors should specify which aspects of the treatment condition were held constant and which were altered in the control condition, as well as the measures taken to establish and test construct validity (see previous sections). This information is crucial for other researchers to evaluate, replicate, and extend the experimental findings. Second, authors should provide an *interpretation rationale* that articulates the interpretations allowed by a selected control condition. The authors should indicate whether they employ a true or pseudo-control design and what conclusions a reader may or may not draw from it, limiting the risk of overinterpretation. Third, scholars should consider a *relevance rationale* to help readers understand how a selected control condition represents a theoretically or practically relevant comparison to a treatment.

Method

Against the backdrop of the previously listed considerations, we conducted a critical review of control conditions in experimental studies that have been published in top-tier management journals. We sought to reveal weaknesses, contradictions, controversies, or inconsistencies by comparing each experiment to good standards of selecting and designing control conditions. In this way, we illustrate current challenges and generate knowledge that can aid future research in addressing those challenges (Aguinis et al., 2023). In conducting our review, we followed a systematic, transparent five-step process that we will explain in the following sections (Aguinis et al., 2018, 2023). Our final data set, a PRISMA flow chart, and a detailed coding scheme are provided on the Open Science Framework (OSF).[1]

Step 1: Scope of Review

Following a common approach to critical reviews (Paré, Trudel, Jaana, & Kitsiou, 2015), we included a broad and representative set of recently published articles to summarize

state-of-the-art practices and uncover present challenges in the selection and design of control conditions across management disciplines. We focused on experimental management research published in top-tier management journals from 2021 to 2023 to represent a broad range of the latest methods, as these can evolve rapidly (Aguinis et al., 2023). As such, we investigated how control conditions are *currently* used in experimental management research and *across* management disciplines.

Step 2: Journal Selection Procedure

We included 10 top-tier management journals (*Academy of Management Journal, Administrative Science Quarterly, Journal of Applied Psychology, Journal of Management, Management Science*[2] *Journal of Organizational Behavior, Organizational Behavior and Human Decision Processes, Organization Science, Strategic Management Journal, The Leadership Quarterly*) based on previous methodological reviews' journal selection (e.g., Bliese, Schepker, Essman, & Ployhart, 2020; Bolinger et al., 2022; Hill et al., 2021). These journals also align with *JOM*'s broad coverage of different management disciplines, allowing us to assess the state-of-the-art of the management discipline. Some of these journals fall at the intersection of management and applied psychology (*JAP*, *LQ*, *OBHDP*, *JOB*), but fortunately, this represents a management domain that frequently uses experimental methods.

Step 3: Article Selection Procedure

We used a manual search process to identify articles with experimental studies (search terms of manual screening: *experiment, manipulation, treatment, condition, intervention, random assignment, randomized controlled trial, causal effect*). Specifically, we read the titles/abstracts of all articles published in the selected journals from 2021 to 2023. If the methodology could not be unambiguously inferred from the title or abstract, we also examined the full text. We included articles reporting one or more experimental studies that used random assignment of participants to two or more conditions, including lab, field, online, and vignette experiments.

We did not include natural or quasi-experiments, which lack random assignment, often rely on naturally occurring groups rather than researcher-designed control conditions, and make different design choices (e.g., Connelly, Sackett, & Waters, 2013). If a paper included multiple experiments, we coded each experiment separately. The final list consists of 958 experiments from 421 study papers.

Step 4: Coding Scheme

We developed a coding scheme for control-condition types, as well as features involved in the selection and design of control conditions. We especially focused on interpretation validity, construct validity, and transparency. To evaluate and refine our coding scheme, we examined example studies for each category and refined our coding until it was exhaustive. The final coding scheme is provided on the OSF.[1] As explained previously, our coding scheme focused on the following aspects.

Control-condition type. Drawing on the methodological literature in management (e.g., Bolinger et al., 2022; Lonati et al., 2018) and clinical psychology (e.g., Mohr et al., 2009), we developed a coding scheme for true and pseudo-control conditions. We included three types of true control conditions (i.e., all-but-treatment, no-treatment, and treatment-as-usual control conditions) and two types of pseudo-control conditions (i.e., opposite-treatment-level and alternative-treatment conditions). When classifying control condition types, we relied on the reviewed authors' own stated intentions. In cases where no such intentions were reported, our coding was based on the coders' judgment, guided by the definitions outlined in the coding scheme. If an experiment encompassed multiple factors, each with various comparison conditions, we ensured that all conditions for each factor were coded.

Interpretation validity. We additionally assessed interpretation validity because the primary objective of experiments in management research is to isolate causal relationships (Highhouse, 2009; Hill et al., 2021; Shadish et al., 2002). We coded whether the authors provided a valid (vs. invalid) interpretation of their results, given the control design they used. Herein, we applied the following coding rule: If the authors used a pseudo-control design (i.e., an opposite-treatment-level or alternative-treatment condition without a true control condition), AND at least once interpreted the effect of the treatment condition (e.g., "the treatment condition increased/decreased Y"), we coded the interpretation as *invalid*. We coded the interpretation as *ambiguous* if the authors interpreted the effect of the treatment condition, but also used language that implied, without explicitly mentioning, a comparison group (e.g., "the treatment condition leads to more Y"). By contrast, we coded an interpretation as *valid* if the effect was consistently interpreted as the difference between the treatment and the pseudo-control condition (e.g., "the treatment condition has a larger effect on Y than the comparison condition") or as the effect of a treatment variable of which high and low levels were manipulated (e.g., "the treatment variable has a positive/negative effect on Y").

Construct validity. We also assessed construct validity, given its importance for interpreting an experimental manipulation. To account for the authors' construct validity considerations, we coded whether they conducted one or more discriminant manipulation checks to test whether their control condition differed from the treatment condition on more variables than the study variable (Ejelöv & Luke, 2020).

Transparency. Upon reviewing some initial studies, it became evident that there were significant variations in how authors justified their selection and design choices. Consequently, we also coded the transparency of these choices. As such, we evaluated whether the authors provided rationales for their selection of the control condition and the design of the study. Specifically, we coded the three rationale types described previously: diagnostic, interpretative, and relevance. In addition, we coded whether the authors justified their control-condition design by referring to another study within the paper or to a literature source.

Study characteristics. In addition, we counted the number of conditions per factor (23% of studies used more than two control conditions), the number of factors (37% of studies manipulated two or more factors, such as an IV and a moderator; half of those studies used

the same control-condition type for all of their factors), and the number of experiments per paper (54% of papers reported multiple experiments). We also extracted the experimental context and delivery method (11% field, 23% laboratory, 66% vignette, etc.), the experimental sample used (57% online panel, 22% students, etc.), the number of participants (median = 297), the dependent variable type, and the experimental stimuli. Moreover, we coded the research field of each paper using the research topic index of the *Journal of Management* (based on title and keywords).

Step 5: Coding Process

The coding was led by the first author, with support from the other two authors and two research assistants (as secondary coders). The first author coded the control conditions and the features of the control condition design in all studies, noting any ambiguities. Afterward, the authors discussed and resolved all ambiguities. Then, the two research assistants, trained by the first author using the coding scheme, each performed a second coding of 10% of the sample (randomly chosen, nonoverlapping subsample; $n_1 = 161$ and $n_2 = 128$). Second-coded control-condition designs were well above sample size recommendations (McHugh, 2012). Interrater agreements between the first author and the research assistants fell well within the range defined as substantial agreement (Landis & Koch, 1977) and aligned with reliability levels commonly accepted in top-tier management research (e.g., Fischer, Dietz, & Antonakis, 2017; Loignon & Woehr, 2018): Interrater agreement was $\kappa = .72$ ($p < .001$) and $\kappa = .74$ ($p < .001$) for control-condition types, $\kappa = .84$ ($p < .001$) and $\kappa = .86$ ($p < .001$) for transparency, and $\kappa = .80$ ($p < .001$) and $\kappa = .74$ ($p < .001$) for interpretation validity for the first and second subset of studies, respectively. The few disagreements were discussed and resolved through consensus between the first author and the research assistants following their independent coding.

Transparency statement about the use of AI: Parts of the manuscript were revised with the assistance of OpenAI's GPT-4.0 and GPT-5.0 to improve reading clarity. All content was reviewed and approved by the authors. The tool did not generate original research ideas, interpret data, or serve as a substitute for the authors' evaluation.

Results

In the following, we present a summary of our findings regarding control conditions, the validity of interpretations, construct validity, and standards for transparency and reporting. We provide a summary of our findings in Table 2 and, below that, a detailed review by control-condition type.

Types of True Control Conditions

We found that 38% of all reviewed studies used at least one true control condition per factor. Thus, about a third of experiments were designed to examine the effect of each factor's treatment condition on the DV.

All-but-treatment control condition. The most frequently used true control-condition type was the all-but-treatment control condition (28%). These controls are often used for

Table 2
Review of Control Conditions in Top-Tier Management Journals From 2021 to 2023

Control-condition type	$\bar{X}$	*AMJ*	*ASQ*	*JAP*	*JOB*	*JOM*	*LQ*	*MS*	*OBHDP*	*OS*	*SMJ*
True control conditions											
All-but-treatment	28%	30%	0%	33%	28%	36%	23%	33%	25%	25%	40%
No-treatment	14%	8%	0%	11%	9%	5%	23%	26%	13%	15%	36%
Treatment-as-usual	15%	18%	0%	13%	6%	36%	10%	14%	12%	29%	12%
Pseudo-control conditions											
Opposite-treatment-level	48%	63%	100%	57%	62%	55%	55%	33%	39%	61%	24%
Alternative-treatment	41%	41%	0%	27%	26%	18%	52%	40%	51%	28%	60%
Number of experiments	958	96	2	120	53	22	60	42	431	97	25
Experiments *not* including at least one true control condition per factor	62%	70%	100%	63%	75%	50%	70%	38%	63%	55%	32%
Experiments making invalid or ambiguous causal claims	20%	28%	50%	29%	34%	14%	7%	17%	16%	24%	24%

Note. % refers to the type of control conditions over all experimental studies in the respective journal. The sum of % may exceed 100%, as experimental studies often include multiple control conditions for multiple manipulated factors.
AMJ = academy of management journal; ASQ = administrative science quarterly; JAP = journal of applied psychology; JOB = journal of organizational behavior; JOM = journal of management; LQ = the leadership quarterly; MS = management science; OBHDP = organizational behavior and human decision processes; OS = organization science; SMJ = strategic management journal.

experimental treatments that involve tasks or activities (e.g., writing or reading) where participants in the control condition complete a similar activity without treatment-related instructions or information (e.g., a bogus task). For instance, in the treatment condition of one experiment on the psychological effects of economic threat, participants were instructed to imagine and write about their future lives to prime a future focus. By contrast, participants in the control condition were instructed to write about anything that came to mind (Sirola, 2023: 9). Another study tested the effects of social skills training on entrepreneurs' business performance. The authors compared two groups: the first received a social skills training module in addition to a baseline training; the second received only the baseline training, "holding constant the rest of the material taught" (Dimitriadis & Koning, 2022: 8636). A different study provided an example of a "placebo" control condition (Curhan, Overbeck, Cho, Zhang, & Yang, 2022). In an experiment on the effects of silent pauses in negotiations, participants in the treatment condition were told that making silent pauses (treatment) is a recommended strategy in negotiations. In contrast, participants in the control condition did not learn about making silent pauses; they were told that simply focusing on the instructed information (placebo treatment) is a recommended strategy in negotiations. Via this deception, the researchers could rule out demand effects because respondents in both conditions were led to believe that they received a treatment to improve their negotiation strategy and, thus, any difference between conditions should be explained by the effectiveness of making salient pauses.

When designing all-but-treatment control conditions, researchers should strike a balance between maintaining sufficient control and avoiding over-controlling, which could exclude parts of the causal mechanism the experiment aims to capture. For instance, one study on the effects of iterative coordination on team innovation kept friendly interactions with a mentor constant across conditions; however, the mentors asked iterative coordination questions in the treatment condition and asked nothing in the control condition (Ghosh & Wu, 2023). By the same token, researchers should take care to avoid introducing new confounds. In some cases, it was unclear whether a control stimulus isolated the unique features of the treatment condition or whether it inadvertently affected other theoretical constructs that may have independently influenced the DV. Consider, for example, a study examining the effects of COVID-19-induced death anxiety. The authors used a control condition involving dental pain, intended to match the COVID-related scenario in terms of intensity and aversiveness but without invoking mortality concerns (Xu, Dust, & Liu, 2023). However, dental pain is not a neutral stimulus. It represents a distinct theoretical construct—an alternative-treatment control condition—that may itself influence psychological outcomes. As such, this condition may not provide a baseline for isolating the effect of death anxiety; instead, it needs to be interpreted as a comparison of the anxiety between two distinct conditions. As described previously, our coding classified this as an all-but-treatment control condition based on the authors' stated design logic—specifically, their explicit aim to hold constant all elements unrelated to the treatment construct. Nonetheless, we flagged this design as conceptually ambiguous because the comparison stimulus may itself introduce additional mechanisms, thereby complicating causal interpretation.

No-treatment control condition. Fourteen percent of studies used a no-treatment control condition. This type of control condition appeared regularly in the few experimental intervention studies in our sample (i.e., only participants in the treatment condition received an

intervention). For example, in an experiment on performance feedback effects, participants in the treatment condition were told that they had won an award for their task performance, whereas participants in the control condition immediately proceeded to the next task without receiving any information about this (potential) reward (Deichmann & Baer, 2023: 5). Notably, some studies used a waitlist control condition as a specific type of no-treatment control. This control group serves as an untreated comparison group during the study, but eventually goes on to receive treatment at a later date. Waitlist control groups are often used when it would be unethical or undesirable to deny participants access to the treatment. For instance, in one experiment on entrepreneurship training, participants assigned to a waitlist control condition completed the training 1 year later than those in the treatment condition (Kotha, Vissa, Lin, & Corboz, 2023: 551). While the no-treatment control effectively creates a counterfactual scenario, it risks introducing confounding variables like demand effects or the effort exerted due to the manipulation or prime.

Treatment-as-usual control condition. Fifteen percent of our sample used a treatment-as-usual control condition. In one experiment, participants in the treatment group were informed about a company's egg-freezing policy. Conversely, participants in the control group received details on the company's parental leave policy, which was established as a typical work–life policy in a pre-study (Flynn & Leslie, 2023: 23). This parental leave policy acted as a counterfactual, allowing inference of the effects of the egg-freezing policy that might not have been observed otherwise. Although this "typical" policy is not a neutral comparison (i.e., it may itself carry evaluative or normative weight), it was chosen to reflect standard practice in the relevant context, based on pretesting. This approach avoided using a control condition in which companies are described as lacking work–life policies, which could have resulted in unjust comparisons. After all, such companies are rare and might unintentionally affect unrelated factors, such as the firm being perceived negatively. In short, this treatment condition creates a "fair" baseline for comparison (Cooper & Richardson, 1986).

In another experiment on the effects of compensation and unethical reciprocity, the authors manipulated compensation at a high level (treatment condition) and at the market wage level of their study population (control condition; Wang, Song, & Zhong, 2022: 2234). Here, the researchers could establish a baseline condition because they had empirical data on the typical salary of the respondents in the sample. Again, note that these experiments use the treatment-as-usual control to establish a baseline to assess what would have happened if the treatment (e.g., high wage) had not been present. However, in most cases, the researchers did not clarify how they ensured their control condition accurately reflected what is typical or expected within the specific study population. For example, in a study on the effects of leader perfectionism, an extremely perfectionistic (treatment condition) and not at all perfectionistic leader (opposite-treatment-level condition) were compared against a *normally* perfectionistic leader (control condition), who would "pursue perfectionism in most things but avoid being unnecessarily overcritical" (Xu, Liu, Ji, Dong, & Wu, 2022: 2098). However, the authors did not clarify why their description of leader perfectionism in the control condition reflected what participants would reasonably view as typical. Without such grounding, the validity of this treatment-as-usual control condition—and its role as a counterfactual—remains uncertain.

Types of Pseudo-Control Conditions

Most experiments in our review (62%) did not include a true control condition for each manipulated factor, but one (or more) pseudo-control conditions.

Opposite-treatment-level condition. Of all studies in our sample that included an opposite-treatment-level condition (48%), which manipulates a different level or category of the treatment variable, a majority of studies employed a pseudo-control design (81% of studies with an opposite-treatment-level condition)—that is, they did not include a true control condition for this factor. Pseudo-control designs can serve the purpose of testing whether an IV affects a DV (Podsakoff & Podsakoff, 2019). For instance, one experiment tested the effects of structural empowerment using a high versus low structural empowerment condition; the authors correctly interpreted their findings by claiming that "lower (vs. higher) social structural empowerment can stifle [. . .] employee psychological empowerment and performance" (Dennerlein & Kirkman, 2023: 1856). In another study, the authors interpreted investors' responses to female vs. male founders as "founder gender preferences," claiming that investors were "[less] interested in ventures with female founders than those with male founders" (Bapna & Ganco, 2021: 2699). Notably, it can be challenging to rule out confounding influences in opposite-treatment-level conditions, especially when constructs are poorly defined at low levels or are correlated with other constructs. Low CSR may, for example, be confounded with unethical practices in organizations. The authors of one study thus took great care to calibrate a low level of CSR, while ruling out confounding influences of unethical practices, by pointing out that the described company "has a few basic CSR initiatives" (Fehr, Gupta, & Guarana, 2021: 178).

Alternative-treatment condition. Of all experiments in our sample that used an alternative-treatment condition (41%), which compares the treatment against another theoretical construct, most did not include an additional true control condition (86% of studies with an alternative-treatment condition). There are cases where it may be more informative to test a treatment against an alternative treatment that has been shown to affect the DV rather than against the counterfactual of not receiving any treatment (Gallistel, 2009). For instance, in a study on the effects of ambidextrous leadership on exploration/exploitation behaviors (Klonek, Gerpott, & Parker, 2023), the authors used transformational leadership as an alternative-treatment condition instead of a neutral control condition. This control-condition choice is reasonable, as ambidextrous leadership theory was explicitly introduced as an innovation-specific leadership style that goes beyond the effects of transformational leadership on innovation (Klonek et al., 2023: 8). Hence, there was theoretical value in comparing ambidextrous leadership with transformational leadership. In a different study comparing human and AI performance feedback, the authors took great care to keep the strength of the manipulation (i.e., the content and format of the feedback) constant across conditions (Tong, Jia, Luo, & Fang, 2021: 1611). This allowed them to attribute the observed effects to the feedback source (AI vs. human), while ruling out alternative explanations such as differences in feedback quality, clarity, or valence.

Meta-Analysis of Within-Study Effects With Different Control Conditions

Next, we conducted a meta-analysis of within-study effects to assess the differences between true control and pseudo-control conditions in relation to the treatment. This analysis allowed us to eliminate potential confounding influences from between-study variables, such as different constructs, paradigms, and measures. The subsample of reviewed experiments included 105 studies that featured at least one true and one pseudo-control condition (the details on the data selection and coding process are provided on the OSF[1]). In studies that included an opposite-treatment-level condition ($k = 49$), the effects derived from comparing the treatment to a condition manipulating the IV at different levels were, on average, 42% larger than comparisons with a true control condition ($F(1, 156) = 47.57$, $p < .001$); this is likely because they manipulated an opposite extreme (vs. a realistic counterfactual). This difference became even more pronounced ($F(1, 126) = 87.76$, $p < .001$) when excluding studies that explicitly examined curvilinear effects where low- and high-level conditions were expected to be similar ($k = 40$). By contrast, in studies that included alternative-treatment conditions ($k = 58$), the effects derived from comparing the treatment to a condition manipulating another theoretical variable were, on average, 19% smaller than comparisons with a true control condition ($F(1, 179) = 111.77$, $p < .001$). A plausible explanation is that alternative-treatment conditions often manipulate constructs that influence the dependent variable in similar directions as the focal treatment, given that they are typically used to evaluate incremental effects beyond an existing intervention. In such cases, authors may bias the true effect by as much as a factor of two when their experiment utilizes a pseudo-control condition and no true control condition, as this design is unsuitable for their conclusions. Thus, depending on researchers' choice of control conditions, the experimental effect not only differs qualitatively in the sense of what it reveals, but also quantitatively in the sense of what effect size it may yield.

Challenges in Control-Condition Design

Conceptually, the challenges of interpretative validity, transparency, and construct validity are equally relevant for all control conditions, but in practice, their prevalence varied across control-condition designs in our sample. Notably, the most frequently used control-condition type (i.e., opposite-treatment condition) was also the most problematic: We observed invalid or ambiguous claims, as well as a lack of transparency and a rationale for construct validity, more often than expected by chance, according to χ^2 tests of independence. Table 3 summarizes the relationships between control-condition designs and the challenges they pose.

Interpretation validity. As mentioned previously, while pseudo-control designs (i.e., using only treatment and pseudo-control conditions) can yield meaningful insights, they do not allow researchers to isolate the effect of the treatment condition. Nonetheless, some authors appear to overstep the bounds of what such designs allow. In our review, we found that 20% of experiments made either invalid (11%) or ambiguous (9%) causal claims that were not supported by the design. Importantly, our critique is not that high–low comparisons are invalid per se; they can yield valuable insights about variation in the IV. The issue arises when researchers make claims about the effect of one condition without a neutral reference

Table 3
Association Between Control-Condition Types and Challenges

	Opposite-treatment-level	Alternative-treatment	All-but-treatment	Treatment-as-usual	No-treatment
Studies make invalid or ambiguous claims	144 (31%)↑	83 (21%)	17 (6%)↓	18 (13%)↓	6 (5%)↓
Studies do not provide a rationale for their control-condition design	318 (69%)↑	220 (56%)↓	166 (63%)	80 (57%)	83 (62%)
Studies do not address construct validity	400 (87%)↑	327 (83%)	206 (78%)↓	121 (86%)	109 (82%)
Studies do not include a confound check	417 (91%)	363 (92%)	241 (91%)	130 (92%)	128 (96%)↑

Note. Absolute and row-wise relative frequencies of challenges by control-condition type are shown. Arrows indicate whether challenges were more (↑) or less (↓) likely than expected to occur in studies with this control-condition type from a χ^2-test of independence.

point or use language that implies such effects. Take the example of a study on the "causal relationship between low power and self-promotional lying," which manipulated low versus high power by asking participants to imagine that they had few (vs. plenty) material resources (Li, Chen, & Hildreth, 2023: 1429). In this case, it is unclear why the authors did not include a true control condition to test the effects of low power against a "normal" case. In any case, such a design does not allow the researchers to conclude that low power increases lying because it could just as well be that high power decreases lying. Testing this would also be important because low and high power may reflect not only quantitative differences in the IV but also qualitative differences (Schaerer et al., 2018). Another study, testing the link between mentors' downward learning and their mentoring effectiveness, asked participants to reflect on a downward vs. upward learning experience but did not include a true control condition. The authors concluded that "reflecting on a downward learning experience increased mentor engagement" (Zhang, Wang, & Galinsky, 2023: 604). However, the observed difference in mentor engagement between the two conditions does not demonstrate that downward learning increased engagement. It could just as plausibly be that upward learning reduced engagement, or that both had effects in opposite directions. The only valid inference is a relative one: Downward learning was associated with higher mentor engagement than upward learning. Another example is a study on the effects of the "Angry Black Woman" stereotype, which compared vignettes featuring male vs. female employees (Motro, Evans, Ellis, & Benson, 2022). While certainly informative, a design that compares men to women does not allow one to determine whether the observed gender differences are due to women being rejected or men being favored—a theoretically relevant distinction (Phillips, Jun, & Shakeri, 2022). In this example, it is impossible to conclude that observing the anger of a female (and Black) individual has a negative impact on performance evaluations because it could just be that observing the anger of a male individual led to more positive performance evaluations.

The other half of problematic cases involved ambiguous causal phrasing. For example, in a study comparing reactions to high- versus low-energy coworkers, the authors concluded:

"Feeling energized by a coworker causes one to be more willing to help that coworker" (Grosser, Sterling, & Piplani, 2023: 1158). This claim is ambiguous. Based on the design, the authors can validly conclude that participants were more willing to help high-energy coworkers than low-energy coworkers. However, the stronger causal claim—that being around high-energy coworkers increases helping behavior—requires comparison against a baseline or neutral energy condition (i.e., a true control group). Without such a condition, it is equally possible that the low-energy coworker decreases helping behavior, or that both conditions influence helping behavior.

Transparency about the selection and design of the control condition. In most experiments (61%), the authors did not explain their choice and design of control conditions for at least one factor. In other cases (19%), the researchers justified their control-condition design by simply copying what previous studies had done, without providing any further explanation. However, simply doing what others have done is not a sufficient rationale. Thus, there is a high chance that authors copied control conditions they did not thoroughly understand. Only in 18% of experiments did the authors explain how their selection of the control condition helped to isolate the effect of the study variable on the outcome (i.e., *diagnostic rationale*). For instance, in a study on gender biases in personnel selection, the authors explained their diagnostic considerations in designing the female versus male conditions more generally: "Beyond the manipulated elements, we used the same precautions to keep the content identical across the conditions (i.e., location, clothing, and additional neutral background elements)" (Roulin, Lukacik, Bourdage, Clow, Bakour, & Diaz, 2023: 465). In another study on employees' perceptions of work–life policies, the authors argued: "We selected IVF and onsite childcare as comparisons because, like egg freezing, they are relatively uncommon (i.e., offered by <20% of companies) and are also linked to parenthood" (Flynn & Leslie, 2023: 32). These authors enhanced their hypothesis test by clearly explaining the reasoning behind their control-condition design and how it effectively eliminated potential confounds.

In a few experiments (6%), the authors specified what interpretations their selected control conditions allow for (i.e., *interpretation rationale*). For example, the authors of one study on the link between workplace power and engagement explained what their pseudo-control design could and could not test: "We can compare power levels only in a relative fashion and cannot distinguish the effects of wielding high power from those of wielding low power in these studies" (Williams, Lopiano, & Heller, 2022: 13). This example is valuable as it aids readers in grasping the range of interpretations that a study allows.

In yet fewer cases (5%), authors explained how selected control conditions represent a theoretically or practically relevant comparison to a treatment (i.e., *relevance rationale*). For example, one study argued for the external relevance of its control condition: "Participants in the control condition proceeded immediately to the negotiation materials. [This control condition] has high external validity as this is what negotiators would normally do" (Masters-Waage, Nai, Reb, Sim, Narayanan, & Tan, 2021: 197).

Construct validity of control conditions. Only 9% of experiments in our review included a test of whether the control condition affected other variables in the same manner as the treatment (i.e., a discriminant manipulation check; Ejelöv & Luke, 2020; Podsakoff et al., 2016). For example, in a critical incident study on workahomeism, the authors included a

discriminant manipulation check across the three conditions (workahomeism vs. presenteeism vs. resting at home). In each condition, participants were asked to imagine waking up in the morning and noticing that they did not feel well. To rule out illness severity as a confound, the authors measured participants' perceived illness intensity and symptoms, enabling them to test whether differences in health status could account for any effects across the three conditions: "We asked the participants to rate their illness in terms of strength and symptoms, which allowed us to test whether their health status could explain differences among workahomeism, presenteeism, and resting at home" (Brosi & Gerpott, 2023: 858). In some studies, the authors used a pretest to address diagnostic concerns in their control-condition selection. For example, in one study that manipulated a leader's voice in audio clips, the authors reported: "To avoid confounds, these audio clips were pretested for equivalence in clarity, professional tone, and fluent delivery. They were also of equal length and only varied in whether the leader communicated how they engage in high voice versus low voice and not on any other parameter" (Taiyi Yan, Tangirala, Vadera, & Ekkirala, 2022: 657). As an alternative, consider the study that investigated the effects of abusive supervision and LMX on self-blame (Tröster & Van Quaquebeke, 2021: 1798). In this case, the authors included a manipulation check that tested differences between the high and low LMX conditions on a measure of the LMX. However, they did not include a confound check, such as testing differences in leader hostility or leader effectiveness produced by the low LMX condition, which stated that the participants "do not have a good relationship" with their leader. Such a confound check would have allowed the authors to rule out confounding influences.

Differences Between Management Research Domains

We found differences across management domains in the use of study designs lacking a true control condition and making invalid claims about their findings. Even fields that used a relatively large number of experiments were not necessarily shielded from inaccurate interpretations. For example, *Organizational Behavior and Human Decision Processes (OBHDP)* published almost half (45%) of all experiments reviewed in this sample. Still, their invalid-interpretation rate (16%) was only slightly below the sample average (20%; Table 2). A positive exception was *Leadership Quarterly (LQ)*, which had the lowest invalid-interpretation rate (7%; Table 2).

Moreover, different research domains prioritized different types of true and pseudo-control conditions. Table 4 highlights the domains where findings were more frequently misinterpreted and where pseudo-control designs were disproportionately common. For instance, we found that alternative-treatment conditions without a true control condition were frequently used in human resource management (HRM; 63%) and behavioral economics (46%). While testing the effects of a theoretically or practically relevant variable may be informative, researchers should still include a true control condition to isolate which condition caused the difference in the outcome. In contrast, OB-related domains frequently relied on opposite-treatment-level conditions without incorporating a true control condition—for example, in studies on leadership (62%), positive and negative employee behaviors (60%), and DEI (58%). DEI studies often compared categories of a demographic characteristic (e.g., male vs. female, Black vs. White, homosexual vs. heterosexual). While this approach can be appropriate for testing hypotheses about group-based disparities, it does not allow researchers to

Table 4
Review of Control Conditions by Research Field

Control-condition type	$\bar{X}$	AE	BE	BP	CT	DEI	EM	ENT	GTN	HRM	L	OM	PM	PNB	SH	SR
True control conditions																
All-but-treatment	28%	32%	21%	30%	38%	41%	22%	37%	30%	32%	15%	18%	13%	21%	44%	30%
No-treatment	14%	10%	21%	7%	17%	1%	8%	21%	24%	12%	17%	41%	27%	8%	15%	12%
Treatment-as-usual	15%	18%	11%	44%	6%	10%	9%	5%	8%	2%	12%	14%	37%	13%	19%	21%
Pseudo-control conditions																
Opposite-treatment-level	48%	35%	29%	56%	23%	62%	62%	37%	41%	39%	68%	18%	42%	67%	30%	58%
Alternative treatment	41%	48%	61%	41%	36%	35%	35%	37%	43%	76%	41%	64%	29%	28%	19%	29%
Number of experiments	958	62	147	27	47	125	85	19	37	41	66	22	52	57	27	142
Experiments *not* including at least one true control condition per factor	62%	55%	65%	52%	51%	71%	76%	63%	57%	71%	76%	36%	38%	79%	30%	56%
Experiments making invalid or ambiguous causal claims	20%	24%	17%	26%	32%	19%	22%	21%	19%	15%	23%	9%	12%	23%	11%	25%

Note. % refers to the type of control conditions over all experimental studies in the respective research field. The sum of % may exceed 100%, as experimental studies often include multiple control conditions for multiple manipulated factors.
AE = Job attitudes, affect, emotions; BE = behavioral economics and decision analysis; BP = business strategy & policy; DEI = diversity, equity, and inclusion; EM = ethics & morality; ENT = entrepreneurship, innovation & creativity; GTN = groups, teams, and networks; HRM = human resource management; L = leadership; OM = operations management; PM = performance management; PNB = positive & negative employee behaviors; SH = work, stress, health, and well-being; SR = the social and relational context of organization.

determine whether the observed effects are driven by the disadvantage of one group, the advantage of another, or both—an ambiguity that can constrain theoretical interpretation (Phillips et al., 2022). While we recognize that testing against a counterfactual may be theoretically and practically irrelevant in such situations, we caution researchers to interpret their findings within the limits of what such a design can show. Meanwhile, treatment-as-usual control conditions were frequently used in business strategy (44%) and performance management (37%), whereas HRM (2%), ethics and morality (9%), creativity (6%), and entrepreneurship (5%) rarely used them. This may limit the psychological realism of experiments in these domains, placing respondents into control conditions that may not accurately reflect employees' typical experience. We can only speculate that studies in business strategy and performance management frequently involve variables where calibrating usual treatment levels is easier: Consider the examples of task difficulty (e.g., Raveendran, Srikanth, Ungureanu, & Zheng, 2023) or monetary incentives (e.g., Fest, Kvaløy, Nieken, & Schöttner, 2021), where researchers may calculate a medium level between high and low treatment levels. In contrast, defining a moderate treatment level for latent and multidimensional variables might be more challenging.

Implications And Recommendations

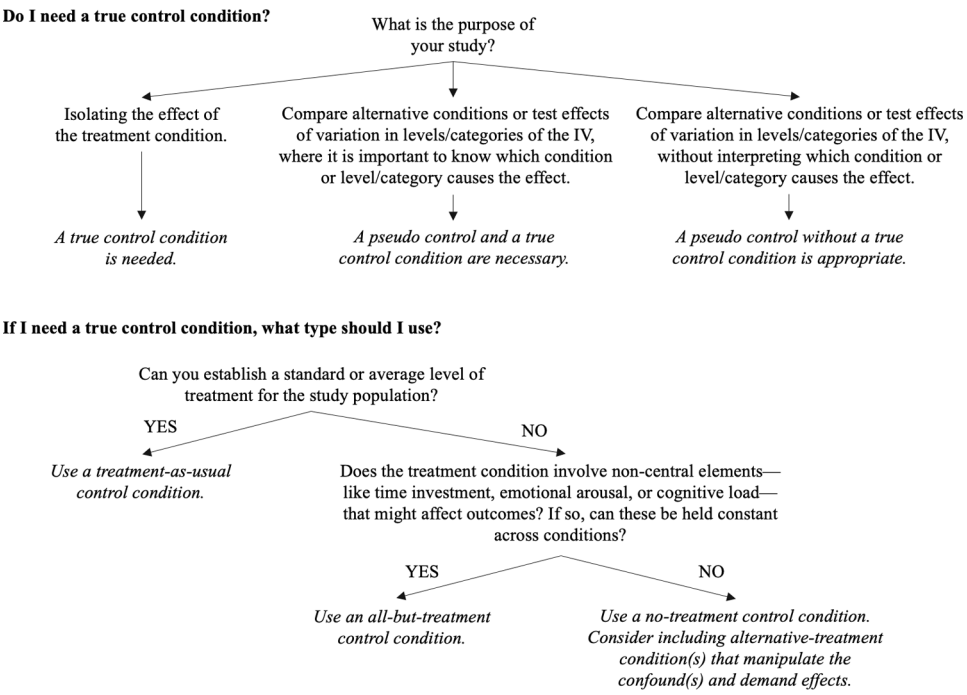
Our analysis of experiments published in top-tier management journals from 2021 to 2023 illuminated the potential to enhance the design and selection of control conditions in management research. Twenty percent of experiments in our sample made invalid or ambiguous causal claims that their control-condition designs did not permit. Furthermore, researchers often did not supply any justification (61%) for their control-condition choice and design. Even fewer studies (9%) tested whether the control condition differed from the treatment condition in ways other than the independent variable. In the following sections, we will review best-practice examples and develop recommendations for primary study researchers, research synthesists, reviewers, and editors.

Selecting the Right Control Condition(s)

Selecting appropriate control conditions is critical to drawing valid inferences from experimental designs. In the following, we will present a discussion of each design type's advantages and limitations, as well as a step-by-step decision tree that guides researchers in choosing suitable control-condition types (Figure 2).

Step one. The first step in selecting appropriate control conditions is to align the experimental design with the specific purpose of the study and the intended interpretation of its results (Antonakis, Bendahan, Jacquart, & Lalive, 2010). This begins by asking: Is the goal to test and interpret the effect of a treatment condition? If so, then it is essential to include a true control condition that provides a meaningful baseline—one that allows for isolating the effect of the treatment from other influences. A true control condition is also necessary if the goal is to compare two levels or categories of an IV or a treatment and an alternative treatment, as well as to be able to interpret the effect of each condition. By contrast, if the goal is to compare two (or more) conditions without distinguishing the effects of each condition, a

Figure 2
Decision Tree for Control-Condition Selection



pseudo-control design (i.e., an opposite-treatment or alternative-treatment condition without a true control condition) is appropriate.

Step two. If researchers determine that a true control condition is necessary, they will then need to select an appropriate format. We recommend that researchers first assess whether a treatment-as-usual control condition is feasible by establishing a standard level of the treatment variable that is appropriate for their study population.

When conceptually feasible, treatment-as-usual controls offer two key advantages over all-but-treatment and no-treatment controls. First, they enhance psychological and mundane realism by anchoring the comparison to participants' everyday experiences (Berkowitz & Donnerstein, 1982). For example, comparing a transformational leadership message to a "standard speech" that reflects typical rhetoric (as in Stock, Banks, Voss, Tonidandel, & Woznyj, 2023) ensures that the control condition is representative of naturally occurring stimuli. Second, treatment-as-usual controls can minimize differences in demand effects across conditions. Because these controls are contextually similar to the treatment, participants are less likely to infer condition-specific hypotheses that might otherwise arise from more artificial or structurally distinct control tasks. For instance, it is often more appropriate to equate the demand characteristics of a high-power writing task with those of an average-power writing task than with those of a power-unrelated writing task, as would be used in an all-but-treatment design (Lonati et al., 2018). When carefully calibrated, treatment-as-usual

controls support more naturalistic responses and meaningful contrasts, thereby strengthening both mundane realism and external validity.

Notably, what counts as "average" or "normal" may vary substantially across research contexts and populations. Thus, a treatment-as-usual condition must be defined by what is typical or expected within the specific study population. Importantly, while treatment-as-usual conditions aim to represent a standard or baseline, they should not be assumed to be affectively or evaluatively neutral. Participants' responses to these conditions may be meaningful and theoretically informative in their own right—for instance, by clarifying how deviations from the norm shape evaluations or behaviors. For some treatments, however, it may be difficult to construct a well-calibrated treatment-as-usual control. As our coding revealed, several studies featured treatment-as-usual controls that were poorly defined, inconsistently implemented, or reflected norms that varied markedly across contexts. In those cases, researchers risk inadvertently introducing substantial ambiguity into their experiments, thereby weakening internal validity and interpretability.

If a treatment-as-usual control cannot be clearly specified, researchers should consider the other two types of true control conditions. A no-treatment control omits the treatment entirely, whereas an all-but-treatment control holds all treatment-unrelated aspects constant while removing only the treatment's unique properties. To make this choice, researchers should assess whether the treatment includes elements that are not conceptually central but may nevertheless influence outcomes, such as time investment, emotional arousal, or cognitive load. When such treatment-unrelated aspects can be held constant, an all-but-treatment control is often preferable because it reduces the risk of confounds by equating extraneous influences across conditions. However, our results made clear that all-but-treatment controls can inadvertently introduce new theoretical content or procedural differences—for example, shorter instructions or lower cognitive engagement—if not carefully designed.

A no-treatment control should only be used if neither a treatment-as-usual nor an all-but-treatment control is viable. No-treatment controls are more exposed to confounding influences because treatment conditions typically involve more effortful, attention-grabbing, or emotionally engaging tasks. Consistent with the examples documented in our results section, several studies in our sample interpreted differences between treatment and no-treatment conditions as treatment effects, even though they were equally compatible with demand effects or other task-engagement differences unrelated to the treatment construct itself.

At the same time, constructing a true control condition that cleanly isolates the treatment's unique features can be more challenging than designing a pseudo-control. After all, holding all nonfocal elements constant without inadvertently introducing new influences is often difficult in practice. As reflected in several examples in our dataset, even minor differences in instructions, task duration, or emotional engagement can compromise construct purity or introduce meaningful confounds. When such confounds cannot be fully avoided, researchers may incorporate additional alternative-treatment conditions to isolate the treatment's unique causal component. For example, Kundro (2023) examined the effect of moral versus logistical framing on creativity. To isolate the moral dimension, the author constructed an alternative-treatment condition in which participants encountered the same constraints as in the treatment condition, but they were framed as logistical rather than moral issues. As such, the researcher could attribute observed differences specifically to the moral framing, rather than to the mere presence of constraints.

Feasibility considerations. The appropriateness of any control condition, whether true or pseudo, depends on both theoretical justification and practical feasibility. True controls offer one way to define a baseline reference point for comparison, but they may also come with feasibility limitations. One limitation is that they often yield smaller effect sizes than opposite-treatment-level controls. Therefore, comparisons against true controls (vs. opposite-treatment-level conditions) often require larger sample sizes to maintain adequate statistical power (Table S2; OSF[1]). Sequential analysis offers one solution to the feasibility issues created by large-*N* requirements. Here, researchers preregister a multiple-stopping rule that allows them to conduct interim significance tests during data collection and stop early if the effect is already significant (Lakens, 2014). Sequential designs require alpha adjustments to account for repeated testing, but the penalties are relatively modest compared to the potential savings in sample size, making this a pragmatic way to retain theoretically aligned designs without compromising interpretive validity. Other practical constraints further complicate the choice of control condition (Buchanan & Bryman, 2007). For example, combining multiple control conditions (true and pseudo) increases the number of experimental cells—and thus the number of participants needed. While such practical constraints can be considerable, our guidance emphasizes that theoretical considerations should always take priority over convenience in experimental design.

In some cases, a true control is neither feasible nor theoretically meaningful—for example, when studying characteristics such as gender or race, where imagining a counterfactual of "no gender" or "no race" might be conceptually fraught. In such situations, the researcher's goal may be to compare relative differences between categories. However, if isolating the effects of these conditions remains theoretically important, experiments can instead manipulate the underlying psychological mechanisms (e.g., stereotype activation; Shadish et al., 2002). For instance, one could cross male and female stimuli with a manipulation that makes stereotypes salient or not, thereby testing whether the observed effect is driven by male- or female-stereotype activation.

Complex scenarios. In practice, studies often feature complex scenarios, such as the inclusion of three or more conditions and multiple factors, or perhaps a complementary series of experiments. While the fundamental guidelines described previously also apply in those scenarios, there are additional considerations that researchers might find relevant.

Across our sample, 9% of experiments used three or more IV levels. When researchers manipulate three or more levels or categories of the IV in a single study, they need to carefully consider which contrast tests are required by their research question. For example, in a study with three levels of abusive supervision (high, low, and none), comparing high versus none (no-treatment control condition) tests whether high abusive supervision increases stress, while low versus none tests whether low abusive supervision reduces stress. Comparing high versus low only reveals whether high abusive supervision leads to more or less stress than low abusive supervision. In such cases, researchers should explicitly state which contrast(s) answer their theoretical question(s) to avoid mixing baseline inferences with relative comparisons.

Another scenario is an interactive design that manipulates a moderator (37% of our reviewed experiments). When researchers manipulate multiple factors in a single study, the control-condition design for each comparison should align with the researchers' specific theoretical question. If the researchers' goal is simply to test whether the IV's effect differs

across levels of the moderator, mixing a true control for one factor and a pseudo-control for the other is acceptable, as each comparison is independent. For example, if the question is whether the effect of leaders' warmth differs between high and low feedback, the control condition for leader warmth (IV) can be manipulated using an average leader (treatment-as-usual control), while feedback quality (moderator) can be manipulated with high- vs. low-quality feedback (opposite-treatment-level condition). Note that in this example, it is not possible to test whether the effect of leader warmth was caused by the high feedback quality condition, the low feedback quality condition, or both. For such a test, the moderator would have required a true control condition. In short, the inclusion control conditions depend on the research questions being tested. Consistency in control types mainly becomes important when researchers intend to compare the absolute sizes of the effects of both manipulated factors on the same conceptual scale. For instance, if leader warmth is compared to a neutral baseline (true control), but high feedback quality is compared to a low-quality baseline (opposite-treatment-level pseudo-control), then a statement such as "leader warmth has a stronger effect than feedback quality" is not warranted because the two effects are calculated relative to fundamentally different baselines.

Finally, about half (54%) of the papers in our sample included multiple experiments, raising important questions about the consistency of control conditions across studies and the meaning of aggregated effects (e.g., in internal meta-analyses). Varying control-condition designs across studies can be appropriate when each study tests a distinct theoretical question (e.g., moving from a simple treatment-control design to adding an alternative-treatment comparison in a later study). However, when integrating effects in an internal meta-analysis, researchers must account for heterogeneity in control-condition types. Importantly, true and pseudo-controls not only yield different effect sizes but also qualitatively different inferences: True controls reveal the effect of the treatment relative to its baseline, whereas pseudo-controls enable comparisons between two levels of the same construct or two constructs. Thus, mixing these effect types blends qualitatively different questions: "What is the treatment's effect?" versus "How do two levels or constructs compare?" Thus, it is crucial that internal meta-analyses avoid pooling effects across heterogeneous control conditions without accounting for these differences.

Maintaining Interpretation Validity

The significant rate of interpretive misalignment (20% of studies) highlights the need for precise claim calibration. Our aim is not to prescribe rigid wording norms but rather to stress the importance of logical alignment: (a) Methods should align with the stated hypotheses, and (b) interpretations should align with what the design can support. We argue that when causal claims are aligned with the type of control condition used, interpretation validity markedly improves. A key principle to keep in mind is that experiments produce evidence through comparisons between discrete groups, not through the estimation of continuous functional relationships. Accordingly, experimental results—specifically those comparing a high to a low condition—support conclusions about differences between conditions rather than justify inferences about linear trends across an assumed continuum. Therefore, interpretive precision is strengthened when conclusions focus on group comparisons rather than extrapolated trends, unless the design explicitly includes the procedures required to test

linear effects (McClelland, 1997). If practical constraints, such as large-*N* requirements, limit the available choices, researchers should not force causal claims that their design cannot sustain. Instead, they should interpret their findings transparently and acknowledge the limits of their design.

Establishing and Testing Construct Validity

The studies in our sample often interpreted the effectiveness of their manipulation based on a significant *p*-value, resulting from a test of the difference between the control and treatment conditions on a measure of the IV (similar to the findings of the review by Ejelöv & Luke, 2020). However, a significant *p*-value does not establish whether the control condition qualifies as a true control. Take anger manipulations as an example: A significant *p*-value merely indicates that anger levels in the control condition differ from anger levels in the treatment condition. It does not show whether the control condition omits the treatment or manipulates a baseline level. To establish the validity of a true control condition, researchers need to demonstrate that participants in the control group experienced no anger, or only a baseline level of anger, rather than merely a lower intensity.

In addition, we observed that few studies explicitly tested for discriminant validity—that is, whether the treatment and control conditions differ only in the intended manipulation and not in other theoretically meaningful ways. While such checks are generally advisable, they become particularly important when researchers use all-but-treatment or no-treatment control conditions. In these designs, there is a heightened risk that unintended constructs may be introduced through the comparison condition, especially when the goal is to hold treatment-unrelated features constant. Across several reviewed experiments, we could not determine whether the comparison stimulus affected only the targeted manipulation or also engaged additional theoretical dimensions. While we do not claim that such confounds were present, our review suggests that incorporating manipulation checks or auxiliary measures could help rule out alternative interpretations and increase confidence in the specificity of the observed effects. In the few cases where this was done (9%), researchers typically interpreted a nonsignificant difference as evidence of similarity. However, it is inappropriate to interpret nonsignificant *p*-values as evidence of equivalence (i.e., the absence of an effect; see Lakens, 2017; Wellek, 2017). That is, researchers may make unwarranted claims in favor of their control condition based on nonsignificant effects on discriminant variables (Ejelöv & Luke, 2020), for which they typically seek evidence of equivalence (i.e., equivalence testing; Lakens, 2017). When equivalence is needed, equivalence testing provides stronger evidence than nonsignificant *p*-values.

Being Transparent About Control-Condition Design

To improve transparency in control-condition design, we recommend that researchers clearly articulate three rationales by adapting this template: "The control condition was designed to hold constant [treatment-unrelated elements, e.g., task framing, format], while varying or omitting [treatment-relevant element, e.g., level of the independent variable or categorical contrast], to test the effect of [treatment construct] on [dependent variable] (*diagnostic rationale*). This design compares [treatment condition] to [control condition] and is intended to test [specific hypothesis, causal question, or theoretical contrast] (*interpretation*

rationale). The control condition is [theoretically / practically] relevant because [justification, e.g., it reflects real-world norms, offers a meaningful conceptual comparison, or isolates competing mechanisms] (*relevance rationale*)."

Of course, this template is not intended as a rigid reporting format. Instead, we hope to prompt researchers to think more systematically about the logic of their design and its implications for causal inference. To illustrate how these rationales can be articulated, consider the following adaptation based on a study reviewed in our sample (Flynn & Leslie, 2023), which used a treatment-as-usual control condition: "The control condition was designed to hold constant the organizational context and communication format, while varying the policy content—specifically, offering egg-freezing benefits (treatment) versus a standard parental leave policy (control)—to test the effect of nontraditional fertility benefits on perceived organizational support (diagnostic rationale). This design compares the egg-freezing policy condition to a treatment-as-usual condition based on pretested perceptions of typical work–life policies. It is intended to test whether adding a novel reproductive benefit enhances perceptions of employer support (interpretation rationale). The control condition is practically relevant because parental leave is widely viewed as a standard workplace benefit, providing a realistic and meaningful baseline for assessing reactions to more novel policies (relevance rationale)." This example illustrates how researchers can make the logic of their control-condition design more transparent, thereby strengthening both methodological rigor and interpretive clarity. This same structure applies to pseudo-control designs when the theoretical aim is to compare levels or mechanisms rather than isolate a baseline effect.

Implications for Different Stakeholders

Our review has implications for primary study researchers (primary stakeholders), as well as for research synthesists, funders and grant writers, reviewers, and editors (secondary stakeholders; Aguinis & Gibson, 2025; Simsek, Li, & Huang, 2022).

Primary Study Researchers. Table 5 summarizes our recommendations for researchers collecting primary data. Across the studies we reviewed, inappropriate or unclear control-condition choices frequently contributed to interpretive ambiguity. Primary study researchers should therefore ensure tight alignment between (a) their hypotheses, (b) what their design can test, and (c) how they interpret their findings. Explicitly aligning specific contrasts with theoretical questions is especially important when using designs with multiple levels, multiple factors, or multiple studies. Our guideline can also support students and early-career researchers in developing a more principled and rigorous approach to selecting and justifying control conditions.

Research synthesists. Our within-study meta-analysis indicated that effect sizes differ systematically across true and pseudo-control conditions; moreover, these differences reflect qualitatively different inferences. True controls estimate the effect of a treatment relative to a baseline, whereas pseudo-controls estimate relative differences between levels of a construct or between two constructs. Consequently, variation in control-condition type across studies can limit comparability of reported effect sizes (e.g., Cohen's d, Hedges's g) and, if unaccounted for, bias meta-analytic conclusions. We therefore encourage research

Table 5
Summary of Implications and Recommendations

Selecting, designing, and interpreting control condition(s):

- Align your experimental design with your research question (e.g., Are you aiming to test the effect of a treatment condition and therefore need a true control condition?).
- Use the decision tree to select appropriate true and/or pseudo-control conditions.
- Avoid interpreting isolated effects of the treatment condition in studies without a true control condition (e.g., Do NOT write: The treatment condition had a positive/negative effect). Instead, you may interpret it in relation to the pseudo-control condition (e.g., The treatment condition had a more positive/negative effect compared to the pseudo-control condition).
- Experiments yield evidence by comparing groups, so conclusions are clearer when framed as group comparisons.

Establishing and testing construct validity:

- Establish "fair comparisons" between your treatment and control conditions by keeping the strength of the manipulation constant across conditions.
- Avoid over-controlling, which could exclude parts of the causal mechanism the experiment aims to capture.
- Be aware of ill-defined mid and lower levels of a treatment variable when designing opposite-treatment-level and treatment-as-usual conditions.
- If your control condition does not keep all potential confounds constant, consider including an alternative-treatment condition that manipulates them.
- Use equivalence testing (a) to assess whether the control condition does not manipulate the treatment (e.g., a measure of the treatment is equal to a value indicating the absence of it) and (b) to test that the treatment and control affect related but unintended variables in a similar manner.

Being transparent about control-condition design:

- Explain how a selected control condition helps to isolate the effect of the study variable on the outcome, e.g., by specifying what aspects were held constant between conditions.
- Highlight what interpretations a control condition allows or does not allow for. For example, does the experiment include a true control condition to establish an effect of the treatment?
- When using an alternative-treatment condition, explain why it represents a theoretically or practically relevant comparison to a treatment.
- Consider using and adapting our template for improved transparency in control-condition design.

synthesists—that is, those conducting meta-analyses or literature reviews to explicitly consider control-condition type in their analyses, for example, by treating it as a potential moderator in meta-analytic models or as an explicit condition for their interpretation of the bigger picture.

Funders and grant writers. Our findings also have implications for the writers and evaluators of grant proposals. Because true control designs often yield smaller effect sizes than opposite-treatment-level or other pseudo-control designs, they typically require larger sample sizes to maintain adequate statistical power. This makes study costs higher for designs that include true controls (see our illustrative power analysis in the supplemental material; Table S2). An awareness of these cost implications can help grant writers justify budget requests and help funders more accurately assess the resource requirements of proposed research.

Reviewers and editors. Our analysis suggests that journals differ in the frequency of misaligned causal interpretations: For example, *Leadership Quarterly* (*LQ*) had the lowest invalid-interpretation rate in our sample (7%). While we can only speculate about the reasons, *LQ*'s explicit emphasis on rigorous methodological practices—including guidance on selecting appropriate counterfactuals—is likely a factor (Wulff et al., 2023). We hope that our guide helps reviewers and editors evaluate whether authors have (a) selected an appropriate control condition, (b) interpreted their findings in a manner aligned with what their design can support, and (c) provided sufficient transparency to prevent misinterpretation.

Conclusion

Scientific understanding, like visual perception, relies on contrast: We only see objects against a background. Yet in experimental research, the focus often rests solely on the treatment (the "object") while the control condition (the "background") becomes an afterthought. Our review of 958 experiments in top-tier management journals (from 2021–2023) suggests that control conditions often receive insufficient theoretical and methodological scrutiny. Hopefully our review serves as a wake-up call: Control conditions should be selected with theoretical intent, designed with methodological precision, and reported with full transparency. After all, it's not just what you test, but what you test against, that defines what you find.

ORCID iDs

Johannes Stark https://orcid.org/0000-0002-3526-911X
Christian Tröster https://orcid.org/0000-0001-8092-5182
Niels Van Quaquebeke https://orcid.org/0000-0002-6141-4659

Notes

1. https://osf.io/8uehf/?view_only=ed723a8a73674bf589a8156f5a7eba33
2. To maintain a focus on management research, we only included experiments published in the entrepreneurship/innovation, organizations, operations management, and business strategy sections.

References

Aguinis, H., & Bradley, K. J. 2014. Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, 17: 351-371.

Aguinis, H., & Gibson, B. C. 2025. Making waves: How to improve scholarly impact performance through stakeholder engagement. *BRQ Business Research Quarterly*, 28: 326-335.

Aguinis, H., Ramani, R. S., & Alabduljader, N. 2018. What you see is what you get? Enhancing methodological transparency in management research. *Academy of Management Annals*, 12: 83-110.

Aguinis, H., Ramani, R. S., & Alabduljader, N. 2023. Best-practice recommendations for producers, evaluators, and users of methodological literature reviews. *Organizational Research Methods*, 26: 46-76.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. 2010. On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21: 1086-1120.

Bailey, R. 2008. *Design of comparative experiments*, vol. 25. Cambridge, UK: Cambridge University Press.

Bapna, S., & Ganco, M. 2021. Gender gaps in equity crowdfunding: Evidence from a randomized field experiment. *Management Science*, 67: 2679-2710.

Berkowitz, L., & Donnerstein, E. 1982. External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist*, 37: 245-257.

Bitektine, A., Lucas, J., Schilke, O., & Aeon, B. 2022. Experiments in organization and management research. In M. Hitt (Ed.), *Oxford research encyclopedia of business and management*. New York, NY: Oxford University Press. doi:10.1093/acrefore/9780190224851.013.284

Bliese, P. D., Schepker, D. J., Essman, S. M., & Ployhart, R. E. 2020. Bridging methodological divides between macro-and microresearch: Endogeneity and methods for panel data. *Journal of Management*, 46: 70-99.

Bolinger, M. T., Josefy, M. A., Stevenson, R., & Hitt, M. A. 2022. Experiments in strategy research: A critical review and future research opportunities. *Journal of Management*, 48: 77-113.

Brosi, P., & Gerpott, F. H. 2023. Stayed at home—But can't stop working despite being ill?! Guilt as a driver of presenteeism at work and home. *Journal of Organizational Behavior*, 44: 853-870.

Buchanan, D. A., & Bryman, A. 2007. Contextualizing methods choice in organizational research. *Organizational Research Methods*, 10: 483-501.

Chatterji, A. K., Findley, M., Jensen, N. M., Meier, S., & Nielson, D. 2016. Field experiments in strategy research. *Strategic Management Journal*, 37: 116-132.

Connelly, B. S., Sackett, P. R., & Waters, S. D. 2013. Balancing treatment and control groups in quasi-experiments: An introduction to propensity scoring. *Personnel Psychology*, 66: 407-442.

Cooper, W. H., & Richardson, A. J. 1986. Unfair comparisons. *Journal of Applied Psychology*, 71: 179-184.

Croson, R., Anand, J., & Agarwal, R. 2007. Using experiments in corporate strategy research. *European Management Review*, 4: 173-181.

Curhan, J. R., Overbeck, J. R., Cho, Y., Zhang, T., & Yang, Y. 2022. Silence is golden: Extended silence, deliberative mindset, and value creation in negotiation. *Journal of Applied Psychology*, 107: 78-94.

Deichmann, D., & Baer, M. 2023. A recipe for success? Sustaining creativity among first-time creative producers. *Journal of Applied Psychology*, 108: 100-113.

Dennerlein, T., & Kirkman, B. L. 2023. The forgotten side of empowering others: How lower social structural empowerment attenuates the effects of empowering leadership on employee psychological empowerment and performance. *Journal of Applied Psychology*, 108: 1856-1880.

Di Stefano, G., & Gutierrez, C. 2019. Under a magnifying glass: On the use of experiments in strategy research. *Strategic Organization*, 17: 497-507.

Dimitriadis, S., & Koning, R. 2022. Social skills improve business performance: Evidence from a randomized control trial with entrepreneurs in Togo. *Management Science*, 68: 8635-8657.

Ejelöv, E., & Luke, T. J. 2020. "Rarely safe to assume": Evaluating the use and interpretation of manipulation checks in experimental social psychology. *Journal of Experimental Social Psychology*, 87: 103937.

Fehr, R., Gupta, A., & Guarana, C. 2021. Rewarding morality: How corporate social responsibility shapes top management team compensation votes. *Organizational Behavior and Human Decision Processes*, 167: 170-188.

Fest, S., Kvaløy, O., Nieken, P., & Schöttner, A. 2021. How (not) to motivate online workers: Two controlled field experiments on leadership in the gig economy. *The Leadership Quarterly*, 32: 101514.

Fiedler, K., McCaughey, L., & Prager, J. 2021. Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science*, 16: 816-826.

Fischer, T., Dietz, J., & Antonakis, J. 2017. Leadership process models: A review and synthesis. *Journal of Management*, 43: 1726-1753.

Flynn, E., & Leslie, L. M. 2023. Progressive or pressuring? The signaling effects of egg freezing coverage and other work–life policies. *Journal of Applied Psychology*, 108(1): 1-26.

Gallistel, C. R. 2009. The importance of proving the null. *Psychological Review*, 116: 439-453.

Gold, S. M., Enck, P., Hasselmann, H., Friede, T., Hegerl, U., Mohr, D. C., & Otte, C. 2017. Control conditions for randomised trials of behavioural interventions in psychiatry: A decision framework. *The Lancet Psychiatry*, 4: 725-732.

Ghosh, S., & Wu, A. 2023. Iterative coordination and innovation: Prioritizing value over novelty. *Organization Science*, 34: 2182-2206.

Grégoire, D. A., Binder, J. K., & Rauch, A. 2019. Navigating the validity tradeoffs of entrepreneurship research experiments: A systematic review and best-practice suggestions. *Journal of Business Venturing*, 34: 284-310.

Grosser, T. J., Sterling, C. M., & Piplani, R. S. 2023. Energized people in prominent places: Political support networks, relational energy, and employee innovation implementation. *Journal of Organizational Behavior*, 44: 1145-1163.

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. 2018. Are manipulation checks necessary? *Frontiers in Psychology*, 9: 998.

Highhouse, S. 2009. Designing experiments that generalize. *Organizational Research Methods*, 12: 554-566.

Hill, A. D., Johnson, S. G., Greco, L. M., O'Boyle, E. H., & Walter, S. L. 2021. Endogeneity: A review and agenda for the methodology-practice divide affecting micro and macro research. *Journal of Management*, 47: 105-143.

Klonek, F. E., Gerpott, F. H., & Parker, S. K. 2023. A conceptual replication of ambidextrous leadership theory: An experimental approach. *The Leadership Quarterly*, 34: 101473.

Kotha, R., Vissa, B., Lin, Y., & Corboz, A. V. 2023. Do ambitious entrepreneurs benefit more from training? *Strategic Management Journal*, 44: 549-575.

Kundro, T. G. 2023. The benefits and burdens of work moralization on creativity. *Academy of Management Journal*, 66: 1183-1208.

Lakens, D. 2014. Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44: 701-710.

Lakens, D. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8: 355-362.

Landis, J. R., & Koch, G. G. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 33: 363-374.

Levine, S. S., Schilke, O., Kacperczyk, O., & Zucker, L. G. 2023. Primer for experimental methods in organization theory. *Organization Science*, 34: 1997-2025.

Lewicki, R. J., McAllister, D. J., & Bies, R. J. 1998. Trust and distrust: New relationships and realities. *Academy of Management Review*, 23: 438-458.

Li, H., Chen, Y. R., & Hildreth, J. A. D. 2023. Powerlessness also corrupts: Lower power increases self-promotional lying. *Organization Science*, 34: 1422-1440.

Loignon, A. C., & Woehr, D. J. 2018. Social class in the organizational sciences: A conceptual integration and meta-analytic review. *Journal of Management*, 44: 61-88.

Lonati, S., Quiroga, B. F., Zehnder, C., & Antonakis, J. 2018. On doing relevant and rigorous experiments: Review and recommendations. *Journal of Operations Management*, 64: 19-40.

Masters-Waage, T. C., Nai, J., Reb, J., Sim, S., Narayanan, J., & Tan, N. 2021. Going far together by being here now: Mindfulness increases cooperation in negotiations. *Organizational Behavior and Human Decision Processes*, 167: 189-205.

McClelland, G. H. 1997. Optimal design in psychological research. *Psychological Methods*, 2(1): 3-19.

McHugh, M. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22, 276-282.

Michopoulos, I., Furukawa, T. A., Noma, H., Kishimoto, S., Onishi, A., Ostinelli, E. G., Ciharova, M., Miguel, C., Karyotaki, E., & Cuijpers, P. 2021. Different control conditions can produce different effect estimates in psychotherapy trials for depression. *Journal of Clinical Epidemiology*, 132: 59-70.

Mohr, D. C., Spring, B., Freedland, K. E., Beckner, V., Arean, P., Hollon, S. D., Ockene, J., & Kaplan, R. 2009. The selection and design of control conditions for randomized controlled trials of psychological interventions. *Psychotherapy and Psychosomatics*, 78: 275-284.

Motro, D., Evans, J. B., Ellis, A. P., & Benson, L., III. 2022. Race and reactions to women's expressions of anger at work: Examining the effects of the "angry Black woman" stereotype. *Journal of Applied Psychology*, 107: 142-152.

Mullen, E., & Monin, B. 2016. Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology*, 67: 363-385.

Paré, G., Trudel, M. C., Jaana, M., & Kitsiou, S. 2015. Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52: 183-199.

Phillips, L. T., Jun, S., & Shakeri, A. 2022. Barriers and boosts: Using inequity frames theory to expand understanding of mechanisms of race and gender inequity. *Academy of Management Annals*, 16: 547-587.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. 2016. Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organizational Research Methods*, 19: 159-203.

Podsakoff, P. M., & Podsakoff, N. P. 2019. Experimental designs in management and leadership research: Strengths, limitations, and recommendations for improving publishability. *The Leadership Quarterly*, 30: 11-33.

Raveendran, M., Srikanth, K., Ungureanu, T., & Zheng, G. L. 2023. How do performance goals influence exploration-exploitation choices? *Organization Science*, 34: 2464-2486.

Roulin, N., Lukacik, E. R., Bourdage, J. S., Clow, L., Bakour, H., & Diaz, P. 2023. Bias in the background? The role of background information in asynchronous video interviews. *Journal of Organizational Behavior*, 44: 458-475.

Schabram, K. F., Myers, C. G., & Hardin, A. E. 2024. Manipulation in organizational research: On executing and interpreting designs from treatments to primes. *Organizational Research Methods*. Advance online publication. doi:10.1177/10944281241300952

Schaerer, M., Du Plessis, C., Yap, A. J., & Thau, S. 2018. Low power individuals in social power research: A quantitative review, theoretical framework, and empirical test. *Organizational Behavior and Human Decision Processes*, 149: 73-96.

Schickore, J. 2024. *Controlled experiments*. Cambridge, UK: Cambridge University Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. New York City, NY: Houghton Mifflin.

Sigall, H., & Mills, J. 1998. Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review*, 2: 218-226.

Simsek, Z., Li, N., & Huang, J. L. 2022. Turbocharging practical implications in management studies. *Journal of Management*, 48: 1083-1102.

Sirola, N. 2023. Going beyond the call of duty under conditions of economic threat: Integrating life history and temporal dilemma perspectives. *Organizational Behavior and Human Decision Processes*, 179: 104292.

Stevenson, R. M., & Josefy, M. 2019. Knocking at the gate: The path to publication for entrepreneurship experiments through the lens of gatekeeping theory. *Journal of Business Venturing*, 34: 242-260.

Stock, G., Banks, G. C., Voss, E. N., Tonidandel, S., & Woznyj, H. 2023. Putting leader (follower) behavior back into transformational leadership: A theoretical and empirical course correction. *The Leadership Quarterly*, 34: 101632.

Taiyi Yan, T., Tangirala, S., Vadera, A. K., & Ekkirala, S. 2022. How employees learn to speak up from their leaders: Gender congruity effects in the development of voice self-efficacy. *Journal of Applied Psychology*, 107: 650-667.

Tenekedjiev, K., Panayotova, D., Daboos, M., Ivanova, S., Symes, M., Panayotov, P., & Nikolova, N. 2025. Quasi-experimental design for medical studies with the method of the Fuzzy Pseudo-control group. *Applied Sciences*, 15: 2076-3417.

Thau, S., Pitesa, M., & Pillutla, M. 2014. Experiments in organizational behavior. In M. Webster Jr. & J. Sell (Eds.), *Laboratory experiments in the social sciences*: 433-447. Cambridge, MA: Academic Press.

Tong, S., Jia, N., Luo, X., & Fang, Z. 2021. The Janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal*, 42: 1600-1631.

Tröster, C., & Van Quaquebeke, N. 2021. When victims help their abusive supervisors: The role of LMX, self-blame, and guilt. *Academy of Management Journal*, 64: 1793-1815.

Wang, L., Song, F., & Zhong, C. B. 2022. High compensation and unethical reciprocity. *Journal of Management*, 48: 2223-2254.

Wellek, S. 2017. A critical evaluation of the current "p-value controversy". *Biometrical Journal*, 59: 854-872.

Wellman, N., Tröster, C., Grimes, M., Roberson, Q., Rink, F., & Gruber, M. 2023. Publishing multimethod research in AMJ: A review and best-practice recommendations. *Academy of Management Journal*, 66: 1007-1015.

Werner, J. M. 1994. Dimensions that make a difference: Examining the impact of in-role and extrarole behaviors on supervisory ratings. *Journal of Applied Psychology*, 79: 98-107.

Whiting, S. W., Podsakoff, P. M., & Pierce, J. R. 2008. Effects of task performance, helping, voice, and organizational loyalty on performance appraisal ratings. *Journal of Applied Psychology*, 93: 125-139.

Williams, M. J., Lopiano, G., & Heller, D. 2022. When the boss steps up: Workplace power, task responsibility, and engagement with unpleasant tasks. *Organizational Behavior and Human Decision Processes*, 170: 104140.

Wulff, J. N., Sajons, G. B., Pogrebna, G., Lonati, S., Bastardoz, N., Banks, G. C., & Antonakis, J. 2023. Common methodological mistakes. *The Leadership Quarterly*, 34: 101677.

Xu, L., Liu, Z., Ji, M., Dong, Y., & Wu, C. H. 2022. Leader perfectionism—Friend or foe of employee creativity? Locus of control as a key contingency. *Academy of Management Journal*, 65: 2092-2117.

Xu, M., Dust, S. B., & Liu, S. 2023. COVID-19 and the great resignation: The role of death anxiety, need for meaningful work, and task significance. *Journal of Applied Psychology*, 108: 1790-1811.

Zhang, T., Wang, D. J., & Galinsky, A. D. 2023. Learning down to train up: Mentors are more effective when they value insights from below. *Academy of Management Journal*, 66: 604-637.