The Leadership Quarterly 34 (2023) 101677

Contents lists available at ScienceDirect

The Leadership Quarterly

journal homepage: www.elsevier.com/locate/leaqua

Common methodological mistakes ☆

Jesper N. Wulff [a], Gwendolin B. Sajons [b], Ganna Pogrebna [c,d,e], Sirio Lonati [f], Nicolas Bastardoz [g], George C. Banks [h,*], John Antonakis [i]

[a] Aarhus University, Aarhus, Denmark
[b] ESCP Business School, Berlin Campus, Germany
[c] AI and Cyber Futures Institute at Charles Sturt University, Bathurst, Australia
[d] University of Sydney, Sydney, Australia
[e] Alan Turing Institute, London, UK
[f] NEOMA Business School – Reims Campus, Reims, France
[g] KU Leuven, Antwerp, Belgium
[h] UNC Charlotte, Charlotte, NC, USA
[i] University of Lausanne, Lausanne, Switzerland

ARTICLE INFO

Keywords:
Open science
Causality
Counterfactuals
Methodology
Endogeneity

ABSTRACT

For scientific discoveries to be valid—whether in theory or empirically—a phenomenon must be accurately described: The scientist must use appropriate counterfactuals and eliminate competing explanations. Empirical work must also use an appropriate design and method, and empirical claims made about the phenomenon must be correctly characterized. Moreover, valid empirical discoveries must be reliable in the sense that scientists who reexamine the data must be able to reproduce the finding or to replicate the effect from data gathered in a similar context. Only discoveries adhering to the above criteria can be scientifically informative, serve as building blocks for theory, or have policy implications. Unfortunately, as several recent surveys of the literature show, much of the published works in the management and applied psychology fields are uninformative; contributing reasons include several intractable problems in the study design and analysis as well as the failure of the field to adopt open science practices. Against this backdrop, we identify common methodological mistakes made in applied work. We group these mistakes into three major categories: (a) study design and data collection (e.g., fit between hypotheses and methods, design, measurement, open science, literature reviews), (b) data analysis (e.g., data preprocessing, choice of estimators, analysis of data, issues concerning endogeneity, and use of instrumental variables), and (c) diagnostics, inferences, and reporting. We also explain how to avoid these issues, so that published work makes for a useful contribution to the scientific record.

To discover how the world works, scientists seek to uncover causal relations among systems of variables. To that end, they should provide a fair test of their hypotheses with appropriate counterfactuals. As scientists accumulate reliable evidence, they build theories. These edifices of knowledge are composed of facts and explanations—to use Forscher (1963) analogy—"bricks."

Some sciences get it right. Physics is at a point where theory is so strong that it can, for the most part, properly guide and evaluate empirical discoveries; no theoretical physicist believed the results of badly executed experiments, where neutrinos were incorrectly thought to go faster than the speed of light (see Reich, 2012). The prowess of the medical sciences was evident when they deployed, in record time, life-saving mRNA COVID-19 vaccines using basic and applied findings from scientists in various fields on the basis of decades of carefully done paradigmatic experiments and theoretical insights (Dolgin, 2021).

Of course, studying phenomena in the social sciences is complex. But so is the human body, the brain, the immune system, weather systems, geophysics, and other phenomena. Does this complexity stop scientists in these fields from building rigorous theories that explain a particular phenomenon? It is obvious that human behavior, perceptions, and decisions are also complex; but they are not random. Instead, they are governed to some degree by systematic variation due to genes, individual differences, culture, preferences, and the social environment. Although not deterministic, this variation can be modeled in some probabilistic framework using a robust scientific design.

Authors note: Authors are listed in reverse alphabetical order. They contributed equally to the design and drafting of this article. We wish to thank Mikko Rönkkö for comments on an earlier draft.

* Corresponding author.
E-mail address: gbanks3@uncc.edu (G.C. Banks).

https://doi.org/10.1016/j.leaqua.2023.101677
Available online 21 February 20231048-9843/

The bricks of scientific knowledge must be carefully made and assembled in the context of some architectural plan; with time, knowledge across disciplines, from physics, to chemistry, to neuroscience, to psychology, to economics, even to the arts, can be put together in a cohesive and integrative way (Wilson, 1998). In each discipline, though, it is essential to ensure that the bricks are not flawed or badly assembled. Alas, as lamented by Forscher (1963) many decades ago, and as echoed by others more recently, the publication game rewards the mass production of "bricks.".

The problems identified by Forscher (1963) are salient in the social sciences as well as in domains of management and applied psychology (Antonakis, 2017; Banks et al., 2016; O'Boyle, Banks, & Gonzalez-Mule, 2017; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). Some branches of the social sciences are practicing their trade better than are others. Economists, for instance, have well-established theoretical insights of the assumptions and behaviors of agents, they have widely accepted and proven econometric methods to guide estimation techniques, and training provided to junior scholars follows a reasonably standardized set of principles (cf. Antonakis, 2017). Interestingly, economists and biologists have many theoretical commonalities (e.g., game theory, signaling theory), and estimation methods (e.g., instrumental-variable regression/Mendelian randomization), suggesting that independent rigorous thinking led to common paradigms in theory and methods.

Why do we face so many problems in knowledge production in other areas of discovery, especially in the social sciences? One of the base disciplines in our field, psychology, currently has many demons to face, following a spate of replication failures (Open Science Collaboration, 2015) among other challenges. Economics appears to be on somewhat sounder footing, and has a higher replicability rate in experimental findings compared to psychology (Camerer et al., 2016). Moreover, researchers seem to have higher confidence in the replication of studies within economics—and much less so with respect to education and psychology—according to prediction markets (Gordon et al., 2020). These recent findings should stimulate us to think: are there ways to improve our discipline? The answer to this question is not obvious and probably multifaceted. Here is a non-exhaustive list of issues we think have probably contributed to the problem:

1. The incentive structures are wrong; they reward quantity of publications, not quality, and disjointed findings (see Antonakis, 2017; Bergh, Perry, & Hanke, 2006; Fanelli & Larivière, 2016; Grote, 2017; Ioannidis, 2016; Larivière & Costas, 2016; Nosek, Spies, & Motyl, 2012).
2. Open science practices have not been fully embraced (Aguinis, Banks, Rogelberg, & Cascio, 2020). Discoveries should only be accepted if they are robust and reliable. Moreover, other scientists should be able to reproduce and replicate these findings, ideally using registered reports (i.e., where the front end of the manuscript is reviewed prior to data being gathered and findings are published regardless of how the cards fall). Unfortunately, though, most journals do not incentivize replications. They fetishize novel and significant results to the detriment of robust and reliable, paradigm-driven discoveries (Antonakis, 2017; Bakker, van Dijk, & Wicherts, 2012). Journals operating "business-as-usual" also expect a theoretical contribution (see next point)! Current practice—incentivizing novel and theoretical contributions and frowning on replications and null results—does not help the field learn what is a valid or invalid discovery.
3. We operate from a weak paradigm (cf. Pfeffer, 1993). Our theories are feeble and untested (Edwards & Berry, 2010; Edwards, Berry, & Stewart, 2016; Edwards & Christian, 2014), oftentimes no more than tautologies and circular arguments, and findings cannot be cohesively assembled (Antonakis, 2017). Builders simply cannot properly evaluate the quality of the bricks because students are not taught formal theory, making it hard for them to see what makes for a theoretical contribution (Antonakis, 2017). Yet, many top journals require a theoretical contribution. Ironically, the propositions—more generally, conjectures that are not derived from first principles or proven theoretical frameworks—in the most cited articles in the flagship theoretical journal of the Academy, the *Academy of Management Review*, are hardly ever tested (Edwards et al., 2016).
4. Students in management and psychology programs—unless strongly quantitatively oriented—are generally not provided the mathematical undergirding to understand the notion of an analytical "proof" and to evaluate whether an estimator does the job appropriately (i.e., asymptotically converges to the true value). An estimator proven to be correct asymptotically (i.e., a "consistent estimator") estimates the parameter of interest accurately. Yet, unproven estimators, statistics, or modeling approaches are often used in a ritualistic fashion in our field (Antonakis, Bendahan, Jacquart, & Lalive, 2010; Rönkkö, McIntosh, Antonakis, & Edwards, 2016). In addition, several surveys of the literature show that the endogeneity problem is rife, and most findings are, simply put, untrustworthy (e.g., see Antonakis et al., 2010, 2021; Fischer, Dietz, & Antonakis, 2017; Hamilton & Nickerson, 2003).
5. The very basis of our bricks, the measures, are bad; in fact, despite most of the field thinking otherwise, we do not generally measure behaviors but perceptions and/or evaluations of behaviors. Behaviors and perceptions are not isomorphic; perceptions are outcomes and not independent variables. Perceptions are driven in part by omitted variables, which if not included in the model will lead to erroneous results. Consequently, we have a massive theoretical confounding and an intractable endogeneity problem in estimation (Antonakis et al., 2016; Banks, Woznyj, & Mansfield, 2021; Fischer, Hambrick, Sajons, & Van Quaquebeke, 2020; Fischer & Sitkin, 2022).
6. Qualitative methods of research are usually not applied in a scientific manner (McDermott, 2023; see also Wright, 2017). That is, counterfactual evidence is usually lacking—comparisons over time and space are often not made (Gerring & McDermott, 2007)—and empirical observations are haphazard and idiosyncratic: it is very difficult to reliably report on or to evaluate a finding regarding phenomena studied under such a research method (Antonakis, 2017; McDermott, 2023). There are no common standards and no ultimate arbiters, akin to proofs, for how analyses should be conducted.

As a result of all the above problems, our field, unfortunately, rewards brickmakers to raucously dump piles of badly made bricks in heaps of rubble. It does not have to be this way. This journal has taken a strong position, making it clear that we will not continue business as usual; we did not care what price we would pay, whether in terms of number of submissions or whether citations to the journal would drop. All we focused on, restlessly, was to publish robust and reliable science (Antonakis et al., 2019). We made it very clear what kinds of articles we solicited, and that we would distinguish our journal from others by ensuring we concretize the sacred tenets of modern science: to publish discoveries that correctly characterize the phenomenon, that are reproducible, replicable, and transparent, and that use the correct estimation techniques (Antonakis et al., 2019).

And, although our chief focus was to publish robust and reliable science, we were pleasantly surprised to see—and this, beyond our wildest expectations—that success did find its way to us. Our two-year impact factor more than tripled in a short time span. In terms of impact factor, we now rival general journals (see Antonakis, 2023)—doing slow science pays off! This focus continues with the editorial team of George Banks.

There is so much to be done to correct the problems of the past; it is now common to see articles in our journal that question the "received wisdom"—badly formulated theories that were established and taught

as if fiat. One by one they face demolition, whether via wrecking-ball type reviews or empirical work (e.g., Alvesson, 2020; Arvate, Galilea, & Todescat, 2018; Banks, Fischer, Gooty, & Stock, 2021; Bechtoldt, Bannier, & Rock, 2018; Fischer et al., 2021; Gottfredson, Wright, & Heaphy, 2020; Hughes, Lee, Tian, Newman, & Legood, 2018). To build solid structures of knowledge we must raze the flimsy structures first: it is unethical and uneconomical to do otherwise.

The need for methods checks

Given our field operates from a weak paradigm, the previous and current editors-in-chief, and the method associate editors of the journal teamed up to write this article (for another example, see Shang & Rönkkö, 2022). The previous editorial team introduced a novelty, perhaps being the first management journal to conduct formal and systematic methodological checks undertaken by a team of dedicated method associate editors. These checks, which are continuing with the incoming team, are designed to address common empirical faults and resolve them before article acceptance; of course, intractable faults lead to desk rejection. These checks benefit authors, the journal, and the field by correcting mistakes, reducing the likelihood for corrections post-publication, and ensuring the work meets the standards for a sturdy brick on which to build our scientific knowledge base. All empirical articles submitted to *The Leadership Quarterly* undergo methods checks prior to acceptance. We believe that much of the journal's recent success depends on these checks. Rather than hogging this knowledge, we have codified and documented it for the benefit of our discipline and beyond.

Table 1 summarizes the most common issues, previously encountered in methods reviews at the journal over the last six years. We present each issue in detail alongside one or more solutions that we have shared with submitting authors. These issues are grouped into three major areas:

1. Design and data collection
2. Data analysis
3. Diagnostics, inferences, and reporting.

Note, our goal is neither to provide an exhaustive list of mistakes and solutions in empirical analysis in general, nor to list all the items the method review team ever focused on. Instead, we focus on issues we have most commonly observed. Suggested solutions are offered to benefit researchers working in leadership and across other social sciences. Our discussion of common methodological problems and ways to mitigate them may be of interest to a broader audience of researchers working across such disciplines as Psychology, Business, Management, Experimental Economics, Sociology, Political Science, Operations Research, and Behavioral Science, to name a few.

Common methodological mistakes

Below we expand on each of the mistakes. We go into detail for common issues that seem ill-understood; we are terse with some of the issues that are very technical in nature and point readers to advanced literature that appropriately discusses the issue at hand. We have numbered them for ease of reference with respect to the summary table.

A. Design and data collection

1. Hypothesis-analysis consistency

1.1. Hypotheses do not reflect or cannot guide the empirical section

Common mistake: Hypotheses—or for exploratory work, research questions—determine measurement and testing. Hypotheses are often poorly formulated in that (a) the theoretical concept they refer to misrepresents how these are later measured (i.e., an author theorizes about leader behavior but measures follower subjective evaluation of the behavior), (b) they are not directly testable, and (c) they do not appropriately map onto the empirical model(s) used to test them.

Solution: When formulating hypotheses, authors need to take greater care of the following points: (a) the incorporated concepts must clearly reflect the concepts measured empirically. For instance, when authors plan to measure leadership via follower evaluations, the theory must be developed around these evaluations of leadership (which are not equivalent to behaviors, Banks, Woznyj, & Mansfield, 2021; Fischer et al., 2020) and the hypotheses must also clearly refer to evaluations; (b) in experimental work, the counterfactual treatment must be explicitly mentioned; (c) the direction of the expected effect/relationship needs to be clearly indicated. Moreover, a hypothesis must match the empirical model used to test it. If the empirical model allows for causal inference, the hypothesis can (and should) use words such as "affects," "impacts," or "effect." In contrast, if the empirical model is not causally identified, authors must refrain from using words that imply causality and instead choose formulations such as "is associated" or "relates to;" and (d) a single hypothesis should neither contain multiple explanatory variables, nor multiple dependent variables. The reason is that the hypothesis is otherwise not directly empirically testable and it is not possible to precisely refer back to it when presenting the results. If it makes theoretical sense to bundle hypotheses (e.g., the effect of one explanatory variable on several outcomes), the hypothesis should be explicitly subdivided (e.g., by using a, b, c for the respective outcomes). Points (a), (b), and (d) are relevant too for exploratory work.

Moreover, authors should explicitly state the results of hypothesis testing or the answers to the research questions in their Results section for each tested hypothesis or question by referring to the number of hypotheses or questions in the Introduction.

2. Experimental design

2.1. Confounded experimental manipulations

Common mistake: To identify the causal effect of an experimental treatment, it is key that the treatment is not confounded (i.e., does not co-vary) with other factors. For instance, when aiming to explore how applicant gender affects callback rates, manipulating gender via men's versus women's names on applicant CVs is only appropriate if people do not perceive the respective names to also vary along other dimensions (such as race, nationality, or socio-economic status) that potentially affect the outcome variable (Kasof, 1993; Simonsohn, 2015). Likewise, when manipulating a certain leadership style (e.g., transformational or charismatic leadership) via a speech to investigate its effect on follower behavior, the speech information content, length, or other features must be equivalent. When experimental treatments are confounded, it is not possible to know whether the estimated effect stems from the treatment or rather just from the variation in the confounds.

Solution: Experimental treatments need to be carefully designed so that they do not co-vary with confounds. Sometimes, potential confounds are objectively measurable (e.g., speech length) and can thus be avoided by design. Other times, they rest on participants' perceptual associations with the treatments (e.g., how applicant names are perceived in terms of race). To rule out the existence of perceptual confounds, it is essential to empirically pre-test the experimental treatments or provide other evidence (e.g., national census data demonstrating a uniform race for the control and treatment group names). Otherwise, authors need to collect the respective evidence ex-post, which can be very costly because a confounded treatment would then require an adjusted follow-up study.

Table 1
Common Issues in Empirical Articles and Suggested Solutions.

Category	Common mistake	Suggested solution
A. Design and data collection		
1. Hypothesis-analysis consistency	1.1 Hypotheses do not appropriately reflect the measured constructs, are not directly testable, or do not map the empirical methods used.	Ensure consistency between theoretical and empirical constructs, clearly state counterfactual treatment (for experiments), identify expected directionality and whether an effect or an association is investigated, avoid multiple explanatory and multiple dependent variables within single hypotheses.
2. Experimental Design	2.1 Confounded experimental manipulations.	Carefully design and empirically pre-test the experimental treatments to ensure manipulation of one variable at a time (so as to isolate treatment effects).
	2.2 Use of demand-driven priming studies.	Perform pre-tests to ensure treatments do not trigger asymmetric demand effects and use non-priming techniques to manipulate the construct of interest.
	2.3 Manipulation checks are placed between manipulation and dependent variable.	If manipulation checks are needed, they should be conducted out-of-sample. When manipulation checks have been placed before the dependent variable, running an additional "reduced form" experiment can test whether study results were biased by the manipulation check.
	2.4 Manipulation checks are confused with attention checks or compliance checks.	Clearly differentiate between the different types of checks in design, terminology, and interpretation.
	2.5 Passing a manipulation check and assuming effect goes only via hypothesized channel.	Complement a manipulation check by discriminant manipulation checks showing that the treatment did not also affect plausible alternative channels.
3. Measurement	3.1 Use of double-barreled items in scaClearly state any data omissions or exclusions providing detailsles.	Measure one concept per item.
	3.2 Improperly documented scale adaptations.	Report any scale adaptation, providing all scale items (modified and unmodified).
	3.3 Use of Cronbach's alpha.	Consider alternatives to Cronbach's alpha like McDonald's Omega (Raykov's rho).
4. Qualitative methods	4.1 Ambiguous coding methodology.	Report objective, reproducible analyses (e.g., via topic modeling).
	4.2 Use of purposeful sampling without having counterfactuals.	Select representative or random counterfactual or matched cases.
	4.3 Data omissions and exclusions are not properly reported.	Clearly state any data omissions or exclusions. Justify omissions or exclusions.
	4.4 Leading interview questions and other biased interviewing techniques.	Avoid leading questions, use neutral language, clearly report all measures used in the study to minimize bias.
5. Open science	5.1 Open science procedures and practices are not or are insufficiently followed.	Where possible, preregister both quantitative and qualitative studies, allow only minor deviations from preregistered study script and transparently report them, share data from published articles, clearly distinguish between confirmatory and exploratory analyses, provide code/analysis file(s) from published articles to simplify replication, share experimental materials or details of data collection/mining.
6. Literature reviews	6.1 Usage of vague or unreproducible procedures in literature reviews.	Ensure reproducibility of literature reviews, clearly describe all procedures of literature search and selection to simplify replication. Follow PRISMA guidelines.
B. Data analysis		
7. Data preprocessing and transformation	7.1 Missing data are removed without appropriate justification.	Clearly document missing data, justify any missing data deletion; use appropriate estimator to handle missing data.
	7.2 Inappropriate or unnecessary use of centering (e.g., mean or scale centering).	Only mean-center when appropriate, provide clear justification for any centering procedures used in the study.
	7.3 Use of unnecessary indices and composites (including hierarchical composites).	Avoid unnecessary indices and composites. Model hierarchical decisions appropriately.
	7.4 Non-linear (e.g., logarithmic) transformations are used inappropriately.	Any non-linear transformations should have a theoretical underpinning and clear justification based on the functional form of the relationship between variables.
	7.5 Log-transforming the dependent variable.	Consider estimating a generalized linear model (GLM) using quasi-maximum likelihood estimation (QMLE).
	7.6 Using ratios in statistical models.	In general, avoid using ratios in statistical models; instead use the numerator and control for the denominator.
	7.7 Unjustified difference scores are used as dependent or explanatory variables.	Avoid difference scores. If the use of difference scores is necessary, run the relevant tests.
	7.8 Continuous variables are arbitrarily categorized (e.g., median splits).	Always treat continuous variables as such. If theoretical categories may be of interest, consider running a finite mixture model.
8. Qualitative methods for text analysis	8.1 Data cleaning stage is not or inappropriately documented.	Clearly justify and document data cleaning in qualitative text analysis.
9. OLS is appropriate for bounded outcomes	9.1 Using GLMs with non-linear links and not considering the use of OLS.	Consider using linear regression by OLS as a default option if the research interest is estimating average marginal effects (rather than focusing on prediction).
10. Regression ranking techniques	10.1 Relative weights analysis is used to rank the strength of predictors in a model.	Only interpret unstandardized partial regression coefficients (obtained from causally identified models) for causal effects.
11. Analyzing experimental data	11.1 Omission of balance checks when randomization to treatment may have been compromised.	Balance checks can provide information on whether randomization to treatments was not successful.
	11.2 The coefficients from regression results are misinterpreted when analyzing 2x2 experiments.	Report the following regression results when analyzing 2x2 experimental data: regression of dependent variable on (i) the two treatments; (ii) the two treatments and controls; (iii) the two treatments and their interaction; (iv) the two treatments, their interaction, and controls. Make sure to correctly interpret the respective coefficients, which change meaning across these specifications.
	11.3 Deleting observations from participants who failed a post-treatment manipulation check.	In general, researchers should report and interpret results using data for the whole sample.

Table 1 (*continued*)

Category	Common mistake	Suggested solution
	11.4 Incorrect analysis and interpretation of non-compliance in randomized experiments.	Avoid dropping subjects from the analysis based on compliance status; conduct an "intention-to-treat" (ITT) analysis and/or estimate the Local Average Treatment Effect (LATE) via instrumental variable regression.
12. Endogeneity: Instrumental variables, measurement error, and testing.	12.1 Endogeneity issues are not appropriately addressed in study design and analysis.	Potential endogeneity bias must be carefully considered in the study design and analysis phase (e.g., via experimental or quasi-experimental methods).
	12.2 Use of incorrect empirical tests as evidence that instruments satisfy the relevance condition.	Report the correct *F*-statistic for instrument(s) excluded from the *y* equation.
	12.3 Instruments are not (as if) random or no theoretical rationale is given to justify they are.	Provide a theoretical rationale for why the instruments used can be considered as (as if) random.
	12.4 No sufficient evidence is given for the exclusion restriction.	Provide a theoretical rationale for why the instruments are expected to satisfy the exclusion restriction. If the model is overidentified, additionally provide an empirical test of overidentification (e.g., Hansen-Sargan test).
	12.5 Endogeneity tests (e.g., Hausman test) are applied without considering their limitations.	Provide a statistical test of endogeneity while also carefully considering the limitations of such tests (e.g., test might be underpowered).
	12.6 Manual execution of two-stage least squares model (2SLS).	Avoid running 2SLS manually; use canned procedures in statistics software instead.
	12.7 Interaction terms containing an endogenous variable are not instrumented.	All endogenous terms should be instrumented. Endogenous interactions composed of one endogenous and one exogenous variable should be instrumented by the instrument for the endogenous predictor and the exogenous variable (i.e., the exogenous moderator).
	12.8 Heckman models are estimated without an excluded instrument.	When estimating a Heckman model, avoid just-identification via functional form alone; include a variable that predicts selection but is excluded from the the outcome equation.
	12.9 Measurement error.	Depending on the availability/number of measures of the same latent variable, (a) consider using a Structural Equation Model (SEM) when several measures are available; (b) consider using error-in-variable regression or constrained SEM when one measure of interest with known reliability is available; (c) consider using instrumental variable estimation when no reliability estimate is available.
	12.10 ITCV or other sensitivity analyses are used as a "test" for omitted variable bias.	Sensitivity analyses should be interpreted cautiously. Only appropriate designs and identification strategies conclusively address omitted variable concerns.
13. Mediation analyses	13.1 Standard mediation models are used when mediators are endogenous.	Discuss the sequential ignorability assumption explicitly. If this assumption is not met, alternative procedures (e.g., instrumental variable analysis) should be used to estimate and interpret causal mediation models.
14. CFA/SEM	14.1 Goodness of fit indices such as CFI, TLI, or RMSEA are used to evaluate SEM or CFA model fit.	Report the chi-square test and only determine the appropriateness of the model fit from this test, which is proven valid; for complex models or small samples use a rescaled chi-square test.
	14.2 Testing structurally unidentified second order latent variable models.	Do not compare first and second order latent variable models with 3 first order factors, unless overidentifying restrictions are made.
	14.3 Using chi-square difference test to compare models that do not fit.	The chi-square difference test can only be applied to compare models that are not rejected by the data.
	14.4 Using modification indices to change CFA/SEM models ex-post.	Avoid indices-based ex-post CFA/SEM modifications (especially do not correlate item disturbances). Modification indices can be used to locate potential misspecifications to provide ideas for how to improve the model in future research.
15. Multilevel (panel) data	15.1 Reporting only results from a random effects (RE) estimator.	Interpreting RE causally requires strong assumptions, which should be discussed. In general, prefer fixed effects (FE) or correlated random effects (CRE) for estimating the effect of level-2 predictors when the RE assumption fails. In some cases (e.g., dynamic panel data model, endogenous level-1 predictors), FE and CRE might not be enough to ensure causal interpretability.
16. Meta-analytic regression	16.1 Inadequate use of fixed effects or random effects meta-analytic regression.	Carefully check the assumptions of fixed effects and random effects meta-analytic regressions before applying them.
	16.2 Overfitting, aggregation bias, ill-defined risk modelling.	Have a sufficient number of studies per examined covariate, avoid regression to the mean, check for ecological fallacy.
17. Control variables	17.1 Personality dimensions are partially omitted as regressors.	If theoretical reasoning suggests controlling for a personality factor in a regression analysis, all other personality dimensions should also be included (e.g., big five personality factors).
C. Diagnostics, inferences, and reporting		
18. Diagnostics	18.1 Harman's single-factor test is used to detect common method variance (CMV).	Reduce the risk of CMV in the study design (e.g., collect dependent and independent variables from different sources) or the study analysis stages (e.g., use instrumental-variable estimation).
	18.2 Outliers are ignored or identified without transparent and appropriate procedures.	Carefully document the procedure used for outlier identification, to ensure replicability of the results. In general, report results with and without outliers and/or consider outlier-robust regression methods.
19. Statistical inference	19.1 Testing on multiple outcomes without appropriate corrections.	Mitigate the problem via Bonferroni correction or related procedures.
	19.2 Results are interpreted as "marginally significant."	Avoid labeling results as "marginally significant." Focus on significant results on an a priori alpha level.
	19.3 Absence of statistical significance is interpreted as "no effect."	Statistically insignificant results simply imply that "no conclusion can be drawn" from the data. To test for "no effect," use additional analyses (e.g., equivalence tests, Bayes factors).
	19.4 Unjustified use of normal standard errors.	Depending on the properties of the data, rely on heteroskedasticity-robust standards errors. Cluster-robust standard errors might also be necessary in multilevel models.

(*continued on next page*)

Table 1 (*continued*)

Category	Common mistake	Suggested solution
20. Reporting	20.1 Insufficient details in figures and tables to ensure replicability.	For regression tables, report (at least in the appendix) all higher order terms, interactions, and control variables used. Label variables in Figures and Tables clearly and consistently, avoiding misleading scales in graphs' axes.
	20.2 Inconsistent reporting of *p*-values and other statistics.	Check reported *p*-values and other statistics carefully before submission.
	20.3 Not reporting uncertainty of estimates in tables and figures.	Report complete results that clarify the magnitude of the reported effect, as well as a measure of the uncertainty of the estimate.
	20.4 Incorrect use of one-tailed tests (especially if estimator is underpowered)	In general report two-tailed tests.
	20.5 Extrapolating fitted predictions beyond data range.	Check that all predictions are reported within the range of the analyzed data and only depict this range in your figures.
	20.6 Reporting standardized regression coefficients.	Report unstandardized regression coefficients or correct for measurement error in *y* before standardizing.
	20.7 Reporting only ANOVA tables.	Report only full regression tables, which provide additional and clearer information compared to ANOVA ones.

2.2. Demand effects in priming studies

Common mistake: Researchers sometimes intend to manipulate a construct via priming techniques. A prominent example is the intent to manipulate power by asking participants to recall and write about an instance in which they had power over others (high power prime) or others had power over them (low power prime), sometimes supplemented by a neutral condition (Duguid & Goncalo, 2015; Galinsky, Gruenfeld, & Magee, 2003; Schaerer, du Plessis, Yap, & Thau, 2018). Priming studies with an obvious manipulation such as the above one are problematic because they are prone to experimenter demand effects (i.e., participants may adapt their behavior to what they think is expected by the experimenter; Orne, 2009; Zizzo, 2010). The reason is that primes can make obvious the study purpose to participants and priming studies often use non-consequential outcomes so that it is "cheap" for participants to act in accordance with what they think is expected from them (Khademi, Mast, Zehnder, & De Saint Priest, 2021; Lonati, Quiroga, Zehnder, & Antonakis, 2018; Sturm & Antonakis, 2015). Most perniciously, such demand effects will often be asymmetric in that the primes provide participants with behavioral cues that go in opposite directions (e.g., behaving more selfishly in the high-power condition and less selfishly in the low power condition). In the case of asymmetric demand effects, it is impossible to determine whether a prime's potential effect on the outcome is really due to the concept the researcher tried to manipulate (e.g., feelings of power) or rather just an artifact of the asymmetric demand effect.

Solution: When priming techniques are used, researchers need to carefully design the experimental conditions such that they are subtle and do not trigger asymmetric demand effects. Pre-tests can provide empirical evidence with regard to participants neither being differently likely to guess the study purpose across conditions nor imputing differential researchers' expectations regarding their behavior in the experiment. Moreover, when appropriate to the study context, incentivized outcome measures can decrease the risk of demand effects. Instead of using obvious primes, use objective manipulations (e.g., manipulated actual power in terms of giving participants the possibility to allocate real resources or not, Khademi et al., 2021).

2.3. Positioning manipulation checks within the experiment

Common mistake: Oftentimes, researchers place manipulation checks between the experimental manipulation and the dependent variable. Doing so may reveal the study purpose to participants and induce them to adjust their behavior to what they think is appropriate or expected by the experimenter (Fayant, Sigall, Lemonnier, Retsin, & Alexopoulos, 2017; Kidd, 1976; Hauser, Ellsworth, & Gonzalez, 2018, Lonati et al., 2018). Khademi et al. (2021) for instance show that asking people to indicate how powerful they feel after a power prime may significantly increase their ability to guess the study hypothesis. Consequently, the researcher loses experimental control.

Solution: If a manipulation is objectively observed or its construct and operationalization are identical (e.g., variation in economic games' parameters), no manipulation check is needed (Lonati et al., 2018). If, in contrast, the manipulation aims at inducing a certain feeling or attitude, a manipulation check may be necessary. In that case, it should preferably be conducted out-of-sample on a sample with similar characteristics (e.g., Hauser et al., 2018). If conducting the manipulation check in another sample is not possible, it may also be conducted after the dependent variable; however, researchers must be aware that participants' experimental behavior may then bias their response to the manipulation check (Lonati et al., 2018). Ultimately, if a manipulation check must be placed before the dependent variable (e.g., when testing a mediator), a second "reduced form" experiment without the manipulation check is needed to confirm that the manipulation check did not bias the results or directly affect the outcome.

2.4. Manipulation versus attention or compliance checks

Common mistake: Authors sometimes confuse "manipulation checks" with "attention checks" (Ejelöv & Luke, 2020). Manipulation checks test whether the manipulation had the intended effect, such as whether training on ethical behavior leads to more ethical behavior. Attention checks, in contrast, test whether participants paid attention while going through the study (e.g., via reverse scaling, response time, or directed queries; Abbey & Meloy, 2017) to assure high data quality. Compliance checks merely indicate if participants did what they were meant to (e.g., write a particular essay; watch a video instead of going for a coffee); however, doing so does not mean the manipulation had the intended effect (see Sturm & Antonakis, 2015).

Solution: Depending on their study, authors may want to use attention, manipulation, or compliance checks. In any case, they must ensure to design, name, and interpret the respective check correctly.

2.5. Discriminant manipulation checks

Common mistake: Manipulation checks are conducted to test whether the manipulation had the intended effect on a specific variable of interest, which is, in turn, hypothesized to affect the dependent variable. Passing such a manipulation check, however, does not rule out the possibility that an effect of the manipulation also worked through alternative channels.

Solution: When it is plausible that the manipulation may have affected the dependent variable (also) via other channels than the construct the researcher wanted to manipulate, it is important to explicitly measure those "discriminant variables" or "discriminant manipulation checks" (e.g., Ejelöv & Luke, 2020, p. 5) and show they were not affected by the manipulation—ideally again out-of-sample.

3. Measurement

3.1. Double-barreled items

Common mistake: The use of "double-barreled" items in scales—that is, an item with two conditions or statements (Furr, 2011; Olson, 2008; Spector, 1992)—can be problematic. Such items are hard to answer because it is not clear how participants should answer when they satisfy only part of the item (e.g., one of the conditions or statements). For example, the following item is not "single-barreled": "My leader usually compliments me or rewards me for work well done especially in tight-deadline situations." Such item is actually triple-barreled because three dimensions require rating: Complimenting, rewarding, and the particular situation. Interpreting answers to double- or triple-barreled items is not straightforward or unambiguous. The same problem applies to items requiring raters to calculate differences or paradoxes in their head (see Myth 3 in Edwards, 2001).

Solution: Double (or triple)-barreled items must be avoided. Rather than relying blindly on published scales, authors should ensure that the different items are unidimensional and that participants can respond unambiguously to all items. If such an item has nonetheless been used for testing purposes, authors must discuss this limitation and make future research suggestions to refine items.

3.2. Scale adaptations

Common mistake: Psychometrically validated scales require a body of evidence that cumulatively allows for the evaluation of the validity and reliability of a scale. Scale adaptations are a common practice in the social sciences (Cortina et al., 2020; Heggestad et al., 2019). However, oftentimes modifications or changes to a scale are not clearly detailed. That is, scale adaptations themselves are not bad and sometimes are even required. The common empirical mistake is that authors often do not provide detail of scale adaptations or supporting validity evidence. Such changes may include adding or removing items from the validated scale, changing the referent, changing the Likert scale, changing the wording of response items, or changing the wording of specific scale items. A survey of psychometricians indicated that some changes are more severe than others (e.g., changing the Likert scale from 5 points to 7 versus dropping several entire items; Heggestad et al., 2019). Regardless, all changes to an existing scale can have implications for research (cf. Bono & McNamara, 2011).

Solution: There are several simple solutions to this common empirical mistake. First, report all scale adaptations regardless of the degree of the change. Clearly, the more one makes changes, the more information needs to be provided in a Methods section and the more supporting validity and reliability information will be required (for recommendations from psychometricians, see Heggestad et al., 2019). Second, authors should fully report scale items in an online appendix, ensuring to follow any relevant copyright and fair use law unless the scales are proprietary in nature. Using proprietary scales does not preclude a thorough description of how the scale items were adapted. Third, it is often necessary to report supporting validity evidence. For instance, one of the most common scale adaptation practices is to drop items in order to shorten a scale. Pareto optimization methods exist which allow authors to consider not only the reliability of the scales, but also convergent and discriminant validity (for a Shiny app, see Cortina et al., 2020).

3.3. Cronbach's alpha

Common mistake: Many authors use Cronbach's alpha as an "internal consistency" coefficient to estimate the reliability of test scores. However, alpha is based on strong assumptions and can provide misleading estimates of reliability (Osburn, 2000; Sijtsma, 2009). Furthermore, alpha does not indicate internal consistency or unidimensionality, and the popular cut-off values of 0.7 or 0.8 are completely arbitrary (Cho & Kim, 2015).

Solution: Alpha should not be an automatic choice for reliability estimation (Cho & Kim, 2015; Trizano-Hermosilla & Alvarado, 2016). Instead, omega and confidence intervals for omega are superior to alpha in several ways: omega makes fewer and more realistic assumptions (Cortina et al., 2020; McNeish, 2018). Dunn, Baguley, and Brunsden (2014) contains an excellent guide for computing omega and interval estimates for omega with supporting code in the open-source software R.

4. Qualitative methods

4.1. Objective coding of data

Common mistake: When coding qualitative data, it is critical that authors use an objective coding method that could be independently replicated by other researchers. We have frequently encountered qualitative studies that provided too little information to help readers understand and critically assess what the authors have done. For instance, the coding procedure or categories were not clearly described, or it was unclear whether the codes were inductively or deductively generated. Also, sometimes the coders were not blind to the study hypotheses, making them vulnerable to confirmation bias. Consequently, qualitative approaches that do not use an objective coding method that can be reproduced or that provide selective quotations to support a particular narrative or position should not be submitted to the journal.

Solution: Authors need to report some transparent and replicable coding method of data (Aguinis & Solarino, 2019). Ideally, coders should be blinded, that is, unaware of the purpose of the study hypothesis; moreover, distinguishing characteristics (e.g., leader names) should be redacted if they may affect coding in stereotypical ways. Authors must also report reliability statistics such as concordance or kappa statistics so that readers can understand how much independent coders agreed beyond chance agreement (Antonakis et al., 2019; Wright, 2017). An ideal way to objectively code data is to use topic modeling (for a tutorial and analysis code in an open-source software see Banks, Woznyj, Wesslen, & Ross, 2018; see also Doldor, Wyatt, & Silvester, 2019; Kobayashi, Mol, Berkers, Kismihók, & Den Hartog, 2018; Oswald, Behrend, Putka, & Sinar, 2020); using deep neural networks in a supervised or unsupervised manner can also prove to be useful (LeCun, Bengio, & Hinton, 2015; Sarker, 2021). Other techniques may be possible, but authors must achieve the same goals in terms of demonstrating analytic reproducibility. Traditional qualitative approaches, such as Qualitative Content Analysis (QCA) can easily be combined with topic modeling (e.g., see Stock, Banks, Voss, Woznjy, & Tonidandel, 2022). In other words, a triangulation approach using multiple analysis techniques can increase confidence in the robustness of the findings.

4.2. Counterfactuals in sampling

Common mistake: Case study research tends to rely on selective and purposeful sampling. In such cases, ensuring counterfactual evidence becomes problematic (Geddes, 2003; McDermott, 2023). Oftentimes there is a misalignment between the data at hand and the inferences made or the research questions. That is, authors often make statements that causal inferences are not the aim, design a qualitative study that does not allow for causal inferences, and then implicitly make causal inferences by "generalizing to a theory" in the write-up of the Results and Discussion section. Moreover, process models (Cloutier & Langley, 2020; Langley, 1999) or studies identifying chains of events and mechanisms (e.g., relations between two concepts) also make causal and temporal claims; however, these claims are not tested in some probabilistic framework (Fairhurst & Antonakis, 2012). Theory of this sort can misguide later empirical research, which stacks up the cards in a way to find significant results to "prove" the theory (Alvesson, 2020).

Solution: One must have contrasting cases, either selected with great care to be representative (akin to what is accomplished via propensity score matching samples) or in another appropriate manner (e.g., randomly if studying many cases). If it is hard to obtain contrasting cases, authors should perform a thought experiment arguing why their results are believable from a counterfactual point of view (Gerring, 2007; Gerring & McDermott, 2007). The idea of such a thought experiment is to theoretically consider whether sampling units not having the characteristics studied in the current sample would lead to a similar or different result. Hypothetical counterfactuals can also be useful if they are selected wisely.

4.3. Data omissions and exclusions

Common mistake: After collecting qualitative data (e.g., data from qualitative interviews, focus groups, member checks, etc.) based on theoretical hypotheses, the authors often need to make reporting decisions (Symon & Cassell, 1998). This process involves identifying answers to the underlying research question, determining repeated words and phrases, establishing unique patterns (outliers) and so forth. In describing this process, the authors may forget to describe the criteria they used for any data omissions and exclusions, or otherwise ignore to clarify the methods and processes by which they settled on a particular final data subset. This approach significantly complicates the review process, because it is often not clear which omission and exclusion criteria were used; failing to report such details can cause serious replication issues.

Solution: It is critical to be transparent about any data omission and exclusion criteria in qualitative (and quantitative) datasets. To that end, it is necessary to clearly document any omissions and exclusions using best practices in qualitative evaluation and research (e.g., Gerring, 2007, 2012; Patton, 1990), which should be presented in the Supplementary Materials accompanying the paper. It is also advisable to identify and apply quality performance metrics and report these metrics in the paper. The exact metrics will depend on the type of qualitative analysis applied. For example, Cuadros-Rodríguez, Pérez-Castaño, and Ruiz-Samblás (2016) provided an extensive list of possible metrics, which could be applied to qualitative studies with multivariate classifications including sensitivity, specificity, false positive and false negative rates, efficiency, and so forth.

4.4. Biased interviewing techniques

Common mistake: Much in the qualitative analysis relies on the way in which the data are collected. Preventing bias is key to getting appropriate data through qualitative interviews. To that effect, leading questions should be avoided, and neutral language should be used (e.g., Roulston & Shelton, 2015; Onwuegbuzie & Leech, 2007). Unfortunately, in some cases, authors may fail to provide their complete interview guide or interview script. At other times, authors might lead respondents and intervene in interpreting what is being said, under the guise that some concepts require the researcher to conceptually expand what is being said. These approaches complicate peer-review as well as make replication and objective coding difficult. At a basic level obtaining bias-free data from respondents who must recall events can be very difficult (Bernard, Killworth, Kronenfeld, & Sailer, 1984) including the problem of putting an intuitive narrative spin on the data (Hastie & Dawes, 2001).

Solution: First, qualitative researchers can preregister their study plan. A study preregistration is simply a plan. All research whether deductive or inductive, quantitative or qualitative, benefits from initial planning. Such a preregistration can also serve to document or acknowledge potential biases or a decision tree of how decisions might be made as a study unfolds iteratively. Study preregistration could be used to register initial research questions or early versions of an interview guide with discussion of how or when a research question or other design features may change as the study inductively unfolds for example.

Second, it is important to check interview guides or interview scripts for leading questions to prevent any bias or prejudice (Morse, Barrett, Mayan, Olson, & Spiers, 2002). It is also necessary to be transparent about the questions used and to avoid leading participants or reinterpreting what they say in vivo or in data reporting. Specifically, the entire interview guide should be made available to the reviewers as a Supplementary Material to the paper so that reviewers can make informed decisions about the appropriateness of the analysis for the research questions asked, as well as allow the reviewers to better understand how the authors ensured that the risk of bias is minimal (e.g., Galdas, 2017). Finally, to avoid retroactive biases and narrative spins, triangulating findings, especially with objectively-coded archival data, would be ideal (Gerring, 2007, 2012).

5. Open science

5.1. Open science principles conformity

Common mistake: Some scientific norms, such as secrecy, particularism, emphasis on quantity of output, and self-interestedness plague social and natural science research (Anderson, Martinson, & De Vries, 2007). The consequence is a lack of analytic reproducibility and robustness, failed replications, and a slowing of scientific advancement. The counter perspective is open science, which represents a values system that seeks to present openness (i.e., collaboration, sharing, exchanging of ideas and materials), transparency, replicability, and reproducibility (Castille, Kreamer, Albritton, Banks, & Rogelberg, 2022). There are several specific open science practices that combine to reduce questionable research practices and to accelerate the cumulation of science. The open science practices adopted will depend on the specific research study at hand. Generally speaking, these practices should promote transparency, reproducibility, and replication, which are the foundational elements of scientific research.

Solutions: There are several simple steps researchers can take to promote open science. First, ideally all study plans are preregistered whether quantitative and qualitative studies, deductive and inductive studies, primary or meta-analytic studies. If a study was preregistered, provide a summary of any deviations from the preregistration. Science is “messy” and so deviations are not inherently good or bad. Some deviations may only be minor whereas others may be more significant. Deviations are not uncommon as scholars confront unexpected challenges. In an inductive study, deviations may be planned intentionally. In a deductive study, the authors seek to minimize or completely avoid deviations. The ultimate goal regardless of the study type should be to accurately and transparently depict the research process as it unfolded.

Second, scholars are strongly encouraged to share data from their published journal article and to document any data or variables excluded. It should be noted that this has long been a requirement from the *American Psychological Association* as indicated in their publication manual. Exceptions to this practice may include the need to protect human subjects or the case of proprietary data. Sharing data is associated with increased reproducibility of science, reduced questionable research practices, and aids with meta-analyses (Hardwick, et al., 2018). However, despite the exhortations of professional associations and journals, data are, in practice, shared rather infrequently (Houtkoop et al., 2018). There are a number of ways for authors to anonymize their data (for a tutorial video see: https://www.youtube.com/watch?v=Sx7TvLGFQLY). Regardless of whether authors share their data, authors should provide a full correlation matrix of all of their variables, including interaction terms needed to reproduce their analyses. Segments of qualitative data shared should also be anonymized. Moreover, at this journal, at any stage of the review process, including post publication, authors must agree to share the data with the editors in a speedy fashion if the data are requested for verification.

Third, authors should clearly distinguish between confirmatory and exploratory analyses (ideally using different subsections). In this case,

there should be alignment between hypotheses and research questions with subsequent analyses.

Fourth, authors should provide analysis code (e.g., R, Stata, Python) along with detailed annotations needed to reproduce their analyses. A good best practice is to ask another member of the author team or a colleague to reproduce the results using the code. This helps to ensure that there are no mistakes and that the conclusions can be reproduced by another scholar with reasonable expertise. R markdown or Stata files are useful for this purpose.

Fifth, transparency checklists serve as reminders to authors about common decisions that need to be transparently disclosed (Aczel et al., 2020). Such checklists also help to promote reproducibility and replication. There are a number of checklists available, such as the 12-item and 36-item checklists (see the shiny app by Aczel et al., 2020). This app produces a PDF output that can be easily uploaded and added to online appendices. Sixth, to truly promote openness, scholars can share or archive study materials to benefit the scientific community and its stakeholders. Examples of commonly shared materials include scale items, experimental materials, research assistant training guides, videos, as well as interview and focus group guides.

Finally, authors should consider open authorship conversations throughout their projects, but especially at the end. Prior to the final submission of a manuscript, author teams should revisit the determination about who was included as an author, who was acknowledged as well as authorship order (and ensure proper citations to past work). This process ensures fairness and open dialogue, especially for junior collaborators around what can be a sensitive and sometimes vague concept. To help avoid issues related to assigning intellectual credit, authors should follow CRediT (https://credit.niso.org/), which encourages reporting specific intellectual contributions on projects.

6. Literature reviews

6.1. Procedures and replicability of literature reviews

Common mistake: Literature reviews, just as any other methodological approach, should be systematic, representative, and reproducible. Sometimes authors do not provide sufficient information to allow others to reproduce their findings. Also, bias may be introduced in the search process and, without sufficient detail, it may be difficult for reviewers and future readers to fully evaluate the conclusions.

Solution: There are several design steps that should be reported during the execution of a literature review. There are also a number of guidelines in existence such as the Preferred Reporting Items for Systematic Reviews and meta-analyses (https://prisma-statement.org/) or the meta-analytic Reporting Standards (https://wmich.edu/sites/default/files/attachments/u58/2015/MARS.pdf). Foundational information includes the exact search terms used (use quotes to clearly indicate the exact words), what databases or sources were leveraged, dates that were searched (month and year) and what were the inclusion/exclusion criteria. Authors should consider creating a flowchart to document the various steps in the search which could be included within a manuscript or in an online appendix. Authors should also explicitly note steps taken to reduce availability or publication bias in both meta-analyses and other types of systematic reviews (Kepes, Banks, McDaniel, & Whetzel, 2012; Kepes, McDaniel, Brannick, & Banks, 2013). Such steps include search strategies as well as modeling strategies for sensitivity analyses. The goal is (a) to ensure reproducibility of the search and extraction of the data, and (b) to demonstrate whether or not certain decision points influenced the conclusions being drawn. If a random sample is taken, such steps need to be explained (for an example see Banks, Fischer, Gooty, & Stock, 2021; see also Fischer et al., 2017). For an easy-to-follow step-by-step guide to systematic literature reviews, see Denyer and Tranfield (2009) as well as Tranfield, Denyer, and Smart (2003).

Finally, the literature reviewed must be appropriately reported and critiqued. For instance, much of the empirical research published does not lend itself to making causal claims; yet, the research is reported as if it did. Authors must very carefully refer to evidence for what it is and not pass off correlational evidence as causal.

B. Data analysis

7. Data preprocessing and transformation

7.1. Missing data

Common mistake: In quantitative studies, authors often remove missing values in a listwise fashion without explaining why. Even if listwise deletion is the default in most statistical software, authors have—at least implicitly—still chosen listwise deletion as their strategy to deal with missing data and should be, thus, able to defend this choice (Newman, 2014). Yet, authors are seldom aware that listwise deletion can often be a sub-par strategy, both in terms of efficiency and consistency of the estimates (King, Honacker, Joseph, & Scheve, 2001; Schafer & Graham, 2002; van der Heijden, Donders, Stijnen, & Moons, 2006).

Solution: It is important to describe the rate and pattern of missingness, as well as the mechanism behind the missing values. Deletion methods may be convenient when the rate of missingness is low (Lan, Xu, Ma, & Li, 2020). However, when the missing rate increases (e.g., 10% or higher), maximum likelihood for missing data or multiple imputation are superior to listwise deletion; the latter often leads to biased estimates and incorrect standard errors (Schafer & Graham, 2002; Sterne et al., 2009). Canned routines to conduct such corrective procedures exist. For instance, Stata contains an extensive suite of multiple imputation commands that can be invoked by using the '*mi estimate:*' prefix (for an excellent documentation of its use, see StataCorp, 2021, Section MI). For complex hierarchical models, it may be advantageous to consider an imputation method that does not require the user to specify an imputation model (Cubillos, Wulff, & Wøhlk, 2022).

7.2. Data centering

Common mistake: It is commonly believed that mean (or scale) centering, or standardization eliminates or reduces collinearity. However, it does nothing to what is called "essential collinearity" in statistics, nor does it change outcomes of the full regression model, the significance of the interaction, or the marginal effects (Dalal & Zickar, 2012; Echambadi & Hess, 2007; Kromrey & Foster-Johnson, 1998). Researchers also center unnecessarily in multilevel models; in particular, grand-mean centering is not useful (see Antonakis, Bastardoz, & Rönkkö, 2021).

Solution: Researchers may choose to mean-center for interpretative purposes if the variables in question do not have meaningful center points (Dalal & Zickar, 2012). Centering is also recommended when researchers estimate moderated structural equation models with latent variables (Marsh, Wen, & Hau, 2004); particular care should be taken with centering decisions in multilevel models (Antonakis et al., 2021).

7.3. Indices or composites

Common mistake: Authors often use indices or other composites to measure their constructs of interest. Composites are built by bundling several constructs into one single measure, typically by adding them up using some weighting scheme. Problematically, such measures pretend that the different underlying constructs are easily combined and assigned the same value to the composite even when those constructs show very different patterns (e.g., a two-dimensional composite that equally weighs both dimensions would give the same value when the first dimension scores high and the second scores low as when those scores were switched). Moreover, authors sometimes even build "hierarchical" composites by combining

two subsequent decisions/behaviors into one single measure. For instance, they measure whether participants select into a certain group ($x1 = 1$ if yes, $x1 = 0$ otherwise) and then whether those in one group showed a specific behavior or had a specific outcome ($x2 = 1$ if yes, $x2 = 0$ otherwise). Those two distinct behaviors are then combined into a single composite c, taking a value of 1 if both $x1 = 1$ and $x2 = 1$ and a value of 0 otherwise). Using either type of composite is methodologically very problematic because each construct is caused by distinct processes and the resulting combined measure is thus "conceptually ambiguous" (Edwards, 2011, p. 373).

Solution: Measuring constructs via unnecessary indices or composites should be avoided altogether. Instead, constructs should be captured via separate unambiguous measures, which should then also be separately entered into the empirical analysis. Regarding hierarchical composites, instead of building composite scores, authors need to estimate an appropriate model (e.g., bivariate probit, Wooldridge, 2010, Chapter 15).

7.4. Non-linear transformations

Common mistake: Researchers often log-transform non-negative and skewed variables to make them "more normally distributed." It is a statistical myth that variables should be log-transformed to make their distributions less skewed: what matters is correctly modeling the functional form (Rönkkö, Aalto, Tenhunen, & Aguirre-Urreta, 2022; Villadsen & Wulff, 2021b). When authors log-transform their dependent variable, they change its relationship to the covariates to comply with criteria that are largely irrelevant (non-normality of the error term is only a problem in small samples; see Wooldridge, 2002, Chapter 5). If the authors state their hypotheses linearly, but estimate a non-linear model, the hypotheses and model become misaligned (for a similar discussion, see Section 1.1 above).

Solution: The use of exponential models should be based on substantive theoretical consideration about a non-linear relationship between the outcome and predictors (Rönkkö et al., 2022; Villadsen & Wulff, 2021b). For instance, Tian, Jiang, and Yang (2022) motivate a non-linear relation by proposing that as a CEO's childhood trauma exposure increases, strategic risk taking will decrease, but at a diminishing rate. Note also how an exponential hypothesis is more precise because it describes not just the direction but also the form of the relationship (Edwards & Berry, 2010). Being specific about the functional form exposes a theory to a more genuine risk of falsification compared to a simple directional hypothesis (Aguinis & Edwards, 2014). In the case of Tian et al. (2022), even if strategic risk taking decreases, the hypothesis would still be rejected if it happens at a constant or increasing rate.

7.5. Log-transforming the dependent variable

Common mistake: Log transforming the outcome is problematic for three reasons. First, ordinary least squares (OLS) is generally inconsistent if used to estimate parameters in a linear regression with a log-transformed outcome (Santos Silva & Tenreyro, 2006; Rönkkö et al., 2022). Second, if the outcome contains zeros, adding an arbitrary constant (e.g., 1) before log-transformation becomes necessary, but biases the estimated parameters (Johnson & Rausser, 1971), goodness-of-fit measures (Ekwaru & Veugelers, 2018), and p-values (Feng et al., 2014). Third, making predictions and inferences about the original untransformed scale of the outcome is not possible without strong assumptions (Wooldridge, 2002, Chapter 6).

Solution: All the issues can be addressed by estimating a generalized linear model (GLM) using quasi-maximum likelihood estimation (QMLE). A GLM with a Poisson distribution and a log-link estimated by QMLE is consistent no matter the true distribution of the data around the conditional mean (Wooldridge, 2010, Chapter 18). Further, GLMs work well for outcomes with many zeroes (Blackburn, 2007) and make it straightforward to compute predictions with correct confidence intervals (Villadsen & Wulff, 2021b). For instance, GLMs can be estimated using Stata's *glm* command. Researchers should test for omitted non-linear relationships using Ramsey (1969) RESET test,[1] Pregibon (1980) link test, and the Box-Cox test (Basu & Rathouz, 2005). A test for the most efficient distribution can be performed using a modified Park test (Manning & Mullahy, 2001). Coefficients should be interpreted appropriately as semi-elasticities for log-level and elasticities for log–log models (Villadsen & Wulff, 2021b). Results should be interpreted using plots that include confidence intervals and the observed data in the plot (Rönkkö et al., 2022).

7.6. Ratios in statistical models

Common mistake: Ratios compare two different quantities such as returns and assets. Whereas ratios can be useful as descriptive measures, using them in statistical models can produce spurious relationships between variables. In fact, when ratios are used as dependent or independent variables, the findings risk being partly or completely driven by the dispersion of the ratio's denominator (Wiseman, 2009).

Solution: Authors should not use ratios in statistical models. Instead of ratios, researchers should focus on the numerator of the ratio and simply control for the denominator (Certo, Busenbark, Kalm, & LePine, 2020). The unscaled version should be modeled (e.g., returns) while controlling for the denominator (e.g., assets), instead of the ratio (e.g., return on assets). Note that this issue is not relevant for fractional variables that are naturally bounded between 0 and 1 and indicate the size of a part relative to the whole (e.g., proportion of total sales attributable to innovation; Villadsen & Wulff, 2021a).

7.7. Difference scores

Common mistake: Difference scores are often reported to study variables operationalizing some type of congruence, fit, or distance. However, if difference scores are used either as dependent or independent variables, explicit tests and/or corrective procedures must be reported to justify their use (Edwards, 1995, 2002; Edwards & Parry, 1993).

Solution: To see the problems related to difference score as a dependent variable, one can notice that, for a difference score $\Delta y = y_a - y_b$, the regression model $\Delta y = \alpha_1 + \alpha_2 x$ assumes implicitly both that $y_a = \alpha_1 + \alpha_2 x$ and that $y_b = \alpha_1 + \alpha_2 x$. That is, using difference scores as a dependent variable implicitly constrains the effect of x and of the constant term to be identical for both y_a and y_b. These constraints, while intuitively compelling, might be empirically untenable (for a related issue, see Section 7.3). Thus, authors should estimate multivariate regression equations to model the effect of x on both dependent variables (i.e., one equation for y_a and one equation for y_b). From this model, authors should test whether the effect of x is identical in magnitude, but opposed in sign, in the two y equations prior to using difference scores (see Edwards, 1995).

Considering the case where a difference score, $\Delta x = x_a - x_b$, is used as an independent variable in a model like $y = \beta_0 + \beta_1 \Delta x$ reveals a similar problem. This model can now be written as $y = \beta_0 + \beta_1 (x_a - x_b)$ and, thus, implicitly constrains the effects of x_a and x_b to have an opposite sign, but the same magnitude. Again, scholars should check whether such constraint is tenable by running a model where x_a and x_b are entered as separate predictors, and then check with a Wald test if their coefficients have, indeed, the same magnitude and opposite sign.

[1] Note, the RESET test is by no means a test for omitted variables (i.e., unobserved confounders).

7.8. Categorizing continuous variables

Common mistake: Despite admonitions from the methodological literature (Cohen, 1983; Fitzsimons, 2008; MacCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993), researchers often transform continuous variables into categorical ones such as by using median splits. This strategy leads to information loss and may produce Type I or Type II errors. In fact, categorizing continuous variables in artificial categories likely leads to inconsistent estimates of the treatment effect. This problem can occur in both primary studies as well as meta-analytic reviews. For instance, in meta-analyses, a continuous moderator may be artificially dichotomized to allow for subgroup analyses.

Solution: Authors should not categorize or dummy-code continuous variables (cf. Aiken & West, 1991). If theoretical categories may be of interest, authors can consider running a finite mixture model (McLachlan & Peel, 2004) to identify whether there are different categories that differentially predict outcomes.

8. Qualitative methods for text analysis

8.1. Data cleaning

Common mistake: Data cleaning is a step that is oftentimes not clearly documented in qualitative and quantitative research. A lack of documentation of the data cleaning or preprocessing steps can be linked to a failure of scientific results to reproduce (Hardwicke et al., 2018; O'Boyle et al., 2017). Text analysis is not an exception.

Solution: There are a few important steps that must be taken to prepare text for analysis. These steps should be reported either directly in the manuscript or in a supplemental file placed in an online appendix. All too often these steps are not reported even though they may influence the results. Authors should take extra care to document and report all decisions made, but especially those which may change the conclusions of the results (for a review, see Banks et al., 2018).

When preparing for computer-aided text analysis or various types of machine learning approaches, common decisions include how quality of writing was handled, what was the average length of responses, and what was the sample size (e.g., number of participants such as the leader, the number of sentences analyzed, and/or the total number of words). Key preprocessing steps to document include explaining how invalid responses were handled, how text was cleaned and tokenized (if appropriate), as well as other potential necessary steps, such as the handling of stop words and sparse terms (for topic modeling), stemming/lemmatization, and uni-/bi-/tri -grams (Banks et al., 2018; Hickman, Thapa, Tay, Cao, & Srinivasan, 2022). In addition to details related to preprocessing, an explanation should be provided on how coders or algorithmic models were trained, an estimate of the reliability of the coded data should be reported, and any other steps that were undertaken to reduce bias should be reported (e.g., de-identify demographics that can bias interpretation of the analysis models and example text).

9. OLS is appropriate for bounded outcomes

9.1. OLS versus GLM

Common mistake: When facing bounded outcomes, many authors discard ordinary least squares (OLS) as an inappropriate estimator and rely only on generalized linear models (GLM) with non-linear link functions. For instance, for binary outcomes, authors often start by estimating a logit or probit model. This approach is, of course, not wrong, yet authors are often not aware that such a modeling strategy comes with several challenges that can be avoided when using OLS. First, interpretation of results is harder. Average partial effects are necessary to gauge the magnitude of the estimated effects (Hoetker, 2007) and for ordered and multinomial logit a graphical interpretation is needed (Wulff, 2015). Second, it is notoriously difficult to extend non-linear models to accommodate endogenous variables or panel data. For instance, instrumental variable estimators for non-linear models are not robust to misspecification of the reduced form (Ramalho & Ramalho, 2017) and make it challenging to accommodate discrete endogenous variables (Wooldridge, 2014).

Solution: If the main purpose is to estimate the average marginal effects, then linear regression by OLS often does a very good job at approximating the effects for bounded outcomes (Wooldridge, 2010, section 15.2). The estimated average marginal effect can be read directly from the estimated coefficient on the variable of interest, making the interpretation easy. Furthermore, extending linear regression by OLS to handle endogenous regressors or panel data is straightforward. For instance, 2SLS estimation of a linear probability model for a binary outcome can be applied no matter if the endogenous variable is continuous, discrete or some combination (Angrist, 2001; Wooldridge, 2014). This versatility makes 2SLS a robust natural starting point for instrumental variables applications with bounded outcomes (Angrist & Krueger, 2001). Still, if the goal is to make predictions or the interest lies in estimates of the partial effects for a wide range of covariate values, especially extreme values, then a non-linear model is needed (Wooldridge, 2010, Chapter 15).

10. Regression ranking techniques

10.1. Relative weights analysis

Common mistake: Relative or dominance weights analysis seeks to order predictors for the purpose of ascertaining which predictor is the strongest driver of a dependent variable y by partitioning the R^2. Such knowledge, if valid, has important policy implications because it tells practitioners how much "bang for their buck" investing in a particular variable (i.e., policy level) they get. Several methods have been devised to rank-order regression coefficients, like Dominance Analysis and Relative Weights Analysis. Despite their widespread use in management and applied psychology, these methods are controversial according to the general statistics literature (see Grömping, 2009, 2015; Thomas, Zumbo, Kwan, & Schweitzer, 2014), and they are also not used in econometrics. The problem with these methods is that there are different ways in which R^2 can be decomposed with correlated predictors (Braun, Converse, & Oswald, 2019). Also, what can inform policy is not the "relative weight" that a predictor has, but rather its causal effect (i.e., how much more y one gets if x increases, all other things being equal). This causal effect is unequivocally captured by the unstandardized partial regression coefficient, under the assumption that the model is correctly causally specified. Current relative weights procedures are contentious, deriving more from ad-hoc intuitions rather than analytic proofs; moreover, applied users typically use the procedures to rank-order regressors in terms of their policy—and hence causal—implications (Kleinbauer, Rönkkö, & Antonakis, 2020).

Solution: Authors wishing to test relative causal effects may conduct a Wald test to compare coefficients, assuming measures are on the same scale; note, this comparison is meaningful for policy only if the effects compared are causally identified to begin with. If the variables are not on the same scale, then authors should use some other approach (e.g., standardization, which has its own shortcomings, see Section 20.6) or compare the variables from a cost-benefit viewpoint (Kleinbauer et al., 2020). Applied users wishing to focus on prediction should turn to modern predictive analytics techniques.

11. Analyzing experimental data

11.1. Balance checks

Common mistake: A necessary condition to be able to interpret the effect of an experimental treatment x on an outcome variable y causally is that randomization of participants to experimental groups "worked." Randomization ensures that, on average, participants do not differ significantly across experimental groups so that the esti-

mated effect of x on y is not driven by the effect of unmodeled participant characteristics on y (e.g., older participants behave systematically differently regarding y) or by the interaction of these characteristics with the experimental treatment (e.g., older participants react significantly stronger to the treatment in terms of y). Yet, authors often omit providing empirical evidence in this regard, at least with respect to observables.

Solution: In large samples, checking whether subjects' observable characteristics are similar across conditions (i.e., a balance test) is unlikely to be needed. However, when there is reason to believe that the randomization device was defective or not directly observable (e.g., randomization not implemented by the researcher, but by an external party) or if attrition is a concern, balance checks can be informative (Athey & Imbens, 2017; Mutz, Pemantle, & Pham, 2019). For instance, one can perform a sequence of pairwise t-tests to compare participant observable characteristics across treatments or predict treatment status with participants' characteristics and use an F-test to study whether characteristics jointly predict the treatment status. One could also use finer measures of distance, like normalized difference in means (Imbens & Wooldridge, 2009). Balance tests can be informative because, although randomization guarantees balance of observable and unobservable characteristics in expectations and over many randomized trials, it does not ensure that each single trial is balanced (Deaton & Cartwright, 2018). Especially in a small sample, there might be particularly large differences in the mean observable characteristics of treated and control individuals, and these differences might pose internal validity threats.[2]

11.2. Analyzing 2 × 2 factorial design experiments using regression analysis

Common mistake: When analyzing 2x2 factorial design experiments using regression analysis, authors can easily misinterpret the coefficients of the experimental treatments and their interaction.

Solution: When authors have a 2 × 2 factorial design experiment with two experimental treatments ($t_1 = 0$ or 1; $t_2 = 0$ or 1), they should report four basic specifications for full transparency:

(1) $y = \alpha_0 + \alpha_1 t_1 + \alpha_2 t_2 + e_1$ (regression of y on the two treatments).
(2) $y = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_c C + e_2$ (regression of y on the two treatments and the controls).
(3) $y = \gamma_0 + \gamma_1 t_1 + \gamma_2 t_2 + \gamma_3 t_1 t_2 + e_3$ (regression of y on the two treatments and their interaction).
(4) $y = \delta_0 + \delta_1 t_1 + \delta_2 t_2 + \delta_3 t_1 t_2 + \delta_c C + e_4$ (regression of y on the two treatments, their interaction, and the controls).

where y is the dependent variable, α, β, γ, and δ are the estimated regression coefficients, C is a vector of controls, and e_1 to e_4 are the respective equations' error terms.

Regarding the interpretation of the coefficients, when no controls are included (models 1 and 3), the models' constants (α_0 and γ_0) denote the baseline value of the dependent variable when participants neither received t_1 ($t_1 = 0$), nor t_2 ($t_2 = 0$). When controls are included (models 2 and 4), the constants (β_0 and δ_0) denote the baseline value of the dependent variable when participants neither received t_1 ($t_1 = 0$), nor t_2 ($t_2 = 0$) and all included covariates C take a value of 0.

In model 1, the coefficients of the experimental treatments are to be interpreted as follows: α_1 is the average effect of receiving t_1 ($t_1 = 1$) compared to not receiving it ($t_1 = 0$), averaged across the two conditions of t_2. In other words, this coefficient compares two experimental groups (both groups that received t_1) with two other experimental groups (both groups that did not receive t_1). Likewise, α_2 constitutes the average effect of receiving t_2 ($t_2 = 1$) compared to not receiving it ($t_2 = 0$), averaged across the two conditions of t_1. Thus, Model 1 gives the average marginal effects, $\frac{\partial_y}{\partial_{t_1}}$ and $\frac{\partial_y}{\partial_{t_2}}$, also known as the "main effects" in ANOVA models. In model 2, the interpretation of β_1 and β_2 is identical, just conditional on the included controls.

In model 3, the interpretation of the treatment coefficients changes substantially. When the interaction between the two treatments is included, γ_1 reflects the effect of receiving t_1 ($t_1 = 1$) compared to not receiving it ($t_1 = 0$), but only for the case in which neither group simultaneously obtained t_2 ($t_2 = 0$). Thus, the coefficient now compares one experimental group ($t_1 = 1, t_2 = 0$) with another one ($t_1 = 0, t_2 = 0$). Likewise, γ_2 reflects the effect of receiving t_2 ($t_2 = 1$) compared to not receiving it ($t_2 = 0$)—only for those who did not also receive t_1 ($t_1 = 0$). The coefficient of the interaction term γ_3 denotes how t_1 changes the effect of t_2 and vice versa (it is thus no direct group comparison). Which of the two interpretations authors should follow (mathematically they are identical) depends on the underlying theory.

Finally, when researchers want to also have the direct group comparison of the group that received neither treatment ($t_1 = 0$, $t_2 = 0$) and the group that received both treatments ($t_1 = 1$, $t_2 = 1$), they can easily do so by testing post-estimation whether $\gamma_1 + \gamma_2 + \gamma_3 = 0$. When they are interested in the direct group comparison between those who received only t_1 ($t_1 = 1$, $t_2 = 0$) and those who received only t_2 ($t_1 = 0$, $t_2 = 1$), authors can likewise test post-estimation whether $\gamma_1 = \gamma_2$. In model 4, the interpretation of δ_1, δ_2, and δ_3 are again identical, just conditional on the included controls.

11.3. Deleting observations based on failed manipulation checks

Common mistake: Researchers often delete observations from participants who failed a post-treatment manipulation check.

Solution: Participant observations should not be excluded based on the results of manipulation checks measured after the treatment assignment. This practice can result in biased results and can threaten study validity (Aronow, Baron, & Pinson, 2019; Montgomery, Nyhan, & Torres, 2018). Instead, researchers should report and interpret results based on the whole sample. In contrast, excluding observations based on a failed *pre*-treatment attention/compliance check does not threaten internal validity per se (Aronow et al., 2019). Researchers can report estimated models that exclude observations in supplementary materials to show the robustness of their main findings.

11.4. Non-compliance in randomized experiments

Common mistake: The key assumption in randomized experiments is that random assignment is the only reason why some subjects are allocated to the treatment or to the control group. However, in some situations, experimenters cannot perfectly enforce the random assignment (e.g., field experiments), facing so-called "non-compliance" issues. Non-compliers are individuals who receive the treatment, even though they were originally assigned to the control group, as well as individuals who do not receive the treatment, despite their original assignment to the treatment group. Non-compliance is a potentially serious problem: compliance status is self-selected by the participant and, as a result, non-compliers might be systematically dif-

[2] If treatments are indeed largely balanced with regard to observables, researchers can have some confidence—though they will never be fully sure—that treatments are also balanced regarding unobservables (i.e., variables that are either not measurable or were not measured as part of the study). Hence, causal interpretation is facilitated, assuming the other necessary conditions for causal interpretation hold. If, in contrast, the tests show clear evidence of unbalancedenss, the risk that the treatments are imbalanced also regarding unobservables is large and estimates thus have a larger risk of being biased, which should explicitly be discussed. In their statistical analyses, authors should in any case–not only ex-post in case of failed randomization–report a model including and a model excluding pre-treatment covariates (see section 11.2). If randomization worked as intended, the inclusion of the additional covariates will not change the estimated treatment coefficient, but potentially increase its precision. Still, care must be used when selecting covariates to include, because different types of "bad controls" (e.g., variables affected by the treatment) can cause different types of biases (see, e.g., Cinelli, Forney, & Pearl, 2022; Montgomery et al., 2018).

ferent from compliers. In turn, experimental results might be invalid if these differences correlate with the outcome. Whereas this issue is well-known in medical sciences and economics (e.g., Duflo, Glennerster, & Kremer, 2007; Gupta, 2011; Athey & Imbens, 2017), it is rarely touched upon in management and psychology (but see Lonati et al., 2018; Sagarin et al., 2014; for an example see Liang et al., 2018). As a result, authors sometimes commit two types of mistakes: (a) dropping non-compliers from the analysis, which leads to biased estimates (see, e.g., Montgomery et al., 2018); (b) correctly analyzing the data including both compliers and non-compliers, but without providing a correct interpretation of the estimated effect.

Solution: The simplest solution to the non-compliance issue is to engage in a so-called "intention-to-treat" (ITT) analysis (Duflo et al., 2007; Athey & Imbens, 2017). Estimating the ITT implies analyzing data from all participants (i.e., both the ones who complied and the ones who did not comply), calculating the difference in the average outcomes based on the original assignment to treatment and control conditions, regardless of whether any given participant has actually received the treatment or the control condition. ITT delivers a causal estimate, yet does not estimate the effect of receiving the treatment per se, but the effect of being in the treatment (compared to the control) group. A second solution is estimating the so-called "Local Average Treatment Effect" (LATE), that is, the average treatment effect on the compliers' sub-population. Again, this quantity has a causal interpretation (at least under some specific assumptions, see Angrist & Pischke, 2009, Chapter 4), but it should not be confused with the average treatment effect. LATE can be estimated with an instrumental variable model where treatment status is instrumented by the random assignment (for more details, additional approaches, and limitations, see, e.g., Sagarin et al., 2014; Athey & Imbens, 2017).

12. Endogeneity: Instrumental variables, measurement error, and testing

12.1. Endogeneity

Common mistake: A common mistake authors make in initial submissions is to ignore the endogeneity problem. Articles having endogeneity issues that are clearly unaccounted for and are not discussed extensively are usually desk-rejected at the journal, unless the editor sees a way to address the issue. Endogeneity may also creep into the manuscript in the revision stage, where new data are included or a different analysis method is used.

Endogeneity refers to a series of empirical problems causing the model disturbance e, which captures unobserved causes of the outcome, to correlate with the measured predictor/s (Antonakis et al., 2010). In the presence of endogeneity, the estimated parameter of interest is uninterpretable because the relationship between the modelled independent variable/s and the outcome could be explained by unobserved causes (the so-called "omitted variables"). For instance, if a particular measured leadership style, x, is modeled as a predictor of an outcome like performance, y, it is likely that omitted variables at the leader level (e.g., intelligence, personality, physical appearance), micro level (e.g., department, firm, or industry factors), or macro level (e.g., culture, laws, time) correlate both with the predictor and the outcome. Having knowledge of and measuring these omitted variables is nothing short of a Herculean task. Note, endogeneity can also be caused by measurement error in the independent variable or simultaneity (see, e.g., Wooldridge, 2002, Chapter 9, 15 and 16) as well as omitted selection or endogenous selection to treatment (Heckman, 1979; Maddala, 1983). For instance, in the case of simultaneity, knowledge of current firm performance may affect how the leader behaves; also, better firm performance (not due to the leader's efforts) may allow the leader to adopt a particular leader style; finally, knowledge of performance may also bias how observers rate the leader.

Solution: Potential endogeneity issues must be considered in the design phase and, if applicable, also discussed in the limitations section of a submission. There are several ways to deal with such issues such as experimental and quasi-experimental methods (e.g., difference-in-differences estimation, regression discontinuity designs, instrumental variable estimation, propensity score matching). Accessible introductory reviews for these techniques in our field include Antonakis et al. (2010), Sieweke and Santoni (2020), Sajons (2020) and Bastardoz et al. (2023). For specific empirical examples in leadership research, readers can refer for instance to Yang, Riepe, Moser, Pull, and Terjesen (2019) for an application of difference-in-differences, to Arvate et al. (2018) or Bastardoz, Jacquart, and Antonkis (2022) for regression discontinuity designs, to Lonati (2020) for instrumental variables, and to Vitanova (2021) for propensity score matching.

12.2. Instrument relevance (or strength)

Common mistake: Authors sometimes use statistical tests incorrectly as evidence that their instruments satisfy the relevance condition (i.e., they have a sufficiently strong effect on the possibly endogenous predictor). Some argue that their instruments are relevant only based on a significant bivariate correlation between the instrument and the potentially endogenous instrumented variable. Others report incorrect first-stage F-statistics, such as the first-stage's overall F-statistic (including controls) or the second-stage's F-statistic (Bastardoz et al., 2023).

Solution: To provide evidence that the relevance condition is satisfied, authors need to test whether their excluded instruments' (i.e., the variables used to instrument the endogenous regressor but not included in the main y equation) joint F-statistic from the first-stage regression–controlling for the included control variables–is higher than Stock and Yogo (2005) critical values.

12.3. Instrument (as if) randomness

Common mistake: Valid instruments must be random (or as-if random, at least controlling for covariates), in that they should not correlate with any omitted factors and, as a result, with the dependent variable equation's disturbance term. Often, authors do not justify why they expect their instruments to theoretically satisfy this condition, which is not empirically testable. Moreover, many confound this condition with the exclusion restriction (Bastardoz et al., 2023).

Solution: Authors need to provide a theoretical argumentation explaining why their instruments are expected to be (as if) random. When the instruments are experimentally manipulated and randomization to experimental treatments worked as intended (see Section 11.1), the condition is satisfied by definition (Antonakis et al., 2010; Sajons, 2020).

12.4. The exclusion restriction

Common mistake: Valid instruments must satisfy the exclusion restriction, which says that the instrument, z, must affect the outcome variable, y, only via the instrumented variable x and not directly or via other channels.[3] Yet, this condition is frequently misunderstood in that there should be no significant effect of z on y (i.e., no significant reduced form effect). In fact, if the instrumented variable x truly impacts the outcome variable y and the instrument z is relevant, then there is—mechanically—also a correlation between z and y (at least in the single-instrument case, see Sajons, 2020). Moreover, when authors empirically test the exclusion restriction via overidentification tests, they sometimes report and interpret these tests very deterministically, with little understanding of their limitations (e.g., constant causal effect assumption, assumption that at least one instrument is in fact valid, assumption that as-if randomness is satisfied; see Bastardoz et al., 2023).

[3] As-if randomness and exclusion restrictions are not clearly differentiated in the traditional error-term notation (i.e., they both imply that the instrument is uncorrelated with the y-equation's disturbance), yet they are conceptually distinct (see Angrist, Imbens, & Rubin, 1996).

Solution: If the model is overidentified such that it has more instruments than instrumented variables, authors should test the exclusion restriction via a test of overidentification such as the Hansen-Sargan test (Hansen, 1982; Sargan, 1958) or the χ^2 test of overall model fit for structural equations models (Antonakis et al., 2010; Bollen, 1989; Jöreskog, 1969). Given that tests of overidentification hinge on the assumption that at least one instrument is in fact valid, authors must, as in the just-identified case, additionally provide a strong theoretical argument for why the exclusion restriction should hold (Antonakis et al., 2010). Details on the estimation procedure can be found in Bastardoz et al. (2023) and Sajons (2020).

12.5. Endogeneity tests

Common mistake: Authors often use endogeneity tests (e.g., Hausman) to justify the use of either instrumental variable- (IV) or OLS-estimation and interpret an insignificant test result as evidence that the predictor variable x is in fact not endogenous. These tests have limitations (Wooldridge, 2015) and may be prone to be underpowered because of a small sample size or when instruments are weak (Hahn, Ham, & Moon, 2011; Hausman, Stock, & Yogo, 2005); also, these tests should not be trusted when the IV estimates are not valid (Baum, Schaffer, & Stillman, 2003).

Solution: Authors can and should provide a statistical test of endogeneity such as the Durbin-Wu-Hausman-test (Durbin, 1954; Hausman, 1978; Wu, 1974). In the interpretation of the results, however, caution is warranted. For instance, authors should explicitly consider the possibility that the endogeneity test was underpowered (and more so in the case of weak instruments) or that the IV model was misspecified (Bastardoz et al., 2023). It is important to also theoretically consider whether a predictor is likely endogenous or not. To provide the full picture, manuscripts should in general report both the efficient (e.g., OLS) as well as consistent (IV) estimates.

12.6. Performing 2SLS estimation manually

Common mistake: It seems intuitive to run two-stage least squares (2SLS) models manually by (a) regressing the potentially endogenous (instrumented) variable on the instrument(s), and (b) regressing the dependent variable on the fitted values from this first-stage (possibly including control variables in both stages). However, performing the two steps manually produces wrong standard errors and thus leads to incorrect statistical inference (Bollen, Kirby, Curran, Paxton, & Chen, 2007). Moreover, such a procedure is also prone to other potential mistakes such as omitting second-stage controls from the first stage, not correlating disturbances of the endogenous regressor with the outcome (which is required when using SEM software to specify an instrumental-variable estimator, Antonakis et al., 2010), or misspecifying the model in the presence of more than one endogenous variable.

Solution: Unless appropriate corrections are made to the standard errors (Wooldridge, 2010, Chapter 5), authors should never run 2SLS models in two different steps, but instead use pre-programmed instrumental variable commands in statistics packages like Stata or R or set up the model correctly when using maximum likelihood estimation.

12.7. Instrumental variables with interactions

Common mistake: When specifying models containing interactions between an endogenous predictor and an exogenous moderator, any interaction term that contains even a single endogenous term might also be endogenous and should be instrumented. Failing to do so will result in biased and inconsistent estimates.

Solutions: Solving the endogeneity problem in the interaction between an endogenous predictor and an exogenous moderator is straightforward and requires instrumenting the endogenous interaction by the product of the instrument for the endogenous predictor and the exogenous moderator (Wooldridge, 2010, Chapter 6). That is, for an instrument z, an endogenous regressor m, an exogenous moderator x, and an outcome y, one should estimate the following system of equations via an instrumental variable estimator:

$$(5)\ m = \alpha_0 + \alpha_1 z + \alpha_2 x + \alpha_3 zx + u$$
$$(6)\ mx = \gamma_0 + \gamma_1 z + \gamma_2 x + \gamma_3 zx + v$$
$$(7)\ y = \beta_0 + \beta_1 m + \beta_2 x + \beta_3 mx + \varepsilon$$

where mx and zx represent, respectively, the interactions between endogenous variable and exogenous moderator and of instrument and exogenous moderator. Aside from the usual caveats related to instrument validity and from the additional difficulties related to the multiple endogenous variable/multiple instrument case (Andrews, Stock, & Sun, 2019; Angrist & Pischke, 2009, Chapter 4; Kleibergen & Paap, 2006), this model presents no specific estimation difficulties. However, it is worth stressing that this solution is valid only if the moderator x is exogenous. Authors should also remember that estimating this system of equations requires all instruments (i.e., z and zx) to enter all first stage equations and the exogenous moderator x to appear in all estimated equations, to avoid incurring other problems (see Wooldridge, 2010, Chapter 9).

12.8. Heckman models without excluded instrument

Common mistake: Heckman selection models can be used to solve non-random sample selection issues (Heckman, 1976, 1979). This type of model is usually identified because of an exclusion restriction similar to the one invoked by instrumental variable models (i.e., an exogenous variable in the selection equation that is excluded from the outcome equation). Differently from traditional instrumental variable models, however, Heckman models are technically identified even if the covariates used in the selection and outcome equations are identical (no exclusion restriction, for details, see Wooldridge, 2002, Chapter 17). This identification strategy leverages the non-linearity of the selection equation. However, this identification strategy based on functional form alone has the important drawback of rendering the inverse Mills ratio calculated from the selection equation potentially extremely collinear with the independent variable(s), causing possibly incorrect inferential results and unstable parameter estimates (see also Hamilton & Nickerson, 2003; Wolfolds & Siegel, 2019).

Solution: Selection models should ideally have at least one exogenous variable in the selection equation that is excluded from the outcome equation to avoid achieving identification only via the functional form (Certo, Busenbark, Woo, & Semadeni, 2016). Overall, we advise against using Heckman models without an excluded instrument, at least in general, inasmuch as we advise against the unreflected use of any identification strategy relying on usually untestable functional forms or distributional assumptions (cf. Angrist & Krueger, 2001).

12.9. Measurement error

Common mistake: Measurement error is a common issue that emerges when a measured variable differs from its true value (e.g., questionnaire items do not measure the relevant latent variable perfectly). Measurement error in model predictors is one of the sources of endogeneity bias (Antonakis et al., 2010). However, despite the longstanding focus on measurement in organizational sciences (e.g., Cortina, 1993), at least four common issues remain prevalent when authors use unreliable measures.

First, authors often assume that measurement error is negligible if predictors have a “large enough” reliability (e.g., Cronbach alpha larger than 0.70) and use averaged items’ responses at the factor level—so called “composite scales”—assuming they are perfectly measured (Cortina et al., 2020). Whereas composites usually contain less measurement error than their individual components, they are not perfectly reliable (Bollen & Lennox, 1991) and, thus, are still

endogenous. Second, authors often assume that measurement error can only bias downwards the regression coefficient of the ill-measured variable. However, multiple mismeasured regressors can also engender positive bias, and endogeneity can "spillover" also to perfectly measured covariates that correlate with the ill-measured one (see Wooldridge, 2002, Chapter 9). Third, authors tend not to address measurement error in non-linear models (e.g., quadratic and interaction models), even though the measurement error issue is particularly problematic in such models (Bohrnstedt & Marwell, 1978). Finally, authors (and reviewers in their evaluation) place a premium on reliability estimates and spend little time on validity (Heggestad et al., 2019). This position is very problematic. For instance, when shortening a validated scale, one may select two different subsets of items (subset A and subset B). Both subsets may have almost identical reliability estimates, but the discriminant and convergent validity estimates may differ substantially (Cortina et al., 2020).

Solution: To appropriately model ill-measured predictors, three main procedures are available. If authors have several measures (e.g., items) of the predictor of interest, a widespread approach is to use Structural Equation Models with latent variables, usually estimated with Maximum Likelihood or alternative techniques (see Bollen, 2019). If authors only have one measure, yet know its reliability, they can use error-in-variables regression or constrain the single indicator's variance of the disturbance in Structural Equation Models (Culpepper, 2012). A different approach is to address endogeneity due to measurement error with an instrumental variable estimator (see the pioneering work of Durbin, 1954; Wald, 1940). Solutions to the measurement error problem in non-linear models are more complex, and require alternative estimators (see, e.g., Brandt, Umbach, Kelava, & Bollen, 2020; Klein & Moosbrugger, 2000). In those cases, however, additional care must be taken, because estimators for non-linear models sometimes make strong distributional assumptions, which need to be carefully checked (Lonati, Rönkkö, & Antonakis, 2020). Finally, validity has to do with whether the measure links to other constructs in a manner specified by the theory in a nomological net; as such, evidence for validity is crucial. If corrections for measurement error can be undertaken, a valid (yet noisy) measure is very useful; however, a reliable but invalid measure has little scientific use.

12.10. Sensitivity analyses for omitted variable bias

Common mistake: Sensitivity analyses (see, e.g., Blackwell, 2014; Oster, 2019) can be used to quantify the robustness of an estimated effect against omitted variable bias. Management scholars have started employing such techniques, relying especially on the so-called "impact threshold of a confounding variable" or ITCV (Frank, 2000). ITCV calculates the correlation that an omitted variable would need to have with both a predictor and an outcome of interest to render their estimated relationship not significant (for details and examples in management, see Busenbark, Yoon, Gamache, & Withers, 2022). However, as noted by Cinelli and Hazlett (2020) and Lonati and Wulff (2023), interpreting the results of this technique is more complex than commonly believed, and can lead to incorrect conclusions about the presence of omitted variable bias even in very simple empirical settings and particularly in cases of multiple omitted variables.

Solutions: Given that omitted variables are unobservable, the only definitive solution is to minimize their impact with specific research designs or clear causal identification strategies (e.g., randomization, natural experiments). When these designs are unavailable, ITCV and other sensitivity analyses can provide some useful indications, yet they shall be interpreted with much care, transparency, and subject-matter expertise (see Cinelli & Hazlett, 2020). Thus, ITCV is not a definitive solution against omitted variable bias and should not be used as a "test for omitted variable bias". In sum, we discourage an excessive or unreflected reliance on any rule of thumb based on ITCV or similar techniques at the journal.

13. Mediation analyses

13.1. Standard meditation models

Common mistake: Another issue related to endogeneity is that, when a mediator is not randomized but only observed by the researcher, typical approaches to mediation analysis (e.g., Baron & Kenny, 1986; Preacher & Hayes, 2004) do not generally lead to causally interpretable results. This unfortunate result emerges because causal mediation makes two strong assumptions, sometimes referred to as "sequential ignorability" (Imai, Keele, & Tingley, 2010). First, the independent variable initiating the causal chain (x) should be randomized or as-if random. Second, the observed mediator (m) must be also as-if random, given x and any observed confounder. Whereas the first assumption is relatively easy to meet (e.g., in an experiment where x is randomized; see Sajons, 2020), the second assumption is a heroic one even in experimental research, because a host of unobserved factors are likely to correlate both with the observed m and with the outcome of interest y (e.g., Emsley, Dunn, & White, 2010; Lonati et al., 2018).

Solution: Given the demanding nature of the sequential ignorability assumption, researchers are advised to avoid running and interpreting mediation models causally, at least in general. If a researcher still wants to conduct a mediation analysis, instrumental variable estimation (which can be interpreted as a full-mediation model, see Section 4 of Antonakis et al., 2010; Antonakis, 2016), sensitivity analyses (Imai, Keele, & Yamamoto, 2010), and parallel design experiments (Imai, Tingley, & Yamamoto, 2013; Pirlott & MacKinnon, 2016; for an example in management, see also Appels, 2022) can be potential solutions. In any case, an explicit acknowledgment of the assumptions behind traditional mediation analysis should be the norm rather than the exception, and so should a transparent discussion about the credibility of these assumptions in the empirical setting at hand.

14. Confirmatory factor analysis (CFA)/Structural equation modeling (SEM)

14.1. Examining goodness of fit

Common mistake: Authors frequently use fit indices to judge the fit of SEM or CFA models based on Hu and Bentler (1999). Relying on goodness or badness of fit indices such as CFI, TLI, or RMSEA to determine whether a model is tenable is neither justified analytically, nor supported empirically. Briefly, indices like RMSEA and CFI are highly influenced by specificities of the model. As is evident from how RMSEA is constructed, it has a higher likelihood to favor complex models especially if tested with large sample sizes, and will arbitrarily reject correct models that are simple or accept wrong models that are complex. Comparative type indexes (e.g., CFI, TLI) compare the worse-fitting model (i.e., a straw man) to the current model and it is not clear what the comparison means and what a correct threshold is especially in the context of an index that has no sampling distribution. The methodological literature has demonstrated these issues quite explicitly (e.g., see Antonakis et al., 2010; Chen, Curran, Bollen, Kirby, & Paxton, 2008; Hayduk, 2014; Hayduk, Cummings, Boadu, Pazderka-Robinson, & Boulianne, 2007; Kline, 2015; McIntosh, 2007; Savalei, 2012; Shipley, 2000).

Solution: Authors need to report and rely on the chi-square test of model fit and its associated p-value (Barrett, 2007; Hayduk, 2014; Hayduk et al., 2007; Ropovik, 2015). A failed model (i.e., significant as per the chi-square test) should openly be reported; yet authors should be tentative and nuanced about the conclusions and policy implications stemming from such a failed model and honestly discuss how the model may be tested better in the future (i.e., make recommendations on how to improve the measurement items or the model).

Authors should also state expressly that ill-fitting models may have estimates that could be biased in unknown ways. Finally, if the sample size is small or the model is complex, corrections are available to ensure an appropriate Type I error rate for the chi-square test (Herzog & Boomsma, 2009; Herzog, Boomsma, & Reinecke, 2007; Jackson, Voth, & Frey, 2013; McNeish, 2020; Swain, 1975; Yuan, Tian, & Yanagihara, 2015).

14.2. Testing the factor structure of 2nd order latent variable models

Common mistake: Authors sometimes test a 2nd order latent variable model and compare it to a 1st order latent variable model. Models with three 1st order factors are mathematically equivalent to models with the same number of 1st order factors that are constrained to load onto the same 2nd order factor (models with two 1st order factors having a 2nd order factor are not identified on the structural model). Thus, modeling a higher-order factor can induce a specious structural relation with outcomes (to the extent that the lower-order factors are not tapping the same construct or differentially predict an outcome).

Solution: A higher order model requires at least four 1st order factors to be tested. A 2nd order latent variable model with three 1st order factors is mathematically equivalent to the 1st order model (Credé & Harms, 2015; Rindskopf & Rose, 1988). Authors should never compare a 1st order to a 2nd order latent variable model unless they make appropriate constraints or have at least 4 lower order factors (see Credé & Harms, 2015; Rindskopf & Rose, 1988).

14.3. Comparing models with a chi-square difference test

Common mistake: Researchers frequently report a chi-square difference test to compare the factor structure of different models. However, a chi-square difference test can only be used to compare two non-rejected models. Comparing two nested ill-fitting models (i.e., two models rejected by the data) does not provide evidence for the soundness of the less ill-fiting model (Ropovik, 2015; Yuan & Bentler, 2004).

Solution: Authors should only rely on a chi-square difference test to compare two models not rejected by the data, that is, two non-significant models according to the chi-square test of model fit. If one model is rejected by the data and one model is not rejected by the data, the model not rejected by the data should be tentatively accepted.

14.4. Modification indices

Common mistake: When a model is rejected by the data, authors change the model post-hoc via modification indices. Modifying the estimated model based on the modification indices—especially by correlating error disturbances of items in an ex-post manner—is rarely advisable because it has low out-of-sample validity and capitalizes on chance (MacCallum, Roznowski, & Necowitz, 1992).

Solution: Use modification indices to uncover potential mispecifications and discuss possible solutions for future research. Do not correlate error disturbances of items in an ex-post manner; correlating error disturbances should only be done based on strong theory and modified models should be validated using a new sample. Note, some suggest splitting a sample randomly to explore and then validate the model (Fabrigar, Wegener, MacCallum, & Strahan, 1999); however, this approach will in general not help because any modeling errors fitted in the first sample can be in principle reproduced in the second sample, given that the samples are equivalent on average.

15. Multilevel (panel) data

15.1. Use of fixed or random effects in multilevel panel data

Common mistake: In multilevel models, authors often report only a random-effects (RE) estimator (e.g., estimated with Hierarchical Linear Modeling, or HLM), which makes the strong random effects assumption (Antonakis et al., 2021; McNeish & Kelley, 2019). One reason for the prevalence of RE models lies in the fact that models with level-2 (i.e., cluster- or group-level) predictors are perfectly collinear with the fixed effects (i.e., the fixed-effects capture all unobserved variability due to the grouping or cluster variable) and, thus, cannot be estimated. If the RE assumption is not met and the fixed effects are not modeled, authors possibly confound their results. That is, if an unmeasured variable that is constant over the panel (i.e., the fixed effect) correlates with the panel-varying predictor x and with the dependent variable y, then part or all of the relation between x and y may be due to the fixed effects. Thus, making policy recommendations on x is simply incorrect (Antonakis et al., 2021; Halaby, 2004; McNeish & Kelley, 2019).

Solution: The RE assumption is testable and should be empirically justified using a Hausman or Wald test (Antonakis et al., 2021). If the RE assumption is tenable, authors can report the RE model, which is an efficient estimator. If the RE assumption is rejected by the data, authors should prefer a fixed effects estimator or a correlated random effects (CRE) model. A CRE model is implemented by including the cluster means of all level-1 variables that vary within the level-2 group or cluster. Importantly, a CRE model includes the "best of both worlds" because it is consistent when the RE assumption is not tenable, while allowing for the inclusion of level-2 predictors. The Wald test of the cluster means equal to zero is akin to a Hausman endogeneity test. If this Wald test is not significant, it indicates that one can omit the cluster means. If this Wald test is significant, authors need to keep the cluster means, which is essentially a fixed-effects estimator (Schunck, 2013).

However, including cluster means or using a fixed-effects estimation is not a silver bullet with respect to causal interpretability. Level-1 predictor variables should be exogenous to ensure an appropriate causal interpretability of the findings; else, authors should rely on instrumental-variable panel models (Wooldridge, 2010, Chapter 11). Also, fixed effects estimators are not enough to guarantee causal interpretability with dynamic panel data models (i.e., models containing at least one lagged dependent variable; for details, see, e.g., Li, Ding, Hu, & Wan, 2021).

16. Meta-analytic regression

16.1. Fixed versus random effects meta-analytic regression

Common mistake: Meta-regression is a statistical tool often used as an extension of standard meta-analysis. When there is sufficient unexplained variation in the result of interest across studies, it may be necessary to investigate whether such heterogeneity may be further explained by variation in study variables or variation in study populations. Although many authors correctly use meta-regression analysis to explore heterogeneity, they often fail to consider the assumptions of fixed versus random effect models. As a result, they often use meta-analytic regression inappropriately (e.g., Petitti, 2000).

Solution: It is important to carefully read all assumptions of the fixed effect and random effect models and apply the correct option for meta-analytic regression analysis. Fixed effect models assume no heterogeneity in effects across studies and solely address within-study sampling error (Borenstein, Hedges, Higgins, & Rothstein, 2021). At the same time, random effects models summarize the uncertainty in the pooled estimate generated from between-variance contribution to total variability, resulting in greater standard errors (Hartung, Knapp, & Sinha, 2008). Therefore, if the task is to capture heterogeneity between study-level variables, fixed effects meta-analytic regression is generally inappropriate for this task.

16.2. Other typical meta-analytic regression considerations

Common mistake: Other issues that are typically observed in studies using meta-analytic regressions include (Gonzalez-Mulé & Aguinis, 2018):

i. meta-regression models may suffer from overfitting, especially if the number of studies per covariate is low, resulting in spurious correlations and incorrect results;
ii. meta-analytic regressions may suffer from ecological fallacy (aggregation bias) when average study participant features and pooled treatment effects may not represent individual participant's characteristics and marginal treatment effects;
iii. meta-regression studies regressing treatment effects against the risk of the outcome reported in included experimental studies are difficult to understand and interpret because the observed risk included as a covariate is also contained in the treatment effect used as the dependent variable.

Solutions: Having a sufficient number of studies per examined covariate in a meta-regression analysis helps mitigate overfitting (cf. Higgins et al., 2019). Avoid regression to the mean, which may otherwise result in covariates and treatment effects being correlated when in fact no relation exists. Use additional analyses to check whether ecological fallacy may be present (e.g., logistic regression).

17. Control variables

17.1. Personality variables as regressors

Common mistake: Researchers frequently focus on one or two (but often not all) personality dimensions. However, not controlling for all Big 5 or HEXACO dimensions creates a potential omitted variable bias (see Antonakis & Dietz, 2011). Also, it is the norm to control for all personality dimensions, and not just the target dimension under study, if one wishes to better understand the unique effect of a target personality dimension (Zaccaro, 2012).

Solution: When applicable, authors should include all Big 5 or HEXACO dimensions as controls because personality dimensions tend to be correlated with predictors and outcomes in most leadership models (Judge, Bono, Ilies, & Gerhardt, 2002). Controlling for the multivariate effects of personality is critical to ensure a causal interpretability of the findings (Antonakis et al., 2010; Zaccaro, 2012; Zaccaro, Green, Dubrow, & Kolze, 2018). When authors have not gathered the data on all personality dimensions, one option is to run a Monte Carlo simulation—generating random variables mirroring the meta-analytic correlations between personality dimensions and the outcome of interest—to estimate the effect of including the Big 5 dimensions along with the other dimensions as predictors.

C. Diagnostics, inferences, and reporting

18. Diagnostics

18.1. Harman's test to detect CMV

Common mistake: Harman's single-factor test should not be used as a post-hoc tool to detect common method variance (CMV). Often researchers cite Podsakoff, MacKenzie, Lee, and Podsakoff (2003) when they justify using Harman's test, but Podsakoff et al. (2003) specifically noted that "despite the fact this procedure is widely used, we do not believe it is a useful remedy to deal with the problem and turn our attention to other statistical remedies that we feel are better suited for this purpose" (p. 889). Researchers may be tempted to use the latent factor to model the common method variance; however, there too are some serious reservations about the use of this technique, as the methodological literature has shown (Antonakis et al., 2010; Richardson, Simmering, & Sturman, 2009).

Solution: The best solution to CMV is to handle it in the ex-ante research design stage. For instance, the dependent variable should be collected from a different source than is the independent variables (Chang, van Witteloostuijn, & Eden, 2010), or the data collection should be spread over time (Ostroff, Kinicki, & Clark, 2002)—note, these designs do not eliminate endogeneity issues stemming from other sources in measured data. Post-hoc analyses may be a last resort, but are rarely convincing on their own (Conway & Lance, 2010). The best solution is to use instrumental variables that can take care of common-method variance and other sources of endogeneity (Podsakoff, MacKenzie, & Podsakoff, 2012) as long as one finds appropriate instruments (Bastardoz et al., 2023).

18.2. Outliers

Common mistake: Outliers—usually defined as observations that "deviate markedly from other members of the sample" (p. 1, Grubbs, 1969)—are a common headache for empirical researchers in leadership, management, and social sciences in general (e.g., Aguinis, Gottfredson, & Joo, 2013; Leys, Ley, Klein, Bernard, & Licata, 2013). Dealing with such observations is quite common, yet it can be complex, often entailing some degree of subjectivity. Thus, authors might fail to report the necessary procedures to ensure the robustness of the results and/or lack accurate descriptions of the outlier detection procedures used.

Solutions: Extreme values should be identified, reported, and treated transparently. Descriptive statistics and graphs (e.g., box plots, histograms, scatter plots, or other representations of the distribution of the raw data) are important tools to identify "univariate outliers" and should be examined, if not reported directly in the manuscript. However, leadership scholars should note that multivariate outliers might exist, too (Rousseeuw & Van Zomeren, 1990). Dealing with such cases is more complex, and requires using, for instance, added variable plots (e.g., Kohler & Kreuter, 2005, Chapter 8; Leys et al., 2013). On the contrary, we strongly advise against a blind use of commonly applied thresholds to delete extreme values (e.g., deleting variables that lie, say, 3 standard deviations above or below the mean). These rules of thumb use descriptive statistics that are themselves affected by the presence of extreme values (i.e., mean and standard deviation). Moreover, thresholds based on standard deviations from the mean are based on the implicit assumption of normal variables (which is untenable in many situations, Micceri, 1989). On the contrary, we urge authors to report their results both with and without potentially extreme values as a robustness check (at least in the appendix; for an example, see Lonati, 2020) or to conduct some additional analyses with regression methods that are, in principle, more robust to extreme values (e.g., quantile regression, MM-estimators; for details, see Koenker & Hallock, 2001; Verardi & Croux, 2009).

19. Statistical inference

19.1. Multiple hypothesis testing

Common mistake: It is standard procedure in leadership and management research to collect multiple predictors and outcomes. When researchers test for the effect of a predictor on multiple outcomes in regression analyses, they face an increased risk of committing a Type I error (i.e., rejecting a null hypothesis that is true in the population). Analyzing multiple outcomes increases the probability that a predictor will significantly affect one of the outcomes by chance.

Solution: To keep a Type I error rate at 5%, authors who use multiple testing on more than one outcome variable must correct their alpha value (see List, Shaikh, & Xu, 2019). The most straightforward correction is named after Bonferroni (1936), who suggested dividing the desired alpha level (.05 conventionally) by the number of tests or comparisons performed (Dunn, 1961). For instance, a researcher using 10 independent variables to predict 5 outcomes should use an alpha value of .05/50 = .001 to keep an overall Type I error rate of 5%. Similar techniques for controlling the family-wise error rate like the Holm-Bonferroni method (Holm, 1979) are also recommended (Aickin & Gensler, 1996).

19.2. "Marginally significant"

Common mistake: Researchers frequently present their results as being "marginally" or "almost significant" (for more examples, see Hankins, 2013). Such terminology is used when a *p*-value is not significant, but falls "close" to a specified threshold (e.g., $p = .06$). This practice is common in psychology (Olsson-Collentine, van Assen, & Hartgerink, 2019), but represents a fundamental confusion between the measure for the Type I error rate (α) and the *p*-value (Hubbard, 2004). The "confusion undermines the rigorous Neyman-Pearson interpretation of limiting error to a prespecified level α. And the role of the value of *p* as a quantitative piece of ongoing scientific investigation (including using null hypotheses that are not a hypothesis of zero effect) favored by Fisher is lost to the decision-making encouraged by a statement of significance or lack thereof." (Kennedy-Shaffer, 2019, p. 86).

First, in the Neyman-Pearson framework, the outcome of a statistical test is either to reject or not reject the null hypothesis—there is no category of "almost rejection" for when a *p*-value is slightly greater than a pre-set alpha (Gibbs & Gibbs, 2015). Second, *p*-values "close" to a threshold do not "trend" or "move" in some direction such that they would become significant if more data were available or the experiment was repeated (Wood, Freemantle, King, & Nazareth, 2014). Third, we find it suspicious that this type of language is only being used to downplay non-significant results. Logically, authors would need to be prepared to say just as many times that a *p*-value of .04 "approached statistical non-significance" as they were to say that a *p*-value of .06 "approached statistical significance" (Mansfield, 2005). That is, there is no meaningful difference between $p = .049$ and $p = .051$. However, if one adopts a prespecified cutoff, one should adhere to that.

Solution: When relying on *p*-values for statistical inference, researchers should decide on whether to adopt a Fisherian or Neyman-Pearson statistical framework. Independent of the chosen framework, researchers need to refrain from using phrases such as "marginally significant" because they are misleading under both frameworks.

19.3. Statistical non-significance

Common mistake: Many authors interpret a statistically non-significant result as a "no effect." Yet, a statistically non-significant result does not imply that there is no effect (Lakens, McLatchie, Isager, Scheel, & Dienes, 2020). Indeed, absence of evidence is not evidence of absence. Absence of evidence could simply mean that there is no information in the data about the relationship (Altman & Bland, 1995). A non-significant result could mean that the experiment is statistically underpowered to detect the effect (Harms & Lakens, 2018).

Solution: Informally, a statistically non-significant result means that the data are not informative about whether there is a difference in the population or not, that is, no conclusion can be drawn. Lakens (2022, Chapter 1.7) referred to the concept of 無 "mu": The answer is neither yes nor no. If authors wish to make conclusive statements about null results, they can do so using equivalence testing through the two one-sided tests procedure (Lakens, Scheel, & Isager, 2018), inference by intervals (Dienes, 2014), Bayes factors (Lakens et al., 2020) or Bayesian estimation of the highest density interval (Kruschke, 2011). For an overview, we refer to Stanton (2021) and Harms and Lakens (2018).

19.4. (Cluster) robust standard errors

Common mistake: Aside from the exogeneity assumption typical of the ordinary least squares (OLS) estimator and of similar techniques (e.g., Maximum Likelihood), these estimators make another assumption on the unobservable disturbance term *e*: Its constant variance, conditional on the predictors. Failing to meet this assumption does not affect the bias or the consistency of the OLS estimator (like the more perilous endogeneity) yet can bias the estimation of the standard errors (Wooldridge, 2002, Chapter 8). Specifically, standard errors based on the assumption of constant (conditional) variance are invalid if the variance of the disturbance changes across sub-groups of the population/as a function of the predictor, *x*. A related issue emerges if errors are correlated within clusters of observations (e.g., correlated over time for the same individual, correlated over subordinates of the same leader, or correlated over space; this issue is clearly related to hierarchical data structures, see also Section 15.1). These two issues —and their co-occurrence—are, at times, misunderstood by applied researchers.

Solution: Solving the non-constant variance issue requires using heteroskedasticity-robust standard errors, which are valid in large samples and are easily implemented by most statistical software (e.g., option vce(r) in Stata; MLR estimator in Mplus; sandwich package for R, see Zeileis, Köll, & Graham, 2020). Whereas dealing with heteroskedasticity is in principle easy, failing to do so is unlikely to completely mislead results (Angrist & Pischke, 2009, Chapter 8). Contrarily, failing to address issues related to clustering can lead to vastly underestimated standard errors. Many believe that corrective procedures for this issue only require using some multilevel model. This intuition is correct, but if within-cluster error correlation is present, so-called "cluster robust" standard errors can be more appropriate, at least if the cluster number is large enough (see Antonakis et al., 2021; Cameron, Gelbach, & Miller, 2011; Cameron & Miller, 2015; MacKinnon & Webb, 2019; McNeish, Stapleton, & Silverman, 2017).

20. Reporting

20.1. Figures/Tables

Common mistake: Authors often report results in a way that does not allow other researchers to understand and/or transparently replicate their results. Frequent issues for Figures include (a) variable names that change across Figures but are not clearly indicated; (b) *x*- or *y*-axes that are not labeled or are unclear; (c) scales that are modified (e.g., logarithmic scale when the effect is linear) or that do not represent the full scope of the variable (e.g., to magnify a small effect); and (d) multiple similar Figures having different scales and thus giving the reader a wrong impression. Frequent issues for Tables include (a) sample size not reported for the corresponding analysis; (b) all variables included in a regression not being reported; (c) the type of standard errors not being precisely reported or mentioned; (d) different estimators are reported in different columns but are not clearly described or labeled; (e) the name of the dependent variable not reported. For ease of reading, numbers should be decimal aligned.

Solution: Authors should report all variables used in the estimated regressions, including the higher-order terms and interactions as well as measured control variables (at least in an appendix). Different estimators should be clearly labeled (by column or in the table note). The dependent variable must be explicitly specified, for instance in the table header. Authors should also report all analyses performed in sensitivity analyses (at least in an appendix), and not only those that turned out to be significant. Figures and Tables should be self-standing so that readers could understand them without reading the full manuscript, which implies labeling the axes and different lines reported.

20.2. Reporting p-values and other statistics

Common mistake: Reporting errors are humane but frequent. In most of the cases, these are omission mistakes whereas in some very rare cases, it can be an attempt at purposefully manipulating the reported results.

Solution: To make sure that LQ papers report the fewest errors possible, we run most accepted articles into the website https://www.statcheck.io. This website reports whether there are inconsistencies in the reporting of *p*-values (although it does not yet take into account Bonferroni corrections and cannot check several types of tables/statis-

tics). Authors should also check their manuscripts themselves prior to submitting. We sometimes manually check (using the original data or the values reported in a manuscript) whether the reported statistics and their associated significance are correctly reported.

20.3. Uncertainty of estimates

Common mistake: Authors often report only mean differences between treatments or whether such difference is statistically significant. Authors also omit to report the uncertainty of their estimates in Tables (e.g., by not reporting standard errors) or Figures (e.g., by not reporting the standard errors around their predicted values).

Solution: Along with the exact *p*-value, authors should report the size of the effect and the uncertainty of the estimate, that is, the standard error. These measures of effect and precision are needed to understand the magnitude of the effect. Similarly, authors should report measures of variability (i.e., 95% confidence intervals) around the predicted values in their graphs.

20.4. One-tailed versus two-tailed tests

Common mistake: Authors sometimes report results from one-sided tests of regression coefficients, presumably to increase statistical power. However, performing one-sided tests due to low power is not a valid reason (Baguley, 2012, Chapter 4).

Solution: To remain agnostic about the direction of effects in the data, we encourage authors to report two-sided tests. If authors can theoretically justify a directional effect, then this strategy may be warranted (Lakens, 2022, Chapter 5.7).

20.5. Out-of-scope predictions

Issue: Authors sometimes report values in Figures that are outside of the scope of values present in their dataset. Authors should be particularly careful not to extrapolate beyond their data range because by doing so they may invent relations where there are none. Note that this issue is particularly frequent when authors use the response surface methodology offered by Edwards and colleagues (Edwards, 2002; Edwards & Parry, 1993)—probably because Edwards' methodology effortlessly provides a graph going from −2 to +2 (or −3 to +3) on the fit construct.

Solution: Authors should only make predictions within the range of their data and, if possible, demonstrate the range in plots (Rönkkö et al., 2022). That is, when reporting on a two- or three-dimensional graph the relationship between their predictors and outcomes, authors should examine the range or bound of their predictor(s) and report predictions only within the available range. In the case of Edwards' methodology, authors should simply truncate their data range where they have data points. This issue is quite serious because misrepresentation of the prediction can mislead policy (see Fischer et al., 2021, Figure 4 and discussion on pp. 11–12).

20.6. Standardized versus unstandardized regression coefficients

Common mistake: Authors sometimes report regression coefficients that have been standardized. Standardization entails subtracting the mean of the dependent variable, $\bar{y}$ from the observation value, y, and then dividing an observation by the standard deviation (i.e., $\sqrt{\widehat{var}_y}$); that is, standardized y, $Std_y = \frac{y-\bar{y}}{\sqrt{\widehat{var}_y}}$. Because the variance depends in part on measurement errors, if the outcome is standardized and if there are measurement errors in y, the regression slopes will be biased. Also, when standardizing y, the risk is to estimate seemingly different effects of the same predictor in different samples, simply because the variances of y might differ. However, regression coefficients with an unstandardized y remain unbiased.

Solution: In the general statistics literature, standardization is controversial and there is general agreement that standardization should be avoided (Baguley, 2009; Criqui, 1991; Greenland, Maclure, Schlesselman, Poole, & Morgenstern, 1991; Greenland, Schlesselman, & Criqui, 1986; Kim & Ferree, 1981). If it is essential to report standardized metrics, measurement error should be removed from y (and of course, always removed from the predictors too). Note, standardization of the predictors does not cause any specific issue.

20.7. ANOVA versus regression tables

Common mistake: Authors sometimes submit regression results in an ANOVA table. Results from ANOVA tables are not self-explanatory and add no information compared to complete and understandable regression tables. Contrary to having regression coefficients, it is not immediately obvious from the ANOVA table how each of the regressors affects the outcome (i.e., one would have to generate marginal means). Moreover, (a) for the case of a factor variable having two levels or for the case of a continuous covariate, the ANOVA table reports average marginal effects for the regressor, and (b) for the case of a factor variable having more than two levels, the ANOVA table reports the Wald test for the dummies comprising the factor variable equaling zero. The differences in interpretation between (a) and (b) can create confusion. Of course, marginal predictions from the models are the same and all results from the ANOVA table can be reproduced in the regression framework.

Solution: Given the flexibility of the standard OLS model, we expect manuscripts to only report regression coefficients—and to interpret them as explained in Section 11.2. Average marginal effects for all regressors, as well as Wald tests, can be easily computed via postestimation commands for regression models.

Conclusion

Reliable evidence is a critical element for building theory and evidence-based practice. This evidence can only be accomplished through appropriate research designs, reproducibility, transparency, and correct analysis techniques. Methods checks have been used to accomplish this goal at *The Leadership Quarterly*. In this article, we have documented the most critical issues we have seen.

Our goal with this article is to help future authors better craft their articles for submission to this journal. Also, we hope that other journals will similarly adopt the values of open science to ensure that robust and reliable findings are reported. In this way, management studies will one day soon—and brick by brick—become a true science.

References

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management, 53*, 63–70.

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharsky, S., Benjamin, D., et al. (2020). A consensus-based transparency checklist for social and behavioural researchers. *Nature Human Behaviour, 4*, 4–6.

Aguinis, H., Banks, G. C., Rogelberg, S., & Cascio, W. (2020). Actionable recommendations for narrowing the science-practice gap in open science. *Organizational Behavior and Human Decision Processes, 158*, 27–35.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16*(2), 270–301.

Aguinis, H., & Edwards, J. R. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies, 51*(1), 143–174.

Aguinis, H., & Solarino, A. M. (2019). Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal, 40*(8), 1291–1315.

Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health, 86*(5), 726–728.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage Publications.

Altman, D. G., & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. *BMJ, 311*(7003), 485.

Alvesson, M. (2020). Upbeat leadership: A recipe for – or against – "successful" leadership studies. *The Leadership Quarterly, 31*(6), 101439.

Anderson, M. S., Martinson, B. C., & De Vries, R. (2007). Normative dissonance in science: Results from a national survey of US scientists. *Journal of Empirical Research on Human Research Ethics, 2*(4), 3–14.

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics, 11*(1), 727–753.

Angrist, J. D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business & Economic Statistics, 19*(1), 2–16.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*(434), 444–455.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives, 15*(4), 69–85.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Antonakis, J. (2016). *Testing mediation: The endogeneity problem and the solution*. Master Tutorial, Society for Industrial and Organizational Psychology. Anaheim, U.S.A.

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly, 28*(1), 5–21.

Antonakis, J. (2023). In Support of Slow Science: Robust, Replicable, and Reproducible. *The Leadership Quarterly*.

Antonakis, J., Banks, G. C., Bastardoz, N., Cole, M. S., Day, D. V., Eagly, A. H., et al. (2019). The Leadership Quarterly: state of the journal. *The Leadership Quarterly, 30*(1), 1–9.

Antonakis, J., Bastardoz, N., Jacquart, P., & Shamir, B. (2016). Charisma: An Ill-Defined and Ill-Measured Gift. *Annual Review of Organizational Psychology and Organizational Behavior, 3*(1), 293–319.

Antonakis, J., Bastardoz, N., & Rönkkö, M. (2021). On Ignoring the Random Effects Assumption in Multilevel Models: Review, Critique, and Recommendations. *Organizational Research Methods, 24*(2), 443–483.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*(6), 1086–1120.

Antonakis, J., & Dietz, J. (2011). Looking for validity or testing it? The perils of stepwise regression, extreme-scores analysis, heteroscedasticity, and measurement error. *Personality and Individual Differences, 50*(3), 409–415.

Appels, M. (2022). CEO Sociopolitical Activism as a Signal of Authentic Leadership to Prospective Employees. *Journal of Management*, 01492063221110207. https://doi.org/10.1177/01492063221110207.

Aronow, P. M., Baron, J., & Pinson, L. (2019). A note on dropping experimental subjects who fail a manipulation check. *Political Analysis, 27*(4), 572–589.

Arvate, P., Galilea, G., & Todescat, I. (2018). The Queen Bee: A myth? The effect of top-level female leadership on subordinate females. *The Leadership Quarterly, 29*(5), 533–548.

Athey, S., & Imbens, G. W. (2017). Chapter 3 - The Econometrics of Randomized Experiments. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of Field Experiments* (pp. 73–140). North-Holland.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*(3), 603–617.

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Palgrave Macmillan.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554.

Banks, G. C., Fischer, T., Gooty, J., & Stock, G. (2021). Ethical leadership: Mapping the terrain for concept cleanup and a future research agenda. *The Leadership Quarterly, 32*(2), 101471.

Banks, G. C., O'Boyle, E., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., et al. (2016). Questions About Questionable Research Practices in the Field of Management A Guest Commentary. *Journal of Management, 42*(1), 5–20.

Banks, G. C., Woznyj, H. M., Wesslen, R., & Ross, R. (2018). A review of best practice recommendations for text-analysis in R (and a user friendly app). *Journal of Business and Psychology, 33*, 445–459.

Banks, G. C., Woznyj, H. M., & Mansfield, C. A. (2021). Where is "behavior" in organizational behavior? A call for a revolution in leadership research and beyond. *The Leadership Quarterly, 101581*. https://doi.org/10.1016/j.leaqua.2021.101581.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*(6), 1173–1182.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences, 42*(5), 815–824.

Basu, A., & Rathouz, P. J. (2005). Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics, 6*(1), 93–109.

Bastardoz, N., Jacquart, P., & Antonkis, J. (2022). Effect of crises on charisma signaling: A regression discontinuity design. *The Leadership Quarterly, 101590*. https://doi.org/10.1016/j.leaqua.2021.101590.

Bastardoz, N., Matthews, M. J., Sajons, G. B., Ransom, T., Kelemen, T. K., & Matthews, S. H. (2023). Instrumental Variables Estimation: Assumptions, Pitfalls, and Guidelines. *The Leadership Quarterly., 101673*. https://doi.org/10.1016/j.leaqua.2022.101673.

Baum, C. F., Schaffer, M. E., & Stillman, S. (2003). Instrumental variables and GMM: Estimation and testing. *The Stata Journal, 3*(1), 1–31.

Bechtoldt, M. N., Bannier, C. E., & Rock, B. (2018). The glass cliff myth? – Evidence from Germany and the U.K. *The Leadership Quarterly, 30*(3), 273–297.

Bergh, D. D., Perry, J., & Hanke, R. (2006). Some predictors of SMJ article impact. *Strategic Management Journal, 27*(1), 81–100.

Bernard, H. R., Killworth, P., Kronenfeld, D., & Sailer, L. (1984). The Problem of Informant Accuracy: The Validity of Retrospective Data. *Annual Review of Anthropology, 13*(1), 495–517.

Blackburn, M. L. (2007). Estimating wage differentials without logarithms. *Labour Economics, 14*(1), 73–98.

Blackwell, M. (2014). A selection bias approach to sensitivity analysis for causal effects. *Political Analysis, 22*(2), 169–182.

Bohrnstedt, G. W., & Marwell, G. (1978). The reliability of products of two random variables. *Sociological Methodology, 9*, 254–273.

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

Bollen, K. A. (2019). Model Implied Instrumental Variables (MIIVs): An Alternative Orientation to Structural Equation Modeling. *Multivariate Behavioral Research, 54*(1), 31–46.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods & Research, 36*(1), 48–86.

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305–314.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8, 3–62.

Bono, J. E., & McNamara, G. (2011). From the editors: Publishing in AMJ–Part 2: Research Design. *Academy of Management Journal, 54*(4), 657–660.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Brandt, H., Umbach, N., Kelava, A., & Bollen, K. A. (2020). Comparing estimators for latent interaction models under structural and distributional misspecifications. *Psychological Methods, 25*(3), 321–345.

Braun, M. T., Converse, P. D., & Oswald, F. L. (2019). The accuracy of dominance analysis as a metric to assess relative importance: The joint impact of sampling error variance and measurement unreliability. *Journal of Applied Psychology, 104*(4), 593–602.

Busenbark, J. R., Yoon, H., Gamache, D. L., & Withers, M. C. (2022). Omitted Variable Bias: Examining Management Research With the Impact Threshold of a Confounding Variable (ITCV). *Journal of Management, 48*(1). 01492063211006458.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics, 29*(2), 238–249.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317–372.

Castille, C. M., Kreamer, L. M., Albritton, B. H., Banks, G. C., & Rogelberg, S. G. (2022). The open science challenge: Adopt one practice that enacts widely shared values. *Journal of Business and Psychology, 37*, 459–467.

Certo, S. T., Busenbark, J. R., Kalm, M., & LePine, J. A. (2020). Divided We Fall: How Ratios Undermine Research in Strategic Management. *Organizational Research Methods, 23*(2), 211–237.

Certo, S. T., Busenbark, J. R., Woo, H. S., & Semadeni, M. (2016). Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal, 37*(13), 2639–2657.

Chang, S.-J., van Witteloostuijn, A., & Eden, L. (2010). From the Editors: Common method variance in international business research. *Journal of International Business Studies, 41*(2), 178–184.

Chen, F., Curran, P. J., Bollen, K. A., Kirby, J., & Paxton, P. (2008). An empirical evaluation of the use of fixed cutoff points in RMSEA test statistic in structural equation models. *Sociological Methods & Research, 36*(4), 462–494.

Cho, E., & Kim, S. (2015). Cronbach's Coefficient Alpha. *Organizational Research Methods, 18*(2), 207–230.

Cinelli, C., Forney, A., & Pearl, J. (2022). A Crash Course in Good and Bad Controls. *Sociological Methods & Research*. https://doi.org/10.1177/00491241221099552.

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 82*(1), 39–67.

Cloutier, C., & Langley, A. (2020). What Makes a Process Theoretical Contribution? *Organization Theory, 1*(1). 2631787720902473.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*(3), 249–253.

Conway, J. M., & Lance, C. E. (2010). What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business and Psychology, 25*(3), 325–334.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98–104.

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., et al. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the Journal of Applied Psychology. *Journal of Applied Psychology, 105*(12), 1351–1381.

Credé, M., & Harms, P. D. (2015). 25 years of higher-order confirmatory factor analysis in the organizational sciences: a critical review and development of reporting recommendations. *Journal of Organizational Behavior, 36*(6), 845–872.

Criqui, M. H. (1991). On the use of standardized regression coefficients. *Epidemiology, 2*(5), 393.

Cuadros-Rodríguez, L., Pérez-Castaño, E., & Ruiz-Samblás, C. (2016). Quality performance metrics in multivariate classification methods for qualitative analysis. *TrAC Trends in Analytical Chemistry, 80*, 612–624.

Cubillos, M., Wulff, J. N., & Wøhlk, S. (2022). A bi-objective k-nearest-neighbors-based imputation method for multilevel data. *Expert Systems with Applications, 204*, 1–9.

Culpepper, S. A. (2012). Evaluating EIV, OLS, and SEM estimators of group slope differences in the presence of measurement error: The single-indicator case. *Applied Psychological Measurement, 36*(5), 349–374.
Dalal, D. K., & Zickar, M. J. (2012). Some Common Myths About Centering Predictor Variables in Moderated Multiple Regression and Polynomial Regression. *Organizational Research Methods, 15*(3), 339–362.
Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2–21.
Denyer, D., & Tranfield, D. (2009). Producing a systematic review. In D. A. Buchanan & A. Bryman (Eds.), *The Sage handbook of organizational research methods* (pp. 671–689). Sage Publications.
Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781.
Doldor, E., Wyatt, M., & Silvester, J. (2019). Statesmen or cheerleaders? Using topic modeling to examine gendered messages in narrative developmental feedback for leaders. *The Leadership Quarterly, 30*(5), 101308.
Dolgin, E. (2021). The tangled history of mRNA vaccines. *Nature, 597*(7876), 318–324.
Duflo, E., Glennerster, R., & Kremer, M. (2007). Chapter 61 - Using Randomization in Development Economics Research: A Toolkit. In T. P. Schultz & J. A. Strauss (Eds.), *Handbook of Development Economics* (pp. 3895–3962). Elsevier.
Duguid, M. M., & Goncalo, J. A. (2015). Squeezed in the middle: The middle status trade creativity for focus. *Journal of Personality and Social Psychology, 109*(4), 589–603.
Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association, 56*(293), 52–64.
Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412.
Durbin, J. (1954). Errors in variables. *Revue de l'institut International de Statistique*, 23–32.
Echambadi, R., & Hess, J. D. (2007). Mean-Centering Does Not Alleviate Collinearity Problems in Moderated Multiple Regression Models. *Marketing Science, 26*(3), 438–445.
Edwards, J. R. (1995). Alternatives to difference scores as dependent variables in the study of congruence in organizational research. *Organizational Behavior and Human Decision Processes, 64*(3), 307–324.
Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods, 4*(3), 265–287.
Edwards, J. R. (2002). Alternatives to difference scores: Polynomial regression and response surface methodology. In F. Drasgow & N. W. Schmitt (Eds.), *Advances in Measurement and Data Analysis* (pp. 350–400). Jossey-Bass.
Edwards, J. R. (2011). The Fallacy of Formative Measurement. *Organizational Research Methods, 14*(2), 370–388.
Edwards, J. R., & Berry, J. W. (2010). The Presence of Something or the Absence of Nothing: Increasing Theoretical Precision in Management Research. *Organizational Research Methods, 13*(4), 668–689.
Edwards, J. R., Berry, J. W., & Stewart, V. (2016). *Bridging the great divide between theoretical and empirical management research*. Kenan-Flagler Business School. University of North Carolina. Working paper.
Edwards, J. R., & Christian, M. S. (2014). Using accumulated knowledge to calibrate theoretical propositions. *Organizational Psychology Review*. 2041386614535131.
Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal, 36*(6), 1577–1613.
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*.
Ekwaru, J. P., & Veugelers, P. J. (2018). The Overlooked Importance of Constants Added in Log Transformation of Independent Variables with Zero Values: A Proposed Approach for Determining an Optimal Constant. *Statistics in Biopharmaceutical Research, 10*(1), 26–29.
Emsley, R., Dunn, G., & White, I. R. (2010). Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Statistical Methods in Medical Research, 19*(3), 237–270.
Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299.
Fairhurst, G. T., & Antonakis, J. (2012). A Research Agenda for Relational Leadership. In M. Uhl-Bien, & S. Ospina (Eds.), *Advancing Relational Leadership Theory: A Conversation among Perspectives* (pp. 433–459). Greenwich, CT: Information Age Publishing.
Fanelli, D., & Larivière, V. (2016). Researchers' Individual Publication Rate Has Not Increased in a Century. *PLoS ONE, 11*(3).
Fayant, M. P., Sigall, H., Lemonnier, A., Retsin, E., & Alexopoulos, T. (2017). On the limitations of manipulation checks: An obstacle toward cumulative science. *International Review of Social Psychology, 30*(1), 125–130.
Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry, 26*(2), 105–109.
Fischer, T., Dietz, J., & Antonakis, J. (2017). Leadership process models: A review and synthesis. *Journal of Management, 43*(6), 1726–1753.
Fischer, T., Hambrick, D. C., Sajons, G. B., & Van Quaquebeke, N. (2020). Beyond the ritualized use of questionnaires: Toward a science of actual behaviors and psychological states. *The Leadership Quarterly, 31*(4), 101449.
Fischer, T., & Sitkin, S. B. (2022). Leadership Styles: A Comprehensive Assessment and Way Forward. *Academy of Management Annals*. https://doi.org/10.5465/annals.2020.0340.
Fischer, T., Tian, A. W., Lee, A., & Hughes, D. J. (2021). Abusive supervision: A systematic review and fundamental rethink. *The Leadership Quarterly, 32*(6), 101540.
Fitzsimons, G. J. (2008). Death to dichotomizing. *Journal of Consumer Research, 35*(1), 5–8.
Forscher, B. K. (1963). Chaos in the brickyard. *Science, 142*(3590), 339.
Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research, 29*(2), 147–194.
Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology*. Sage Publications.
Galdas, P. (2017). Revisiting bias in qualitative research: Reflections on its relationship with funding and impact. *International Journal of Qualitative Methods, 16*(1). 1609406917748992.
Galinsky, A. D., Gruenfeld, D. H., & Magee, J. C. (2003). From power to action. *Journal of Personality and Social Psychology, 85*(3), 453–466.
Geddes, B. (2003). *Paradigms and sand castles: Theory building and research design in comparative politics*. University of Michigan Press.
Gerring, J. (2007). *Case study research: Principles and practices*. Cambridge University Press.
Gerring, J. (2012). *Social science methodology: A unified framework* (2nd ed.). Cambridge; New York: Cambridge University Press.
Gerring, J., & McDermott, R. (2007). An experimental template for case study research. *American Journal of Political Science, 51*(3), 688–701.
Gibbs, N. M., & Gibbs, S. V. (2015). Misuse of 'trend' to describe 'almost significant' differences in anaesthesia research. *BJA: British Journal of Anaesthesia, 115*(3), 337–339.
Gonzalez-Mulé, E., & Aguinis, H. (2018). Advancing theory by assessing boundary conditions with meta-regression: A critical review and best-practice recommendations. *Journal of Management, 44*(6), 2246–2273.
Gordon, M., Viganola, D., Bishop, M., Chen, Y., Dreber, A., Goldfedder, B., et al. (2020). Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme. *Royal Society Open Science, 7*(7).
Gottfredson, R. K., Wright, S. L., & Heaphy, E. D. (2020). A critique of the Leader-Member Exchange construct: Back to square one. *The Leadership Quarterly, 31*(6), 101385.
Greenland, S., Maclure, M., Schlesselman, J. J., Poole, C., & Morgenstern, H. (1991). Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology*, 387–392.
Greenland, S., Schlesselman, J. J., & Criqui, M. H. (1986). The fallacy of employing standardized regression-coefficients and correlations as measures of effect. *American Journal of Epidemiology, 123*(2), 203–208.
Grote, G. (2017). There is hope for better science. *European Journal of Work and Organizational Psychology, 26*(1), 1–3.
Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician, 63*(4), 308–319.
Grömping, U. (2015). Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics, 7*(2), 137–152.
Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics, 11*(1), 1–21.
Gupta, S. K. (2011). Intention-to-treat concept: a review. *Perspectives in Clinical Research, 2*(3), 109–112.
Hahn, J., Ham, J. C., & Moon, H. R. (2011). The Hausman test and weak instruments. *Journal of Econometrics, 160*(2), 289–299.
Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology, 30*, 507–544.
Hamilton, B. H., & Nickerson, J. A. (2003). Correcting for endogeneity in strategic management research. *Strategic Organization, 1*(1), 51–78.
Hankins, M. (2013). Still Not Significant. Probable Error. https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/.
Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica, 50*, 1029–1054.
Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., et al. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science, 5*(8), 180448.
Harms, C., & Lakens, D. (2018). Making "null effects" informative: Statistical techniques and inferential frameworks. *Journal of Clinical and Translational Research, 3*(2), 382–393.
Hartung, J., Knapp, G., & Sinha, B. K. (2008). Meta-regression. In J. Hartung, G. Knapp, & B. K. Sinha (Eds.), *Statistical meta-analysis with applications* (pp. 127–137). John Wiley & Sons, Inc.
Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications.
Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology, 9*, 998.
Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society, 46*(6), 1251–1271.
Hausman, J. A., Stock, J. H., & Yogo, M. (2005). Asymptotic properties of the Hahn-Hausman test for weak-instruments. *Economics Letters, 89*(3), 333–342.
Hayduk, L. A. (2014). Shame for disrespecting evidence: the personal consequences of insufficient respect for structural equation model testing. *BMC Medical Research Methodology, 14*, 1–10.
Hayduk, L. A., Cummings, G., Boadu, K., Pazderka-Robinson, H., & Boulianne, S. (2007). Testing! testing! one, two, three - Testing the theory in structural equation models! *Personality and Individual Differences, 42*(5), 841–850.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economics and Social Measurement, 5*(4), 475–492.
Heckman, J. J. (1979). Sample Selection Bias As A Specification Error. *Econometrica, 47*(1), 153–161.
Heggestad, E., Scheaf, D., Banks, G. C., Hausfeld, M. M., Tonidandel, S., & Williams, E. (2019). Scale adaptation in organizational science research: A review and best-practice recommendations. *Journal of Management, 45*(6), 2596–2627.
Herzog, W., & Boomsma, W. (2009). Small-sample robust estimators of noncentrality-based and incremental model fit. *Structural, Equation Modeling, 16*(1), 1–27.
Herzog, W., Boomsma, W., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling, 14*(3), 361–390.
Hickman, L., Thapa, S., Tay, L., Cao, M., & Srinivasan, P. (2022). Text preprocessing for text mining in organizational research: Review and recommendations. *Organizational Research Methods, 25*(1), 114–146.
Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.
Hoetker, G. (2007). The use of logit and probit models in strategic management research: Critical issues. *Strategic Management Journal, 28*(4), 331–343.
Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics., 6*(2), 65–70.
Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data Sharing in Psychology: A Survey on Barriers and Preconditions. *Advances in Methods and Practices in Psychological Science, 1*(1), 70–85.
Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.
Hubbard, R. (2004). Alphabet Soup: Blurring the Distinctions Betweenp's anda's in Psychological Research. *Theory & Psychology, 14*(3), 295–327.
Hughes, D. J., Lee, A., Tian, A. W., Newman, A., & Legood, A. (2018). Leadership, creativity, and innovation: A critical review and practical recommendations. *The Leadership Quarterly, 29*(5), 549–569.
Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309–334.
Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science, 25*(1), 51–71.
Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 5–51.
Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(1), 5–86.
Ioannidis, J. P. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly, 94*(3), 485–514.
Jackson, D. L., Voth, J., & Frey, M. P. (2013). A Note on Sample Size and Solution Propriety for Confirmatory Factor Analytic Models. *Structural Equation Modeling, 20*(1), 86–97.
Johnson, S. R., & Rausser, G. C. (1971). Effects of Misspecifications of Log-Linear Functions When Sample Values Are Zero or Negative. *American Journal of Agricultural Economics, 53*(1), 120–124.
Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183–202.
Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: a qualitative and quantitative review. *Journal of Applied Psychology, 87*(4), 765–780.
Kasof, J. (1993). Sex bias in the naming of stimulus persons. *Psychological Bulletin, 113*(1), 140–163.
Kennedy-Shaffer, L. (2019). Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing. *The American Statistician, 73*(Suppl 1), 82–90.
Kepes, S., Banks, G. C., McDaniel, M. A., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods, 15*(4), 624–662.
Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. (2013). Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS. *Journal of Business and Psychology, 28*, 123–142.
Khademi, M., Mast, M. S., Zehnder, C., & De Saint Priest, O. (2021). The problem of demand effects in power studies: Moving beyond power priming. *The Leadership Quarterly, 32*(4), 101496.
Kidd, R. F. (1976). Manipulation checks: Advantage or disadvantage? *Representative Research in Social Psychology, 7*(2), 160–165.
Kim, J.-O., & Ferree, G. D. (1981). Standardization in Causal Analysis. *Sociological Methods & Research, 10*(2), 187–210.
King, G., Honacker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review, 95*(1), 49–69.
Kleibergen, F., & Paap, R. (2006). Generalized reduced rank tests using the singular value decomposition. *Journal of Econometrics, 133*(1), 97–126.
Klein, A. G., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65*(4), 457–474.
Kleinbauer, T. R., Rönkkö, M., & Antonakis, J. (2020). Examining the Use and Utility of Dominance and Relative Weights Analysis. *Paper presented at the Academy of Management Proceedings, 2020*(1), 21797.
Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.
Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational Research Methods, 21*(3), 733–765.
Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives, 15*(4), 143–156.
Kohler, U., & Kreuter, F. (2005). *Data analysis using Stata*. Stata Press.
Kromrey, J. D., & Foster-Johnson, L. (1998). Mean Centering in Moderated Multiple Regression: Much Ado about Nothing. *Educational and Psychological Measurement, 58*(1), 42–67.
Kruschke, J. K. (2011). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science, 6*(3), 299–312.
Lakens, D. (2022). *Improving Your Statistical Inferences*. https://lakens.github.io/statistical_inferences/. https://doi.org/10.5281/zenodo.6409077.
Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving Inferences About Null Effects With Bayes Factors and Equivalence Tests. *The Journals of Gerontology: Series B, 75*(1), 45–57.
Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269.
Lan, Q., Xu, X., Ma, H., & Li, G. (2020). Multivariable data imputation for the analysis of incomplete credit data. *Expert Systems with Applications, 141*.
Langley, A. (1999). Strategies for theorizing from process data. *Academy of Management Review, 24*(4), 691–710.
Larivière, V., & Costas, R. (2016). How Many Is Too Many? On the Relationship between Research Productivity and Impact. *PLoS ONE, 11*(9).
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.
Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766.
Li, J., Ding, H., Hu, Y., & Wan, G. (2021). Dealing with dynamic endogeneity in international business research. *Journal of International Business Studies, 52*(3), 339–362.
Liang, L. H., Brown, D. J., Lian, H., Hanig, S., Ferris, D. L., & Keeping, L. M. (2018). Righting a wrong: Retaliation on a voodoo doll symbolizing an abusive supervisor restores justice. *The Leadership Quarterly, 29*(4), 443–456.
List, J. A., Shaikh, A. M., & Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics, 22*(4), 773–793.
Lonati, S. (2020). What explains cultural differences in leadership styles? On the agricultural origins of participative and directive leadership. *The Leadership Quarterly, 31*(2), 101305.
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.
Lonati, S., Rönkkö, M., & Antonakis, J. (2020). Violation of Distributional Assumptions in Latent Interaction Models. *Academy of Management Proceedings, 2020*(1), 18911.
Lonati, S., & Wulff, J. N. (2023). *A Critical Evaluation Of The Impact Threshold Of A Confounding Variable In Management Research*. Unpublished working paper.
MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modification in covariance structure-analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490–504.
MacCallum, R. C., Zhang, S. B., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40.
MacKinnon, J. G., & Webb, M. (2019). *When and how to deal with clustered errors in regression models*. Working Paper 1421, Economics Department, Queen's University.
Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.
Manning, W. G., & Mullahy, J. (2001). Estimating log models: To transform or not to transform? *Journal of Health Economics, 20*(4), 461–494.
Mansfield, L. (2005). The reading, writing, and arithmetic of the medical literature, part 2: Critical evaluation of statistical reporting. *Annals of Allergy, Asthma & Immunology: Official Publication of the American College of Allergy, Asthma, & Immunology, 95*(4), 315–322; quiz 322, 380.
Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9*(3), 275–300.
Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*(1), 181–190.
McDermott, R. (2023). The scientific study of small samples. *The Leadership Quarterly., 101675*.
McIntosh, C. N. (2007). Rethinking fit assessment in structural equation modelling: A commentary and elaboration on Barrett (2007). *Personality and Individual Differences, 42*(5), 859–867.
McLachlan, G. J., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.
McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433.
McNeish, D. (2020). Should We Use F-Tests for Model Fit Instead of Chi-Square in Overidentified Structural Equation Models? *Organizational Research Methods, 23*(3), 487–510.
McNeish, D., & Kelley, K. (2019). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods, 24*(1), 20–35.
McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*(1), 114.
Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science, 62*(3), 760–775.
Morse, J. M., Barrett, M., Mayan, M., Olson, K., & Spiers, J. (2002). Verification strategies for establishing reliability and validity in qualitative research. *International Journal of Qualitative Methods, 1*(2), 13–22.
Mutz, D. C., Pemantle, R., & Pham, P. (2019). The perils of balance testing in experimental design: Messy analyses of clean data. *The American Statistician, 73*(1), 32–42.
Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods, 17*(4), 372–411.
Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia. *Perspectives on Psychological Science, 7*(6), 615–631.
O'Boyle, E. H., Jr., Banks, G. C., & Gonzalez-Mule, E. (2017). The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management, 43*(2), 376–399.
Olson, K. (2008). Double-barreled question. *Encyclopedia of Survey Research Methods, 10* (9781412963947), n145.
Olsson-Collentine, A., van Assen, M. A. L. M., & Hartgerink, C. H. J. (2019). The Prevalence of Marginally Significant Results in Psychology Over Time. *Psychological Science, 30*(4), 576–586.
Onwuegbuzie, A. J., & Leech, N. L. (2007). Validity and qualitative research: An oxymoron? *Quality & Quantity, 41*(2), 233–249.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716-4711/4718.
Orne, M. T. (2009). Demand characteristics and the concept of quasi-controls. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifacts in behavioral research* (pp. 110–137). Oxford University Press.
Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5*(3), 343–355.
Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics, 37*(2), 187–204.
Ostroff, C., Kinicki, A. J., & Clark, M. A. (2002). Substantive and operational issues of response bias across levels of analysis: An example of climate-satisfaction relationships. *The Journal of Applied Psychology, 87*(2), 355–368.
Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resource management: Forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior, 7*, 505–533.
Pashler, H., & Wagenmakers, E. J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science, 7*(6), 528–530.
Patton, M. Q. (1990). Qualitative evaluation and research methods. *Sage*. Publications.
Petitti, D. B. (2000). *Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine, Oxford University Press: New York*. UK: USA and Oxford.
Pfeffer, J. (1993). Barriers to the advance of organizational science: Paradigm development as a dependent variable. *The Academy of Management Review, 18*(4), 599–620.
Pirlott, A. G., & MacKinnon, D. P. (2016). Design approaches to experimental mediation. *Journal of Experimental Social Psychology, 66*, 29–38.
Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903.
Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control It. *Annual Review of Psychology, 63*(1), 539–569.
Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*(4), 717–731.
Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics, 29*, 15–24.
Ramalho, E. A., & Ramalho, J. J. S. (2017). Moment-based estimation of nonlinear regression models with boundary outcomes and endogeneity, with applications to nonnegative and fractional responses. *Econometric Reviews, 36*(4), 397–420.
Ramsey, J. B. (1969). Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis. *Journal of the Royal Statistical Society. Series B (Methodological), 31*(2), 350–371.
Reich, E. S. (2012). *Embattled neutrino project leaders step down*. Nature News. 2 April, 2012.
Richardson, H. A., Simmering, M. J., & Sturman, M. C. (2009). A Tale of Three Perspectives: Examining Post Hoc Statistical Techniques for Detection and Correction of Common Method Variance. *Organizational Research Methods, 12*(4), 762–800.
Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23*(1), 51–67.
Rönkkö, M., Aalto, E., Tenhunen, H., & Aguirre-Urreta, M. I. (2022). Eight simple guidelines for improved understanding of transformations and nonlinear effects. *Organizational Research Methods, 25*(1), 48–87.
Rönkkö, M., McIntosh, C. N., Antonakis, J., & Edwards, J. R. (2016). Partial least squares path modeling: Time for some serious second thoughts. *Journal of Operations Management, 47–48*, 9–27.
Ropovik, I. (2015). A cautionary note on testing latent variable models. *Frontiers in Psychology, 6*(1715), 1–8.
Roulston, K., & Shelton, S. A. (2015). Reconceptualizing bias in teaching qualitative research methods. *Qualitative Inquiry, 21*(4), 332–342.
Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*(411), 633–639.
Sagarin, B. J., West, S. G., Ratnikov, A., Homan, W. K., Ritchie, T. D., & Hansen, E. J. (2014). Treatment noncompliance in randomized experiments: statistical approaches and design issues. *Psychological Methods, 19*(3), 317–333.
Sajons, G. B. (2020). Estimating the causal effect of measured endogenous variables: A tutorial on experimentally randomized instrumental variables. *The Leadership Quarterly, 31*(5), 101348.
Santos Silva, J. M. C., & Tenreyro, S. (2006). The log of gravity. *Review of Economics and Statistics, 88*(4), 641–658.
Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society, 26*(3), 393–415.
Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science, 2*(6), 420.
Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement, 72*(6), 910–932.
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.
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.
Schunck, R. (2013). Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models. *The Stata Journal, 13*(1), 65–76.
Shang, G., & Rönkkö, M. (2022). Empirical research methods department: Mission, learnings, and future plans. *Journal of Operations Management, 68*(2), 114–129.
Shipley, B. (2000). *Cause and correlation in biology: A user's guide to path analysis, structural equations, and causal inference*. Cambridge University Press.
Sieweke, J., & Santoni, S. (2020). Natural experiments in leadership research: An introduction, review, and guidelines. *The Leadership Quarterly, 31*(1), 101338.
Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika, 74*(1), 107–120.
Simonsohn, U. (2015). *How to Study Discrimination (or Anything) With Names; If You Must*. https://datacolada.org/36.
Spector, P. E. (1992). *Summated rating scale construction: An introduction* (Vol. 82). Sage.
Stanton, J. M. (2021). Evaluating Equivalence and Confirming the Null in the Organizational Sciences. *Organizational Research Methods, 24*(3), 491–512.
StataCorp (2021). *Stata 17 Base Reference Manual*. Stata Press.
Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ (Clinical Research Ed.), 338*.
Stock, G., Banks, G. C., Voss, N., Woznjy, H., & Tonidandel, S. (2022). Putting leader (follower) behavior back into transformational leadership: A theoretical and empirical course correction. *The Leadership Quarterly*, 101632. https://doi.org/10.1016/j.leaqua.2022.101632.
Stock, J., & Yogo, M. (2005). Asymptotic Distributions of Instrumental Variables Statistics with Many Instruments. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg* (pp. 109–120). New York: Cambridge University Press.
Sturm, R. E., & Antonakis, J. (2015). Interpersonal power: A review, critique, and research agenda. *Journal of Management, 41*(1), 136–163.
Swain, A. J. (1975). *Analysis of parametric structures for variance matrices (doctoral thesis)*. Adelaide: University of Adelaide.
Symon, G. E., & Cassell, C. E. (1998). *Qualitative methods and analysis in organizational research: A practical guide*. Sage Publications Ltd..
Thomas, D. R., Zumbo, B. D., Kwan, E., & Schweitzer, L. (2014). On Johnson's (2000) relative weights method for assessing variable importance: A reanalysis. *Multivariate Behavioral Research, 49*(4), 329–338.
Tian, L., Jiang, Y., & Yang, Y. (2022). CEO childhood trauma, social networks, and strategic risk taking. *The Leadership Quarterly, 34*(2), 101618.
Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*(3), 207–222.
Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best Alternatives to Cronbach's Alpha Reliability in Realistic Conditions: Congeneric and Asymmetrical Measurements. *Frontiers in Psychology, 7*, 769.
van der Heijden, G. J. M. G., Donders, A. R. T., Stijnen, T., & Moons, K. G. M. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology, 59*(10), 1102–1109.
Verardi, V., & Croux, C. (2009). Robust regression in Stata. *The Stata Journal, 9*(3), 439–453.
Villadsen, A. R., & Wulff, J. N. (2021). Are you 110% sure? Modeling of fractions and proportions in strategy and management research. *Strategic Organization, 19*(2), 312–337.
Villadsen, A. R., & Wulff, J. N. (2021). Statistical Myths About Log-Transformed Dependent Variables and How to Better Estimate Exponential Models. *British Journal of Management, 32*(3), 779–796.
Vitanova, I. (2021). Nurturing overconfidence: The relationship between leader power, overconfidence and firm performance. *The Leadership Quarterly, 32*(4).
Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics, 11*(3), 284–300.
Wilson, E. O. (1998). *Consilience: The unity of knowledge* (1st ed.). Knopf: Distributed by Random House.

Wiseman, R. M. (2009). On the use and misuse of ratios in strategic management research. In D. D. Bergh & D. J. Ketchen (Eds.), *Research Methodology in Strategy and Management* (Vol. 5, pp. 75–110).

Wolfolds, S. E., & Siegel, J. (2019). Misaccounting for endogeneity: The peril of relying on the Heckman two-step method without a valid instrument. *Strategic Management Journal, 40*(3), 432–462.

Wood, J., Freemantle, N., King, M., & Nazareth, I. (2014). Trap of trends to statistical significance: Likelihood of near significant P value becoming more significant with extra data. *BMJ, 348*.

Wooldridge, J. M. (2002). *Introductory econometrics: A modern approach: South-Western, Div*. of Thomson Le.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT Press.

Wooldridge, J. M. (2014). Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Journal of Econometrics, 182*(1), 226–234.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources, 50*(2), 420–445.

Wright, P. M. (2017). Making great theories. *Journal of Management Studies, 54*(3), 384–390.

Wu, D.-M. (1974). Alternative tests of independence between stochastic regressors and disturbances: Finite sample results. *Econometrica: Journal of the Econometric Society, 42*(3), 529–546.

Wulff, J. N. (2015). Interpreting results from the multinomial logit model: Demonstrated by foreign market entry. *Organizational Research Methods, 18*(2), 300–325.

Yuan, K.-H., & Bentler, P. M. (2004). On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*(5), 737–757.

Yuan, K.-H., Tian, Y., & Yanagihara, H. (2015). Empirical correction to the likelihood ratio statistic for structural equation modeling with many variables. *Psychometrika, 80*(2), 379–405.

Zaccaro, S. J. (2012). Individual differences and leadership: Contributions to a third tipping point. *The Leadership Quarterly, 23*(4), 718–728.

Zaccaro, S. J., Green, J. P., Dubrow, S., & Kolze, M. (2018). Leader individual differences, situational parameters, and leadership outcomes: A comprehensive review and integration. *The Leadership Quarterly, 29*(1), 2–43.

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: an object-oriented implementation of clustered covariances in R. *Journal of Statistical Software, 95*, 1–36.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics, 13*, 75–98.

Yang, P., Riepe, J., Moser, K., Pull, K., & Terjesen, S. (2019). Women directors, firm performance, and firm risk: A causal perspective. *The Leadership Quarterly, 30*(5), 101297.