CAMBRIDGE
UNIVERSITY PRESS

FOCAL ARTICLE

In praise of Table 1: The importance of making better use of descriptive statistics

Kevin R. Murphy

Department of Work and Employment Studies, Kemmy Business School, University of Limerick, Ireland
Corresponding author. Email: Kevin.R.Murphy@ul.ie

Abstract

As data analytic methods in the managerial sciences become more sophisticated, the gap between the descriptive data typically presented in Table 1 and the analyses used to test the principal hypotheses advanced has become increasingly large. This contributes to several problems including: (1) the increasing likelihood that analyses presented in published research will be performed and/or interpreted incorrectly, (2) an increasing reliance on statistical significance as the principal criterion for evaluating results, and (3) the increasing difficulty of describing our research and explaining our findings to non-specialists. A set of simple methods for assessing whether hypotheses about interventions, moderator relationships and mediation, are plausible that are based on the simplest possible examination of descriptive statistics are proposed.

Keywords: descriptive statistics

As scholarship in the organizational sciences has developed and matured, the range and complexity of data analytic methods has expanded considerably. A casual perusal of any issue of leading journals in this field is likely to make this evident, but the point is driven home vividly if you systematically examine current publications. For example, I reviewed all of the papers in *Journal of Management* and *Journal of Applied Psychology* published in 2018 (Volumes 44 and 103, respectively). There were several papers that were editorials, recommendations regarding research methods, meta-analyses, or theory-development papers, but there were over 120 papers that used quantitative methods to test the principal hypotheses advanced by the authors. Approximately 15% of these papers relied on ordinary least squares (OLS) regression or some variant (e.g., ANOVA) to test their main hypotheses.[1] Approximately 10% used regression with qualitative dependent variables (e.g., logistic, probit, or tobit analyses). Moderated or mediated regression methods were quite popular (approximately 18%); advanced variants of regression (e.g., polynomial regression, two-stage regression, random coefficient models, panel regression) were equally popular (approximately 19.0%). Latent variable analyses and multilevel analyses were common (approximately 15% and 17%, respectively), and other methods (e.g., event history/ hazard models, visualization methods) were also used (approximately 6%).[2] On the whole, this

---

[1]In multistudy papers, I used the last study that tested a substantive hypothesis to classify the analytic methods used.

[2]These classifications are necessarily subjective and incomplete because many studies used multiple methods and some used quite different methods to test different hypotheses. However, this brief summary does give a concrete picture of the range and complexity of the methods used in papers published in this journal.

quick review reinforces the idea that complex and sophisticated methods of data analysis have become common in organizational research.

The increasing complexity and diversity of data-analytic methods in organizational research has created several problems in our field including (a) incorrect application and interpretation of these analyses, (b) increasing reliance on significance testing, and (c) increasing difficulty in interpretation and accompanying gaps between science and practice. In this paper, I will argue that we could make progress on resolving all of these problems if we readjusted the emphasis given to descriptive versus inferential statistics in making sense of the data we collect. I will argue that current research pays scant attention to descriptive statistics and that shifting our focus from the relatively complex analyses that have come to characterize much of the research published in the organizational sciences to simple and informative presentations of descriptive statistics would increase the value and interpretability of our research.

Are descriptive statistics ignored, and if so, who cares?

To illustrate the treatment of descriptive statistics that is typical of our field, I selected a recent issue of the two leading journals noted earlier (i.e., *JOM*, Vol. 44, No. 8; *JAP*, Vol. 103, No. 10) that included a number of papers that use analyses of empirical data (as opposed to literature reviews, editorials, etc.) to test study hypotheses. Virtually every paper that used the analysis of primary data to test hypotheses included a statement like "Means, standard deviations, and correlations among the variables measured in this study are shown in Table 1." The question was what else, if anything, these studies had to say about the descriptive statistics that were dutifully presented in their articles. As in my review of the statistical methods that were used in these journals, I focused on Study 1 or the principal study in papers that reported multiple studies.

There were 17 papers in these two issues that used analyses of primary data to test study hypotheses. Every one of these papers presented detailed descriptions of the samples and measures employed, but their discussion of descriptive data were scant. For example, in *JAP*, there were five studies that used data to test specific hypotheses, and three of them said nothing beyond the fact that there was a Table 1 or its equivalent. One study cited descriptive statistics to justify a decision to not use particular control variables. One paper had a full sentence describing descriptive statistics, but this paper did not include a table of these statistics. In *JOM* there were 12 papers that used data analyses to test important hypotheses. Across 11 of these papers, there was a total of 15 sentences referring to Table 1 or its equivalent; two of these 11 papers made no reference whatsoever to the contents of Table 1. In one study (Desai, 2018), a full paragraph and a part of another discussed how the correlations among some of the variables measured in that study had implications for tests of particular models.

In some studies, information that at least refers back to the descriptive statistics is included in figures, but in virtually all of the studies reviewed above, the authors skip over their descriptive statistics and go straight to a discussion of their more complex analyses. I would argue that this is a serious error for two reasons. First, significant experience and skill are often required to concretely interpret the coefficients and statistics that many of these analytic procedures provide. Sometimes, authors provide information that can help the reader make sense of the meaning of model parameters. For example, Giennik et al. (2018), in describing some of their results, note,

> [T]he path model shows a positive and significant effect of entrepreneurial intentions on entrepreneurial activity ($B = .041$, $p < .001$). This result provided support for Hypothesis 2. The marginal effect of entrepreneurial intentions was 0.05, suggesting that an increase of one unit of entrepreneurial intentions increased the probability of engaging in entrepreneurial activity by five percentage points. (p. 1074)

However, this is not the norm. Rather, the norm is to indicate which coefficients are statistically significant or which model provides a significantly better fit than another does, reducing the

Table 1. Proportion of JAP and JOM Analyses That Include Effect Size Estimates

Analytic method	Proportion
OLS regression	.735
Multilevel modeling	.373
Logistic regression	.400
SEM	.230
Other	.000

interpretation of study results to an exercise in "stargazing"—that is, scanning tables to see which models and which coefficients are starred as statistically significant. These tables virtually always present the values of fit statistics and coefficients, but it is often difficult for organizational researchers to interpret these values in any substantive way.

Three reasons for concern with the relative emphasis on inferential versus descriptive statistics

There are several trends in the use of inferential statistics and statistical hypothesis testing in contemporary research in the organizational sciences that combine to make our research increasingly difficult to conduct appropriately, understand, and apply. Three strike me as particularly worrisome.

Increasingly complex analyses

Research articles in the organizational sciences usually pay a great deal of attention to inferential statistics (e.g., tests of parameter significance, model fit). Unfortunately, as the complexity and diversity of data analytic methods has increased, the likelihood of incorrect application and interpretation of these analyses has also increased. For example, Cortina et al. (2017) reviewed over 700 structural equation models that were presented in 75 separate papers in top journals, asking a relatively simple question: Were the degrees of freedom for the models these papers purport to test correct? They found that in nearly 40% of the cases where there was sufficient information to correctly calculate degrees of freedom, the reported *df* were wrong and that there were often reasons to question whether the models actually being tested were the same as those being discussed in the paper. In a related vein, Green et al. (2016) examined methodological and statistical issues raised by reviewers in the organizational sciences and noted that they often include concerns about a variety of common analytic techniques (e.g., factor analysis, structural equation modeling, hierarchical linear models). These analytic methods are widely used and widely taught in our field; frequent errors in applying these long-established methods do not bode well when considering the likelihood or errors in less familiar analytic methods.

Organizational researchers seem to have a difficult time understanding, much less properly analyzing, even the most common types of hypotheses if those hypotheses involve complex relationships among multiple variables. For example, Aguinis et al. (2017) document the frequent confusion and errors in tests of some of the most common hypotheses in the organizational sciences—that is, hypotheses involving moderator and mediator variables. Not only are there frequent errors of commission in applying complex statistical methods; there are also frequent errors of omission. For example, DeSimon et al.(2019) document how researchers routinely focus on the simplest details of meta-analytic studies, reporting only whether an effect exists and (sometimes) its approximate strength, ignoring the wealth of analytic information available in a modern meta-analysis (e.g., variation in effect sizes, systematic vs. random variation, boundary conditions).

Organizational researchers routinely *use* complex methods of data analysis, but it is far from clear that they *understand* them. As analyses become more complex, there is every reason to believe that researchers' ability to correctly apply and interpret them will continue to degrade.

Overreliance on null hypothesis significance testing

There is a growing trend in applying analytic methods that rely exclusively on the outcomes of null hypothesis significance tests, most often tests of the significance of individual model parameters, to judge findings, with little attention given to (and often, little opportunity to compute and present) effect-size estimates or other concrete indicators of what a significant finding might actually mean. This can be illustrated by examining the analyses presented in the papers in the 2018 volumes of *Journal of Management* and *Journal of Applied Psychology* discussed earlier. I classified the analytic methods of papers that performed statistical significance tests of primary data (meta-analyses, literature reviews, and comment papers were excluded) into one of five groups: (a) regression—including variants of OLS regression (e.g., ANOVA, *t* tests) and two-stage regression; (b) multilevel models; (c) logistic regression—including TOBIT and LOGIT analyses; (d) structural equation models (SEMs); and (e) other—generalized linear models, panel analyses, and random coefficient models. I examined each of these papers to determine whether some measure of effect size was present in analyses that included statistical significance tests. Table 1 shows the proportion of studies in each category that presented some sort of effect size measures (e.g., *d*, eta squared, R^2, pseudo R^2).

As this table suggests, null hypothesis tests conducted in the broad framework of regression are usually (73.5%) accompanied by some sort effect-size indicator. There is an important caveat here. There is little evidence in most of the studies I reviewed that effect-size estimates were actually taken into account or that they had any clear influence on the interpretation of null hypothesis tests in these papers. At least, however, there was some possibility that effect-size information *might* have been used. In every other category of analyses I reviewed, effect-size information was rarely presented and was virtually never discussed. None of the generalized linear model analyses of analyses in the "other" category included effect-size estimates; for some analyses, it is not clear that such an indicator can be created. In logistic and multilevel analysis, effect-size measures are often available, but more often than not they are not presented at all, much less taken into account when evaluating the significance of model parameters. In SEM analyses, the goal of the analysis was often to explain a dependent variable or a set of dependent variables in terms of some sets of exogenous variables, mediators, or moderators, but less than a quarter of the analyses using this technique presented any information about how well the model actually explains those DVs. In the great majority of SEM studies, the reader might learn how well the model fit the data but have little idea how well it explained the DVs. These results suggest a clear and disturbing trend. As analytic methods move from the familiar and comfortable ground of OLS regression toward a range of other analytic methods, the likelihood that effect-size measures will at least be available to aid in interpreting statistical significance tests is likely to decrease and reliance on statistical significance alone is likely to become more important in evaluating study results.

Research on the statistical power of null hypothesis tests makes one discouraging fact clear; the outcomes of these tests depend much more on the size of the sample than on the strength or nature of the effect being tested (Cohen, 1988; Kraemer & Thiemann, 1987; Maxwell et al., 2008; Murphy et al., 2014). When you reject the null hypothesis that some parameter in a statistical model is zero, you learn two things. The value of the parameter is substantially larger than its standard error (in most cases, at least 1.96 times as large) and the sample size was large enough to result in a small standard error.[3] When you fail to reject the null hypothesis, all you learn is that your sample was not large enough to give you sufficient statistical power to reject that particular

---

[3]Most standard error statistics involve some function of *N* in the divisor.

null hypothesis. With a large enough sample, virtually *any* null hypothesis will be rejected (Murphy et al., 2014).

Effect-size measures are widely advocated as an accompaniment to many types of significance tests (Applebaum et al., 2018; Kelly & Preacher, 2012; Valentine & Cooper, 2003; Wilkinson & Task Force on Statistical Inference, 1999). Many of the analytic methods that are used in current organizational research do not provide meaningful information about effect sizes, forcing researchers to rely on tests of null hypotheses to draw inferential conclusions. In many cases, there is hardly any point in doing these tests. Regardless of the hypothesis or model being tested, the null hypothesis will be rejected if the sample is very large and will not be rejected if the sample is small.

There is an extensive literature dealing with conceptual and statistical arguments against over-reliance on tests of statistical significance (See, for example, Cohen, 1994; Cortina & Dunlap, 1997; Meehl, 1978; Murphy et al., 2014; Schmidt, 1996), and there are many points of controversy in this debate. There is, however, agreement that significance tests by themselves are rarely enough to allow you to understand the meaning of data and that these tests should be accompanied or augmented, where possible, with information that will make the meaning of those tests more concrete. Unfortunately, at the very time that scientists across a wide range of disciplines are rising up against the uncritical use of significance tests as tools for evaluating research results (Amrhein et al., 2019; Wasserstein & Lazar, 2016; Wasserstein et al., 2019), organizational scientists are increasing their reliance on these dubious tools.

Our research is increasingly difficult to understand and explain

There is a substantial literature dealing with the research–practice gap in the organizational sciences (Caetano & Santos, 2017; Rynes, 2009). One of the recurring themes in this literature is the potential conflict between rigor and relevance. I believe that the increasing use of complex analytic methods has both enabled and encouraged increasingly complex theories and formulating complex research questions, sometimes leading to research findings that are difficult to understand or apply. I will cite the first empirical article published in the 2018 issue of *Journal of Applied Psychology*. Wolfson et al. (2018) studied the effects and boundary conditions of informal field-based learning (IFBL). In their discussion of the principal results of Study 3, they note,

> We hypothesized and found nuanced relationships between promotion focus and IFBL behaviors, as well as between IFBL behaviors and performance improvements. We showed that at high levels, non-punitive climate has an amplifying effect on the IFBL–changes in job performance relationship, whereas it reverses at low levels and demonstrates a negative relationship. (p. 27)

This conclusion is based almost entirely on statistically significant interaction effects in a multi-level analysis. As Murphy and Russell (2017) note, moderator effects of this sort are notoriously difficult to replicate, which suggests that considerable caution should be observed in interpreting and applying this sort of result.

I did not cite Wolfson et al. (2018) because it was unusually complex or difficult to interpret; I decided to describe the first study reporting empirical results in that issue of *JAP* regardless of the methods that were used or the results reported. In many ways, this paper is exemplary in its attempts to make complex findings more concrete by illustrating regions of significance in a number of figures, showing where moderator effects might make a difference and where they are less likely to do so. Nonetheless, it might be quite difficult to describe nuanced findings like this and their action implications to organizational decision makers. You could certainly recommend that organizations strive for a less punitive climate, but describing how and when the effects and determinants of IFBL might flip their directions and what organizations should do about this is likely to be a real challenge.

Complex research involving complex effects is sometimes necessary and can even be critically important, but when conclusions about these effects rely entirely on null hypothesis, tests of complex models can be difficult to interpret or explain. As a result of the proliferation of these methods, I believe we are well on the road to creating science that will be increasingly difficult to translate into practice. More to the point, we are well along the road to making important decisions about what the data mean based on analyses that most researchers and practitioners may not fully understand and would be hard pressed to explain in simple and concrete terms.

Making better use of Table 1

Why do researchers say so little about (and often completely ignore) the descriptive statistics they present in Table 1? One possibility is that they simply do not know what to say. That is, we have well-rehearsed routines for presenting models and coefficients but no clear template for what to do about the contents of Table 1.

I propose two uses for the information in Table 1, both of which would improve manuscripts, reducing the likelihood that the producers and consumers of research will misinterpret the results of studies. First, researchers should routinely discuss the contents of Table 1 in their Methods section. The contents of Table 1 provide critical information about what happened when the methods by which measures are defined and data are collected hit the realities of the context in which those data are obtained. A rigorous and elegant set of methods might not provide useful information if, for example, there is extensive range restriction for key variables in the samples employed.

Proposition: Descriptive statistics are the result of the application a set of research methods to a particular sample and context. They should be discussed in a level of detail commensurate with the discussion devoted to the methods themselves.

Discussing descriptive statistics as part of the Methods section of the paper would help the readers understand whether and to what extent well-developed methods yield information that at least has the potential to shed light on the primary questions with which the paper deals. In particular, a close examination of descriptive statistics can help to determine whether the main boundary conditions of a model or the predictions coming out of a model are likely to be met.

Table 1: Methods or results?

The Methods sections of articles often include detailed information about both the samples and the measures used in a study, but descriptive statistics are almost always relegated to the first sentence or two of the Results section and often never discussed again. It is time to break the rigid barrier between Method and Results.

I propose that the contents of Table 1 should be routinely discussed in the Methods section. The distributions of key measures are as much a part of the sample characteristics as the demographic variables that are routinely discusses in this section. The intercorrelations among variables are as much a part of understanding the measures that are used as the coefficient alpha values that are routinely presented when discussing each scale. If you feel that you *must* maintain a strict distinction between Methods and Results, move Table 1 into the Methods section.

Table 1 not only provides data about characteristics of the sample and of the measures employed, it also provides important information about whether or not the boundary conditions that define the study have been met. If the study deals with the way organizations respond to adverse weather events but the data were all collected on sunny days, this is a methodological failing, not something to be stuck in Table 1 and ignored. If a study deals with the way people respond to merit-based pay raises but everyone gets the same performance ratings (and therefore the same raises), your methods, *as applied in this study*, have not provided the sort of data that will allow you to ask the research question that motivated the study. This might reveal a problem with

your measures (e.g., performance ratings in the organization you sampled might be so inflated that there are few differences in overall ratings), the sample (perhaps you have only sampled top performers), or both. The point is that the statistics in Table 1 provide important information about whether or not the methods used will allow you to sensibly ask your research question.

Moving Table 1 (at least conceptually) into the Methods section, by discussing in detail how the data that have been collected provide information about the sample, the measures, and the positioning of the study in terms of the processes hypothesized, provides both the producers and consumers of research a much more detailed and concrete understanding of the data. This understanding will be critical when it comes time to reach conclusions about what the data mean.

There are many good guides for interpreting and illustrating descriptive data. Bedeian (2014), for example, offers advice on what to look for when examining descriptive data (e.g., pay attention to distributions and outliers, be alert to unlikely results, such as correlations that exceed the reliabilities of the variables involved). More than 40 years ago, Tukey (1977) described a number of strategies for using data visualization to make sense of the meaning of data and advocated careful and thoughtful examination of descriptive data before moving on to complex analyses. In a series of beautifully designed and illustrated books, Tufte (2001) provided essential guidance on the visual display of quantitative information. The effective examination of descriptive data is a critical step in bringing methodology from the abstract to the concrete by showing what actually happens when finely tuned methods come into contact with contexts, populations, and situations that might either enhance or limit the value of the data that are actually obtained.

Using Table 1 to evaluate hypothesis tests

Second, the descriptive statistics and their close derivatives are critical to the interpretation of virtually every claim that can be made in a research article. As analyses become more complex (with the resulting risk of mistakes in analysis and interpretation) and become more reliant on null hypothesis testing, the importance of making whatever demonstration that can be made based on simple descriptive statistics of the reasonableness of one's hypotheses is likely to increase.

Proposition: Any result that is established based on a complex data analysis that cannot be shown to be at least plausible based on the types of simple statistics shown in Table 1 (e.g., means, standard deviations, correlations) should be treated as suspect and interpreted with the utmost caution.

That is, hypothesized effects or relationships should first and foremost be illustrated based on simple correlations, mean differences, or graphic presentations of data in their simplest form possible. The idea of using descriptive statistics to at least demonstrate the tests is based on two assumptions. First, it is assumed that the data that are collected or analyzed by researchers are meaningful and relevant. To be sure, data sometimes require corrections or controls to be sensibly interpreted, but it is important to understand that the practices of statistical correction and control have been hotly debated and that corrections and controls can do as much to obscure the meaning of your data than they can to illuminate it (Becker, 2005; Berneth et al., 2018; LeBreton et al., 2014; Murphy & DeShon, 2000; Schmidt et al., 2000). Data may need to be combined or transformed before they can be sensibly interpreted; items must be grouped into scales, and sometimes data that are distributed in unusual ways may require transformation or rescaling. However, the first assumption behind the idea of feasibility tests is that the data should at some point be sufficiently meaningful so that direct and simple analyses of those data will shed at least *some* light on what the data mean.

The second assumption behind feasibility tests is that although complex analyses may lead to different, and sometimes more valid, conclusions than a simple examination of the data would suggest, the results of analyses that are several steps removed from the data should be treated with some skepticism if they plainly contradict the descriptive data. This does not necessarily mean that conclusions based on complex analyses are incorrect. Failure to demonstrate the reasonableness of

Table 2. Using Simple Statistics to Evaluate the Feasibility of Several Classes of Hypothesis Tests

---

1. Intervention studies
   a. Is $M_{\text{treatment}} > M_{\text{control}}$?
   b. Is d large enough to care about?
2. Moderator studies
   a. Is r_{xz} small?
   b. Are r_{xy} and r_{zy} small enough not to create a very large R^2 when X and Z are used to predict Y?
   c. Is $r_{xy} > r_{zy}$?
3. Mediator studies
   a. Are both r_{xm} and r_{my} larger than r_{xy}?
   b. Are r_{xm} and r_{my} closer in value to $\sqrt{r_{xy}}$ than to the value of r_{xy}?

---

one's hypotheses before moving on to a complex analysis does not necessarily diminish the contribution and value of a more complex analysis, but it does point out the potential for complex analyses to mislead the researcher, and it should place the onus on the researcher to argue why complex results that might not be easily understood based on an inspection of the descriptive data should nevertheless be believed.

Below, I describe three simple approaches to using descriptive statistics and simple data analysis tests that correspond with questions that are often asked in research in applied psychology and the organizational sciences: Did treatments or interventions work? Does a third variable (Z) moderate the relationship between two other variables or sets of variables (X and Y)? Does a does a third variable (M) mediate or explain the relationship between two other variables or sets of variables (X and Y)?

The core idea behind the approach suggested here is that authors should demonstrate, using the simplest descriptive statistics available and appropriate, that their core hypothesis is at least feasible before moving on to complex analyses that might be sufficiently removed from the data to allow for the possibility of serious misinterpretation. Three examples are summarized in Table 2 and are described below.

Intervention studies

If you propose that some treatment or intervention will lead to higher or more favorable scores on some main dependent variable, it is a good idea to first ask whether people who receive this treatment *do* in fact receive what appear to be higher scores. At the simplest level, this might involve nothing more than a comparison of the mean scores of people who received the treatment with mean scores of comparable people who did not; if mean scores in the treatment group are the same as or lower than those in the control group, the hypothesis that treatments make things better is not on the face of it feasible. In longitudinal studies, feasibility tests might involve something as simple as noting whether the posttest mean is greater than the pretest mean. If posttest scores are lower than pretest scores, the hypothesis that things improve over time if not feasible.

A slightly more sophisticated type of feasibility test is to calculate an effect-size measure. There is a large and sophisticated literature dealing with alternative effect-size measures and their meaning (e.g., Cumming, 2014; Kelly & Preacher, 2012; Valentine & Cooper, 2003), but at this stage simple measures may be all that are needed. Something as simple as a d statistic (i.e., the difference between treatment and control group means, divided by some sort of standard deviation estimate) will do nicely, as would any number of variants of statistics that express group differences in terms of the percentage of variance in the dependent variable that is explained (Murphy et al., 2014). The use of effect-size measures in research of this sort is widely advocated (e.g., the journal reporting

standards of the American Psychological Association have long called for presenting effect-size estimates along with the results of tests of statistical significance; Applebaum et. al., 2018), but many of the analytic methods used in the organizational sciences provide little if any information about the strength of effects. When these types of analysis are used, it is especially important to demonstrate that it is at least feasible that treatments have a nontrivial effect before proceeding to more complex analyses.

A simple examination of pretest versus posttest or treatment versus control group means would save authors from the embarrassment of concluding that their interventions are working when they are in fact making things worse, an error I have encountered multiple times in reviewing research that went straight to complex analyses without taking the descriptive statistics in Table 1 seriously. Even if mean scores in the treatment group *are* higher than are those in the control group, it can be difficult to interpret these differences without the sort of contextual information that effect-size measures include. It is surely better to have a general sense of whether treatment effects are relatively large or trivially small before moving on to complex analyses that may not provide this information. A demonstration that scores seem to get better after an intervention, and that this effect may be large enough to pay attention to, provides a quick and simple test of the feasibility of your hypotheses and justifies moving on to more complex analyses.

A variety of benchmarks have been proposed for describing effects or relationships as trivially small (Bosco et al., 2015; Murphy et al., 2014), and it may be difficult to set a hard and fast definition of trivially small effects; treatments that lead to small changes in a dependent variable may nevertheless be useful if the effects of those changes are practically important. Nevertheless, it is useful to have at least some evidence that there *might* be an effect worth caring about before proceeding to analyses that are so far removed from the data that they may not provide this information.

If your hypothesis is that treatments or interventions make things better, your results section should start with a brief review of the contents of Table 1 to determine whether scores do in fact improve with the intervention and whether these improvements are large enough in comparison with the overall variability of scores to care about. That is, before you do anything complicated with your data, you should first be required to show that the values of d or some related statistic are in the predicted direction and are large enough to be meaningful.

An example

Suppose your study is designed to evaluate a training intervention. Before launching into something as complex as a longitudinal latent variable analysis, where conclusions might be drawn based on coefficients several steps removed from the primary data you collected, it is very valuable to have a statement like "After receiving this training, performance improved substantially ($M_{pre} = 2.5$, $M_{post} = 3.7$), and this difference in performance was large enough ($d = .37$) to be potentially meaningful." This statement lays the groundwork for more complex analyses, which may be more informative but which also may not give a concrete indication of how much performance improved, and in some cases, whether performance is in fact better after training than before.

Moderator hypotheses

One of the most common hypotheses in research in the organizational sciences is that the relationship between two variables (or sets of variables) X and Y changes as a function of scores on a third variable, Z. Moderator effects and interactions are an important component of many models and theories (Aguinis, 2002; Aguinis et al., 2005; Latham & Pinder, 2005), and numerous papers have discussed methods for detecting and estimating interactions and moderators in primary research, meta-analysis, and meta-regression (e.g., Aguinis, 1995, 2002; Aguinis et al., 2001;

Aguinis & Pierce, 1998; Aguinis & Stone-Romero, 1997; Bobko & Russell, 1994; Gonzalez-Mulé & Aguinis, 2018; McClelland & Judd, 1993; Rogers, 2002). This literature also includes several papers describing best-practice recommendations for studies that involve moderation effects (e.g., Aguinis & Gottfredson, 2010; Aguinis et al., 2011; Carte & Russell, 2003; Edwards, 2009).

There are well-known reasons for skepticism about moderator hypotheses (Murphy & Russell, 2017). Moderator effects in the behavioral sciences and in management research are often extremely small,[4] and the statistical power of tests of moderators or interactions in organizational research is often correspondingly low (Aguinis, 1995; Aguinis al., 2005; Aguinis & Stone-Romero, 1997; Sackett et al., 1986; Shieh, 2009). Much of the research on the power of moderator tests here has been conducted using moderated multiple regression, but the problems of small effect sizes and low levels of power are equally serious when meta-analytic and multilevel methods are used for testing moderator hypotheses (Aguinis et al., 2013; Aguinis et al., 2011; Aguinis & Pierce, 1998; Mathieu et al., 2012).

Assume that reasonable measures and research designs have been employed and that there is a good theoretical reason to pursue moderator hypotheses. Is there anything authors can do, based on the simplest examination or analysis of the descriptive statistics typically found in Table 1, to demonstrate that moderator hypotheses are at least feasible? Murphy and Russell (2017) suggest some simple and concrete possibilities.

The most common method of testing for moderator effects involves comparing an additive main effect and a multiplicative model using ordinary least squares regression where

$$\hat{Y} = b_0 + b_1 X + b_2 Z, R^2_{additive};$$

$$\hat{Y} = b_0 + b_1 X + b_2 Z + b_3 XZ; R^2_{multiplicative};$$

and

$$\Delta R^2 = R^2_{multiplicative} - R^2_{additive.}$$

Rejection of the hypothesis that $b_3 = 0$, or that $\Delta R^2 = R^2_{multiplicative} - R^2_{additive} = 0$, is usually taken as evidence that a moderator effect is present.

In evaluating the feasibility that a moderator can be reliably detected in a study, four factors (N, the reliability of X and Z, the correlation between X and Z, and the strength of the linear effects of X and Z as predictors of Y) need to be considered (Murphy & Russell, 2017). Two of these, sample size and the reliability of the measures employed, are best thought of as indicators of the adequacy of the research methodology for testing moderator hypotheses. For example, sample size is one of the primary determinants of the power of most statistical tests (Murphy et al., 2014), and the likelihood that a moderator will be detected when N is small is quite low; Murphy and Russell (2017) note that if you are using moderated multiple regression to test for moderators, you are likely to need a sample of several thousand subjects to detect the types of moderator effects that are typically found in the organizational literature (Aguinis et al., 2005). Reliability also matters. Moderated multiple regression requires the formation of cross-product terms (i.e., $X \times Z$), and if *either* X or Z show even moderately low levels of reliability, the product of these two terms will be highly unreliable, reducing your ability to detect potential moderators.

In addition to these two methodological factors (i.e., N and reliability of X and Z), there are two aspects of the descriptive statistics presented in most studies that bear heavily on the feasibility of moderator hypotheses. First, as Murphy and Russell (2017) note, the likelihood of detecting a moderator becomes lower the stronger the linear relationship between the Y variable and the independent variable(s) and the moderator(s). This suggests that if you hope to detect a moderator effect, both r_{xy} *and* r_{yz} should be reasonably small. The rationale is simple; relatively large

---

[4]Aguinis et al (2005) suggest that 95% of all reported studies found that moderator effects accounted for less than 7/10 of 1% of the outcome variance.

correlations between Z and Y and between X and Y will imply correspondingly large multiple correlations between the combination of X and Z and Y.

Second, as Murphy and Russell (2017) also note, the likelihood of detecting a moderator decreases as the correlation between the independent variable(s) and the moderator(s) goes up. This suggests that if you hope to detect a moderator effect, r_{xz} should also be reasonably small. Once again, the rationale is simple: Relatively large correlations between X and Z will produce cross-product terms that are so highly correlated with X and Z that the likelihood that these cross products will capture unique information is low. If you wish to have a realistic hope of detecting a moderator effect, r_{xz} should be small (Murphy & Russell, 2017).

It is easy to say that r_{xy}, r_{xz}, and r_{yz} should be small if you want to have a reasonable chance of Z moderating the relationship between X and Y, but *how* small? There is probably no certain answer, but it is possible to lay out two considerations when evaluating these correlations. First, the feasibility of a moderator hypothesis is likely to depend more strongly on r_{xz} than on r_{xy} or r_{yz}. The correlation between main effects and cross products is highly sensitive to the value of r_{xz} (Cohen et al., 2002; Cortina, 1993), and unless the correlation between X and Z is small, the cross-product term that is used in testing the moderator hypotheses will necessarily be highly correlated with both X and Z and the likelihood of an incremental contribution will necessarily be low. There are certainly steps that can be taken to reduce this collinearity (e.g., centering X and Z; Iacobucci et al. [2016] discuss various ways in which centering influences the outcomes of moderated regression), but if X is even moderately correlated with Z, these steps will have only a limited effect. This suggests that it is most critical to show that r_{xz} is reasonably small before proceeding to test moderator hypotheses. Second, r_{xz} should probably be smaller than r_{xy}. The rationale here is more conceptual than statistical, if the goal is to predict Y. If Z is a better predictor of Y than X, this suggests that the authors may be barking up the wrong tree. Note that tests of moderator hypotheses via moderated multiple regression are indifferent to which variable gets the label X and which gets the label Z, and the finding that Z moderates r_{xy} also implies that X moderates r_{zy}. All in all, if we are going to relegate one of the variables (i.e., X or Z) to the role of moderator, a case can be made that it should be the one that does not contribute quite so much as an independent predict.

Putting together the considerations laid out above, I propose that authors who wish to pursue moderator tests should, in their discussion of Table 1, ask themselves whether r_{xz} can plausibly be described as small and whether $r_{xy} > r_{zy}$. As with the other tests discussed here, finding out that the answer is "No" does not automatically mean that moderator hypotheses are implausible. However, it does indicate that the data seem to lean strongly in the direction of rejecting the idea of moderation, and if a more complex analysis suggests that moderation has occurred, the onus should be on the author to argue why we should believe the more complex analysis rather than believing a simpler look at these same data.

An example
A study that proposes that individual differences in agreeableness (Z) moderate the relationship between a leader's willingness to engage in behaviors that show consideration (X; on the assumption that more agreeable leaders will be more effective and convincing in this regard than less agreeable ones) and leader effectiveness (Y) might note the following:

> In this study, $r_{xy} = .25$, $r_{xz} = .12$, and $r_{zy} = .14$. Based on the benchmarks that are suggested by Bosco et al. (2015), r_{xz} and r_{zy} represent small to moderately large correlations, whereas r_{xy} is approximately the median effect size for leadership studies. The correlations between the mediator and the independent variable are not so large as to create excessive collinearity with cross-product terms, and the combined validities of the independent variable and the mediator are not so large as to create statistical barriers to demonstrating a moderator effect.

Mediation hypotheses

The hypothesis that the effect of some variable X on an outcome Y is mediated by some third variable M is very common in psychology (MacKinnon et al., 2007) and management (Rosopa & Stone-Romero, 2008). For example, Wood et al. (2008) reviewed 25 years of research in the organizational sciences and summarized the results of over 400 papers (about 50% experimental or quasi-experimental studies) that tested for over 700 mediation effects. The 2014 issue of the *Academy of Management Journal* published 60 papers (excluding editorials); more than 25% of these explicitly proposed and tested mediator hypotheses.

Mediation models are often phrased in causal terms (i.e., X causes M, which in turn causes Y), at least in their formal structure, but they need not be. Tests for spurious correlation (e.g., the number of armed robberies in a city is correlated with the number of churches, but both of these are strongly influenced by the total city size) can be take the same statistical form as tests for mediation. There is a robust literature dealing with methods for estimating and testing mediation effects (e.g., Baron & Kenny, 1986; James et al., 2006; Judd & Kenny, 1981; MacKinnon et al., 2007; Maxwell et al., 2011), with the implementation these methods (Preacher & Hays, 2004), and with the difficulties in drawing the sorts of causal inferences that are often implicit in mediator models (e.g., Rosopa & Stone-Romero, 2008; Shrout & Bolger, 2002).

Mediation hypotheses are often built on a compelling theoretical rationale, and they are undoubtedly important in the behavioral sciences. However, the hypothesis that M fully mediates the relationship between X and Y is a demanding one, and mediation studies often fail, in the sense that they more often than not lead to the conclusion that M might *partially* mediate r_{xy} but cannot fully account for this relationship (James et al., 2006; Judd & Kenny, 1981; Maxwell et al., 2011; Wood, et al., 2008). The statistical power of the procedures for evaluating full versus partial mediation is often low (MacKinnon et al., 2002; Mallinckrodt et al., 2006; Maxwell et al., 2011), particularly when mediation tests are performed in small samples (Preacher & Kelly, 2011), and it is likely that in some papers that the failure to reject the hypothesis that the link between X and Y, once M is taken into account, is zero is a reflection of inadequate power rather than an indication of full mediation. Thinking about *why* tests of the hypothesis that M fully mediates r_{xy} so frequently fail helps in articulating the (surprisingly strict) conditions under which a full mediation hypothesis is plausible and in identifying a simple feasibility test.

There is an obvious and simple test for the feasibility of mediation hypotheses, but this test is often overlooked in the haze of complex mediation models. For a mediation hypothesis to be feasible, it is of foremost importance to demonstrate that both the proposed mediator and the independent variable are indeed correlated with the dependent variable. That is, if r_{xm} or r_{xy} are close to zero, the idea that M mediates r_{xy} is simply implausible. An inspection of the correlations among independent variables, mediators, and dependent variables allows you to go considerably farther than this simple test. A simple decomposition of the semipartial correlation between X and Y controlling for the relationship between M and Y shows that there are two other simple feasibility tests that can be performed by inspecting the table of correlations among independent variables, mediators, and dependent variables.

Semipartial correlation is not the most common or the most powerful method for testing mediation models, but it provides an *extremely* powerful conceptual tool for explaining how and why mediation hypotheses succeed or fail (Baron et al., 2000; MacKinnon, 2008; MacKinnon et al., 2007) and for articulating what it takes for a mediation model to fully succeed. In particular, this method allows you to specify the conditions under which the hypothesis that M mediates the relationship between X and Y is feasible.

If M mediates the relationship between X and Y, this implies that the X–Y relationship should tend toward zero once the effects of M on Y are controlled. That is, the semipartial correlation between X and Y controlling for the correlation between M and Y should approach zero if M truly mediates r_{xy}. The formula for this semipartial correlation, $r_{x(y.m)}$, is shown in Equation 1.

$$r_{x(y.m)} = \frac{r_{xy} - r_{xm}\,r_{my}}{\sqrt{1 - r_{my}^2}}. \tag{1}$$

The formula shown in Equation 1 illustrates quite clearly what it takes for the hypothesis that M mediates the relationship between X and Y to be feasible. As the formula shows, $r_{x(y.m)}$ will equal zero if and only if $r_{xy} = r_{xz} \times r_{zy}$. This can in turn happen only if (a) *both* r_{xz} and r_{zy} are considerably larger than r_{xy} or (b) one of these two correlations is smaller than r_{xy} and the other is *much* larger than r_{xy}.

For example, full mediation can be demonstrated whenever both r_{xm} and r_{my} are equal to $\sqrt{r_{xy}}$. In that case, r_{xy} will equal $r_{xm} \times r_{my}$. If either r_{xm} or r_{my} are smaller than $\sqrt{r_{xy}}$, the other one must increase proportionally for full mediation to work. For example, if $r_{xy} = .50$ (which means that $\sqrt{r_{xy}} = .707$) and $r_{xm} = .65$, r_{my} must be .77 for the value of $r_{x(y.m)}$ to drop to zero.

It can be argued that requiring the value of $r_{xm} \times r_{my}$ to be exactly equal to r_{xy} might pose an unduly stringent standard for demonstrating that the mediator effects are at least plausible. Other methods of analysis take more information into account, and it might be possible to demonstrate, using a more comprehensive method of analysis, that M fully mediates, or almost fully mediates, the relationship between X and Y. The analysis presented here does, however, suggest a very simple feasibility test for mediation hypotheses. The hypothesis that M mediates the relationship between X and Y is simply not feasible unless both r_{xm} and r_{my} are larger than r_{xy}, preferably closer in value to $\sqrt{r_{xy}}$ than to the value of r_{xy}.

An example
Suppose a study proposes that turnover intentions (M) mediate the relationship between burnout (X) and turnover decisions (Y). To demonstrate the feasibility of a mediator hypothesis, the authors might note, "The correlation between turnover intentions and turnover decisions is substantial ($r_{my} = .55$). Burnout is correlated with turnover decisions ($r_{xy} = .29$) and is even more correlated with turnover intentions ($r_{xy} = .48$). This pattern of correlations suggests that the hypothesis that turnover intentions mediate the relationship between burnout and turnout decisions is plausible; the semipartial correlation of burnout and turnover decisions controlling for turnover intentions is quite close to zero ($r_{x(y.m)} = .03$)".

Conclusions

Table 1 is the Cinderella of tables. It is often overlooked, but it can be of immense value. One key to making better use of Table 1 is to get over the strict distinction between Methods and Results. The statistics in Table 1 often tell about primary aspects of the sample and of the measures that were used. Careful attention to means, standard deviations, and intercorrelations tells you a great deal about whether your measures and sample are capturing the phenomena you are trying to study. Returning to an earlier example, suppose you are testing a hypothesis about how people respond when they are dissatisfied. Correlations between your satisfaction measure and other measures in your study are likely to shed light on the validity of these measures. Means and standard deviations will help you determine whether your study actually includes many people who *are* dissatisfied. If you shuffle Table 1 off to the Results section and proceed (as most studies seem to do) to ignore it, you might miss a lot of information that would be useful for understanding what your findings actually mean.

Table 1 often provides you will all of the information needed to make an initial assessment of the reasonableness of your hypotheses. In intervention studies, they can save you from getting everything backwards or from mistaking a trivial effect for an important one because it is "significant." In moderation and mediation studies, knowing ahead of time that your data make it very

unlikely that a moderation or mediation hypothesis can possibly be right would save us all from a lot of embarrassment. For example, if you know based on Table 1 that a proposed mediator is not related to the dependent variable in any meaningful way, it is unwise to pursue mediation hypotheses by using complex analytic tools whose end results may be hard to concretely interpret. On the other hand, a paper that starts with a simple demonstration that the proposed effects are large enough to care about or that the proposed process seems to make sense when you look at simple and easily understood statistics will make the results of subsequent analyses both more credible and more understandable. There are many tables in an article that might include important and useful information; Table 1 is almost always one of these.

Organizational research is moving in the direction of an increasing reliance on complex analyses that are frequently performed or interpreted incorrectly. Even when the analysis is done right and interpreted correctly, the likelihood that main results will depend entirely on the outcomes of a significance test is increasing. In almost all of the contexts in which organizational researchers use these tests as a primary criterion for evaluating study results, all that they are really learning is whether their sample was large enough to allow them to reject the null hypothesis. That is, these tests tell you more about the methods and the sample than they do about the substantive phenomenon being investigated.

Complex statistical methods not only open more possibilities for errors and misinterpretation but also make our research increasingly difficult for nonspecialists to understand. Explaining hypothesis tests and demonstrating their reasonableness in terms of the simplest possible statistics is likely to increase the understanding and applicability of our research in organizations and in the broader scientific community. The aircraft design team that created the legendary F-5 fighter jet had a motto that might be profitably applied to the results section of virtually every academic article in the organizational sciences—that is, "add simplicity and lightness." We don't even need the lightness; organizational scientists who can successfully add simplicity to their work are much more likely to have a meaningful influence than are their colleagues who write for a narrow audience of similarly trained specialists. Table 1 is a great place to find simplicity.

Over the last 35 years, I have reviewed or served as action editor for several thousand research articles, and my experience (which may not be a reliable indicator and might not be replicated by others) has led me to several conclusions about the state of science in organizational research. I firmly believe that when the primary findings of a study cannot be explained, or at least hinted at, based on the sort of simple statistics that are typically shown in Table 1, the likelihood that the work has been done and interpreted correctly *and* that it is sufficiently important to matter to a nonspecialist is very low. We should and can make much better use of simple statistics in arguing for our hypotheses, and a great place to start is Table 1. The current practice is to provide and then largely ignore Table 1. I suggest that we reverse our thinking. Table 1 is usually the most important table and often the *only* important table in manuscripts. If N is either very large or very small, the outcomes of significance tests are a foregone conclusion and the only useful tool for making sense of results will be the descriptive statistics presented in Table 1. Let's give this table the attention it deserves!

References

Aguinis, H. (1995). Statistical power problems with moderated multiple regression in management research. *Journal of Management*, **21**, 1141–1158.

Aguinis, H. (2002). Estimation of interaction effects in organizational studies. *Organizational Research Methods*, **5**, 207–211.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, **90**, 94–107.

Aguinis, H., Boik, R. J., & Pierce, C. A. (2001). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression. *Organizational Research Methods*, **4**, 291–323.

Aguinis, H., Edwards, J. R., & Bradley, K. J. (2017). Improving our understanding of moderation and mediation in strategic management research. *Organizational Research Methods*, **20**, 665–685.

Aguinis, H., & Gottfredson, R. K. (2010). Best-practice recommendations for estimating interaction effects using moderated multiple regression. *Journal of Organizational Behavior*, **31**, 776–786.

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, **39**, 1490–1528.

Aguinis, H., Gottfredson, R. K., & Wright, T. A. (2011). Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior*, **32**, 1033–1043.

Aguinis, H., & Pierce, C. A. (1998). Testing moderator variable hypotheses meta-analytically. *Journal of Management*, **24**, 577–592.

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology*, **82**, 192–206.

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, **567**, 305–307.

Applebaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force Report. *American Psychologist*, **73**, 3–25.

Baron J., Hershey J. C., & Kunreuther, H. (2000). Determinants of priority for risk reduction: The role of worry. *Risk Analysis*, **20**, 413–427.

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**, 1173–1182.

Becker, T. E. (2005). Potential problems in statistical control variables in organizational research: A qualitative analysis with recommendations. *Organizational Research Methods*, **8**, 274–289.

Bedeian, A. G. (2014). "More than meets the eye": A guide to interpreting the descriptive statistics and correlation matrices reported in management research. *Academy of Management Learning and Education*, **13**, 121–135.

Berneth, J., Cole, M. S., Taylor, E. C., & Walker, H. J. (2018). Control variables in leadership research: A qualitative and quantitative review. *Journal of Management*, **44**, 141–160.

Bobko, P., & Russell, C. J. (1994). On theory, statistics, and the search for interactions in the organizational sciences. *Journal of Management*, **20**, 193–200.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, **100**, 431–449.

Carte, T. A., & Russell, C. J. (2003). In pursuit of moderation: Nine common errors and their solution. *MIS Quarterly*, **27**, 1–22.

Caetano, A., & Santos, S. C. (2017). The gap between research and professional practice in work and organizational psychology: Tensions, beliefs, and options. In E. Nelva, C. Torres & H. Mendonça (Eds.), *Organizational psychology and evidence-based management* (pp. 1–22). Springer International.

Cohen. J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen. J. (1994). The earth is round ($p <. 05$). *American Psychologist*, **49**, 997–1003.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge.

Cortina, J. M. (1993). Interaction, nonlinearity, and multicollinearity: Implications for multiple regression. *Journal of Management*, **19**, 915–922.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, **2**, 161–173.

Cortina, J. M., Green, J. P., Keeler, K. R., & Vandenberg, R. J. (2017). Degrees of freedom in SEM: Are we testing the models that we claim to test. *Organizational Research Methods*, **20**, 350–378.

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

Desai, V. M. (2018). Third-party certifications as an organizational performance liability. *Journal of Management*, **44**, 3096–3132.

DeSimone, J. A., Köhler, T., & Schoen, J. L. (2019). If only it were that easy: The use of meta-analytic research by organizational scholars. *Organizational Research Methods*, **22**, 867–891.

Edwards, J. R. (2009). Seven deadly myths of testing moderation in organizational research. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences* (pp. 143–164). Routledge.

Giennik, M. N., Zacher, H., & Wang, M. (2018). Age in the entrepreneurial process: The role of future time perspective and prior entrepreneurial experience. *Journal of Applied Psychology*, **103**, 1067–1085.

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**, 2246–2273.

Green, J. P., Tonidandel, S., & Cortina, J. M. (2016). Getting through the gate: Statistical and methodological issues raised in the reviewing process. *Organizational Research Methods*, **19**, 402–432.

Iacobucci, D., Schneider, M. J., Popovich, D. L., & Bakamitsos, G. A. (2016). Mean centering helps alleviate "micro" but not "macro" multicollinearity. *Behavior Research Methods*, **48**, 1308–1317.

James, L. R., Mulaik, S. A., & Brett, J. M. (2006). A tale of two methods. *Organizational Research Methods*, **9**, 233–244.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, **5**, 602–619.

Kelly, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, **17**, 137–152.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects?* Sage.

Latham, G. P., & Pinder, C. C. (2005). Work motivation theory and research at the dawn of the twenty-first century. *Annual Review of Psychology*, **56**, 485–516.

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, **7**, 478–500.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Erlbaum.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, **58**, 593–614.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, **7**, 83–104.

Mallinckrodt, B., Abraham, W. T., Wei, M., & Russell, D. W. (2006). Advances in testing the statistical significance of mediation effects. *Journal of Counseling Psychology*, **53**, 372–378.

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen. G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, **97**, 951–966.

Maxwell, S. E., Cole, D. A., & Mitchell, M. A. (2011). Bias in cross-sectional analyses of longitudinal mediation: Partial and complete mediation under an autoregressive model. *Multivariate Behavioral Research*, **45**, 816–845.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, **59**, 537–563.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, **114**, 376–390.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of psychology. *Journal of Consulting and Clinical Psychology*, **46**, 806–834.

Murphy, K. R., & DeShon, R. (2000). Inter-rater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, **53**, 873–900.

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (4th ed.). Taylor & Francis.

Murphy, K. R., & Russell, C. J. (2017). Mend it or end it: Redirecting the search for interactions in the organizational sciences. *Organizational Research Methods*, **20**, 549–573.

Preacher, K. J., & Hays, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in mediation models. *Behavior Research Methods, Instruments and Computers*, **36**, 717–731.

Preacher, K. J., & Kelly, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, **16**, 93–115.

Rogers, W. M. (2002). Theoretical and mathematical constraints of interactive regression models. *Organizational Research Methods*, **5**, 221–230.

Roposa, P. J., & Stone-Romero, E. F. (2008). Problems with detecting assumed mediation using the hierarchical multiple regression strategy. *Human Resource Management Review*, **18**, 294–310.

Rynes, S. (2009). The research-practice gap in industrial-organizational psychology and related fields: Challenges and potential solutions. In S. Kozlowski (Ed.), *Oxford handbook of industrial-organizational psychology*. Oxford University Press.

Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderators in meta-analysis of correlational data: A Monte-Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology*, **71**, 302–310.

Schmidt, F L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, **1**, 115–129.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, **53**, 901–912.

Shieh, G. (2009). Detecting interaction effects in moderated multiple regression with continuous variables: Power and sample size considerations. *Organizational Research Methods*, **12**, 510–528.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, **7**, 422–445.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Valentine, J. C., & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes.* What Works Clearinghouse.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *American Statistician*, **70**(2), 129–133.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "*p* < 0.05. *American Statistician*, **73**(supp. 1), 1–19.

Wilkinson. L., & **Task Force on Statistical Inference.** (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, **54**, 594–604.

Wolfson, M. A., Tannenbaum, S. I., Mathieu, J. E., & **Maynard, M. T.** (2018). A cross-level investigation of informal field-based learning and performance improvements. *Journal of Applied Psychology*, **103**, 14–36.

Wood, R. E., Goodman, J. S., & **Beckman, N.** (2008). Mediation testing in management research: A review and proposals. *Organizational Research Methods*, **11**, 270–295.