The Eight Steps of Data Analysis: A Graphical Framework to Promote Sound Statistical Analysis

Dustin Fife[1]

[1] Rowan University

Abstract

Data analysis is a risky endeavor, particularly among those unaware of its dangers. In the words of Cook and Campbell (1976; see also Shadish, Cook, & Campbell, 2002), "Statistical Conclusions Validity" threatens all research subjected to the dark arts of statistical magic. Although traditional statistics classes may advise against certain practices (e.g., multiple comparisons, small sample sizes, violating normality), they may fail to cover others (e.g., outlier detection and violating linearity). More common, perhaps, is that researchers may fail to remember them. In this paper, rather than rehashing old warnings and diatribes against this practice or that, I instead advocate a general statistical analysis strategy. This graphically-based eight step strategy promises to resolve the majority of statistical traps researchers may fall in without having to remember large lists of problematic statistical practices. These steps will assist in preventing both false positives and negatives and yield critical insights about the data that would have otherwise been missed. I conclude with an applied example that shows how the eight steps reveal interesting insights that would not be detected with standard statistical practices.

Keywords: statistical assumptions, NHST, confirmatory data analysis, graphical data analysis, fishing, *p*-hacking

The field of psychology has been forced to participate in methodological introspection, of sorts. This introspection began late in the 20th century as methodologists vehemently protested the knee-jerk focus on *p*-values Null Hypothesis Significance Testing (NHST) encourages (Cohen, 1994; Harlow, Mulaik, & Steiger, 2016; Jones, 1952). The American

I wish to thank those who assisted in reviewing this manuscript, including Tom Dinzeo, Polly Tremoulet, Jeffrey Greeson, Yoav Zeevi, Christine Simmons, and Joseph Rodgers, as well as the anonymous reviewers and editor.

Correspondence concerning this article should be addressed to Dustin Fife, 201 Mullica Hill Road Glassboro, NJ 08028. E-mail: fife.dustin@gmail.com

Psychological Association (APA) suggested several alternatives, including a stronger focus on estimation (i.e., identifying the strength of the effect, as well as the size/direction of the parameters of interest; Wilkinson & Task Force on Statistical Inference, 1999), rather than the probability of the data in the face of no effect (a strange hypothesis indeed!).

This forced introspection not only continues today, but the replication crisis has heightened its necessity. A recent attempt to replicate 100 different studies (from three top journals in psychology) resulted in poor replication metrics (Open Science Collaboration, 2015); estimates of effect sizes were half as strong in the replications than in the original studies and 61% of the attempted replications failed to produce statistical significance. (For an alternative perspective on the replication crisis, see Shrout & Rodgers, 2018 and Maxwell, Lau, & Howard, 2015). Some have suggested this replication crisis was caused (at least partially) by researchers' over-reliance on NHST (and other statistical practices; Cumming, 2014; Pashler & Wagenmakers, 2012). In this paper, however, I will not rant and rave about NHST; Others have already highlighted its problems (Cohen, 1994; Cumming, Fidler, & Thomason, 2001; Schmidt, 1996; Trafimow, 2017). Rather, my purpose is primarily to bridge the gap between known best practices in data analysis and actual statistical application. In so doing, I hope to provide a step-by-step strategy targeted at applied researchers for developing a deeper understanding of one's data.

A Potential Misunderstanding

Arguably, what spawned the first methodological "crisis" in the 1990s was Jacob Cohen's thorough and acerbic critique of significance testing (Cohen, 1994). Near the conclusion of his article, he said, "don't look for a magic alternative to NHST, some other objective mechanical ritual to replace it. It doesn't exist" (p. 1001). I wonder if this statement is too easy to misunderstand. I believe what Cohen was trying to suggest was there is no *one* alternative to NHST; rather statistical analysis requires a rather large toolbox, where each tool is adapted to the circumstances under which it is most appropriate. The tool might be, for example, Bayes factor, confidence intervals, effect sizes, single-subjects designs, preregistration, and/or graphical data analysis. I fear that when some read Cohen's words, they interpret it to mean "NHST is wrong and there is nothing else researchers can use."

This is a rather discouraging interpretation, yet this interpretation is easy to make. In the past, methodologists have been quick to critique, but much slower to offer alternative recommendations. What exacerbates the problem is the fact that few agree about what ought to replace NHST (e.g., Cumming, 2014; Kruschke & Liddell, 2018). Again, the reason clear alternatives have failed to emerge is likely because the alternative is, more than likely, a toolbox rather than a specific tool. I am inclined to think few methodologists believe any *one* alternative is ideally suited to all circumstances, and, as such, have shied away from offering step-by-step instructions for best-practices because it feels too mechanical. On the other hand, applied researchers generally need some sort of structure. What I will introduce provides the structure without the mechanics. My framework is designed to train a researcher to look for cues in data that the toolbox needs to be adjusted. While NHST seems to allow one to follow a sequence of steps and arrive at an unambiguous conclusion, my approach allows one to follow a sequence of steps designed to gather evidence in order

to arrive at a conclusion that may be ambiguous. NHST permits one to "turn off" their brain. My approach re-activates it.

Although the approach and many of the graphical tools I present are new, most of the suggestions are not; for decades statisticians have advocated for non-mechanistic judgments, assessing model assumptions, graphical data analysis, and sensitivity analyses.[1] Despite this fact, practices have changed little over the years. Researchers rarely evaluate statistical assumptions (Hoekstra, Kiers, & Johnson, 2012) or plot their data (Healy & Moody, 2014; Levine, 2018), and researchers still rely heavily on significance testing (Counsell & Harlow, 2017; Coyne, 2016). Perhaps part of the problem is that these suggestions are relatively de-centralized. In this paper, I offer a centralized set of guidelines that have been cobbled from various sources that synthesize decades of suggestions into one unified step-by-step framework. This framework will not only protect against false conclusions, but will also free researcher's minds from rigid NHST thinking that is endemic in psychology.

In the following sections, I first review several reasons NHST practices are pervasive in psychology, then discuss potential causes for the replication crisis. Next, I review how the eight steps of data analysis encourages a greater focus on estimation and "listening" to one's data. Finally, I conclude with an example where I show how the eight steps deepened understanding of data.

Should Psychology Abandon p-values?

For decades, some methodologists have suggested significance tests ought to have no place in psychological journals (Cohen, 1994; Harlow et al., 2016; Schmidt, 1996; Valentine, Aloe, & Lau, 2015). Yet there seems to be little evidence researchers aren't using p-values to make decisions, nor does there seem to be much of a visible shift in statistical practices (Counsell & Harlow, 2017; Coyne, 2016; Cumming et al., 2007). Despite passionate and cogent arguments against NHST, several obstacles remain and *will* remain, no matter how red-faced methodologists get. These include:

- **Social pressures**. The entire field of psychology understands p-value speak and a researcher may decide not to venture outside NHST practices for fear of having an otherwise publishable paper relegated to the file drawer.
- **Habit**. For many researchers, they have been doing NHST for decades. For these people, shifting away from such rote practices is counter-intuitive (and difficult).
- **Learning**. Abandoning p-values in favor of some other statistical practice may require considerable time and effort that most researchers do not have.
- **p-values reduce ambiguity**. Without p-values, it would open the field to disagreement about what constitutes a scientifically significant finding. A rigid cutoff of 0.05 acts as an operational definition for a relationship that ought to be noticed (and published).

[1] Anecdotally, I've shared this manuscript with both statisticians and applied researchers. The statisticians tend to say, "Of course that's how data analysis ought to be done. We've been advocating for that for years!" On the other hand, applied researchers say, "I never thought to do data analysis this way."

- **p-values provide a "translational mechanism" from theory to data**. As argued by Cortina and Landis (2011), NHST bridges theoretical language into data analysis, and back again into theoretical language. This translation, they argue, is ill-defined, at best, and non-existent, at worst, in other statistical frameworks.

The guidelines I introduce help side-step (and sometimes address head-on) all of these concerns. No researcher needs to learn additional statistical techniques (except when the eight steps reveal that traditional methods fail to model the data), or remove p-values from their method section . Rather, I advocate a simple shift in focus that will add richness to one's statistical practices. Doing so will, over time, shift the culture away from p-values and toward a greater focus on estimation and attending to messages our data have long been trying to tell us.

Potential Causes of Replication Crisis

To diagnose the cause of the replication crisis, it is advantageous to think of the current predicament as a systematic, discipline-wide habit of committing Type I errors. Some authors (e.g.; Rothman, 2010) have suggested abandoning statistical NHST will fix all Type I (and Type II) errors. This may be technically correct, but eliminating NHST will not necessarily alter the number of false positives/negatives when researchers utilize other metrics for decision-making. Regardless of whether p-values or some other metric are used, the sources of false positives (and false negatives) are the same. Shadish, Cook, and Campbell (2002), noted two characteristics of data in particular that may inflate Type I error rates (and/or false positives): (1) Violated assumptions of statistical tests (e.g., normality, homogeneity, linearity, independence), and (2) Fishing/multiple testing. Various authors have commented on both of these issues and I will briefly review each in turn.

Violated Assumptions of Statistical Tests

Linear models (e.g., regression, ANOVA, t-tests, structural equation models) are the most common models in psychology. For these models to behave appropriately (i.e., for p-values to actually reflect the probability of a Type I error under the null), the data must meet four key assumptions: independence[2], normality, homoskedasticity, and linearity.[3] When decisions of significance are determined based on p-values, some of these assumptions are more critical than others (e.g., if normality is violated, the probability of committing a Type I error under a statistical significance decision criteria, given the null, tend to stay close to 0.05; while violations of independence will lead to significant departures from 0.05). When violated, Type I error rates may remain fairly close 0.05, or may deviate substantially in either direction. Likewise, when violated, Type II errors may also be inflated.

[2]Independence is a serious assumption that, when violated, will result in substantial bias in estimating standard errors. However, independence is more of a design issue than a characteristic of the data. Consequently, I will not address how to evaluate independence.

[3]Linearity is actually not an assumption of ANOVAs/t-tests. Or, rather, linearity is an assumption, though it is guaranteed to be met with categorical predictors.

The previous paragraph is how most textbooks write of statistical assumptions. I tend to think of them differently, particularly since I rarely use p-values for decision-making. If these assumptions are violated, it means the researcher has simply chosen the wrong statistical model, a condition easily fixed by choosing another. Yes, p-values will not remain at 0.05 if assumptions are violated, but it also means the researcher has simply chosen a model that is not appropriate. It would be like choosing to compute the mean on highly skewed data; one can do it but the information gleaned may be misleading. If the wrong model is chosen, one might have a false positive or negative.

The sensitivity of linear models to these assumptions have been well documented (e.g., Maxwell & Delaney, 2004; Osborne, 2013) as well as the consequences of violating these assumptions (Micceri, 1989). Yet rarely do researchers mention whether they checked for the appropriateness of linear models (Hoekstra et al., 2012). And because most researchers do not provide the datasets used for analysis[4], it is unknown the degree to which psychological research has been corrupted by violated assumptions.

The solution to the problem, as I mention in detail later, is a greater focus on estimation and graphical data analysis. Graphics allow the researcher to determine at a glance whether assumptions have been violated.

Although violated assumptions may inflate false positives, I suspect a far more common cause is multiple testing. I will address this issue in the following section.

Fishing/Multiple Testing/p-hacking

Most researchers are likely familiar with the problem of multiple testing: when there are four groups, for example, it would be unwise to perform a t-test comparing each and every group (group 1 vs. group 2, group 2 vs group 3, etc.). Likewise, it would be unwise to compute dozens of Cohen's d values and only interpret those that are larger than 0.5. Though the probability of one test being significant under the null is 0.05 (or the probability of one large d is small), the probability of rejecting one among several is much higher (much like the probability of rolling at least one six over the course of 10 rolls is much higher than the probability of getting it on the first roll). Indeed, this problem is sufficiently well-known that if any researcher were to submit a paper and report they performed 107 t-tests, the paper would likely be rejected.

Yet multiple testing likely happens all the time in psychology, but in more nuanced ways (Simmons, Nelson, & Simonsohn, 2011). Suppose, for example, a researcher collected 10 covariates that could potentially muck up the relationship between the IV and the DV. Said researcher might include a covariate, then run the analysis. If the treatment effect is non-significant, the researcher may decide another covariate is more appropriate to use as a control. This practice may continue until one of the covariates yields statistical significance on the treatment effect. This is another form of multiple testing. Likewise, if the researcher measured several dependent variables, then performed statistical analysis on each DV until

[4]This statement is based on extensive (and frustrating) personal experience.

significance was achieved, this too constitutes multiple testing, yet it is not so explicitly and universally condemned as running multiple t-tests.

The term that encapsulates the new ways in which some researchers participate in multiple testing is called "p-hacking," and p-hacking includes not only the practices I have mentioned (measuring multiple DVs and covariates and running several models until significance is obtained), but also several others, including adding more observations until significance is reached, and dropping an experimental condition. With a discipline so focused on p-values, it is simple to see why so many researchers exercise "researcher degrees of freedom" (knowingly or unknowingly) to achieve the "gold standard" of $p < 0.05$.

Various authors (e.g., Nelson, Simmons, & Simonsohn, 2018; Wagenmakers, Wetzels, Borsboom, Maas, & Kievit, 2012) suggest researchers voluntarily report their planned statistical analysis *a priori* (e.g., through the "as predicted" platform: https://aspredicted.org/), then be explicit in the paper about any modifications to the original plan. This is a great idea, when possible. However, it is extremely rare that researchers are prepared to make such detailed plans in advance; Also preregistration won't address the statistical issues mentioned previously[5] (i.e, violating of statistical assumptions) or invite researchers to consider the uncertainty associated with statistical analysis. Few researchers are ready to propose such risky hypotheses; few analyses go exactly as planned and preregistration best works when modifications to the original plan are unlikely. In these situations, preregistration is well-equipped to prevent p-hacking. Indeed, that is exactly what it was designed to do. It was not designed, however, to deepen one's understanding of data. One could, presumably, preregister a hypothesis, test the hypothesis, and replicate it, and still be misled if that person is not attending to the messages the data are trying to voice. The eight steps of data analysis I propose, on the other hand, were designed to give voice to data. These steps were not designed to prevent p-hacking. In other words, neither the eight steps nor preregistration alone will fix the replication crisis. Together, however, they will shift the focus away from simply publishing a (potentially spurious) finding toward building sound scientific and mathematical models of reality.

Type II Errors and the File-Drawer Problem

The replication crisis highlights the fact that psychology may be inundated with false positives. Unfortunately, it is more difficult to estimate the prevalence of false negatives. Researchers may spend months collecting data only to have a p-value not reach statistical significance. Some may abandon the project, while others might participate in p-hacking until significance is achieved.

The framework I propose will alleviate the problem of both false positives and negatives. For example, a pattern may not reach statistical significance for several reasons that would be detected under this framework, including outliers that pull means to be more similar, strong

[5] Preregistration will certainly invite authors to consider how they will handle violations of assumptions (Wicherts et al., 2016), but preregistration alone will not tell researchers how to detect and handle such violations. The eight steps will guide researchers making these decisions. As such, the two (preregistration and the eight steps) work hand in hand.

non-linear patterns that are poorly represented by a straight line, and violated statistical assumptions that render traditional tests overly conservative. In either case, whether we publish spurious findings or abandon non-significant results, spending more time with our data via the eight steps of data analysis will aid in shifting our focus to what the data are actually trying to tell us.

Guiding Principles of Data Analysis

Before I explain the eight steps, let me first introduce three guiding principles of data analysis:

1. *Plot raw data whenever possible.* One can easily be deceived when a graphic displays only summaries of the data (e.g., means as dots and standard errors as lines). For example, suppose Figure 1 came from an experiment where subjects were randomly assigned to watch a neutral video or a violent video. Further, suppose the subjects were subsequently measured on aggression. If one were to simply interpret the left plot, they might believe the type of video had a large effect on aggression. Yet when we overlay the "jittered"[6] raw datapoints (right panel), we see that the aggression scores are bimodal in the violent group. Perhaps that bimodality is caused by gender (e.g., males report higher aggression after watching the video, while females do not). This would be an important discovery that would be masked if one simply graphed the summaries rather than the raw data. Furthermore, the right panel shows the mean for the violent video group is quite misleading; the mean falls at a place where data are quite sparse.

2. *Utilize sensitivity analyses whenever needed.* Often times our data throw us curve balls that require making an ad hoc decision. For example, our data may require a transformation to render the residuals more normal, an outlier may require deleting, or a missing value may require imputing. One would hope the conclusions gleaned from data remain largely unaffected by whatever decision we make. The only way to determine whether our results are sensitive to our decisions is to run the analysis both ways. Others have called this a "multiverse" analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). For example, suppose a researcher decides an outlier ought to be deleted. The researcher should then run the analysis both with and without the outlier deleted and determine the degree to which the results change. The researcher might also investigate other strategies for dealing with the outlier (e.g., treating it as missing then imputing that value). Under this situation, I strongly recommend the researcher report the results of all three analyses (at least in a supplemental section) and comment on whether the results are sensitive to the decision made.

   Note that this sensitivity analysis is comparing two models that test the *same* hypothesis, rather than two models that test *different* hypotheses (i.e., a model comparison;

[6] Jittering means to add random noise to a categorical variable (in this case, Neutral and Violent, which may be coded as 1 and 2). Jittering categorical values prevents overlap of datapoints and makes it easier to see the distribution of the datapoints. See Fife (2019a) for examples of how to produce jittered plots in the point-and-click software called Jamovi.

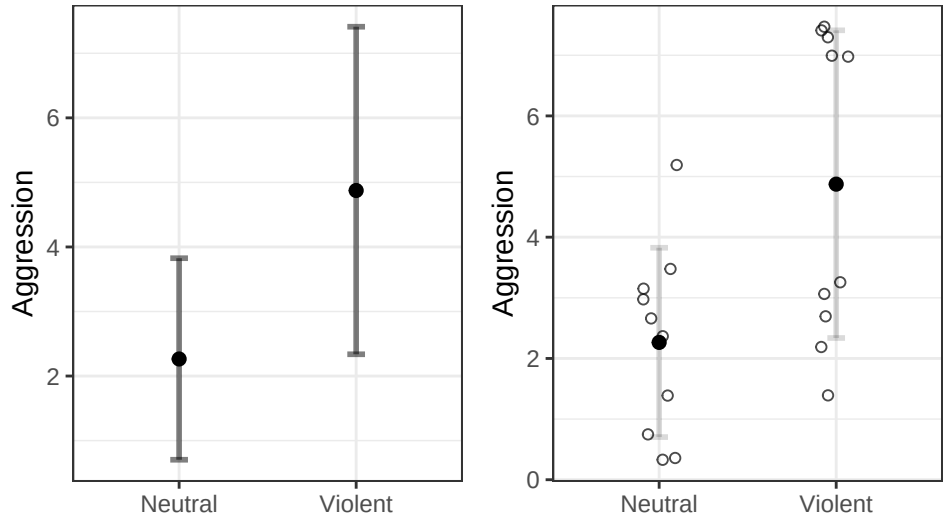

Figure 1. Two plots of the same data. In the left panel, only means and standard deviations are reported. In the right panel, the raw data points overlay the standard deviation bars. When possible, graphical displays of raw data are always preferred over graphical summaries.

Rodgers, 2010). Model comparisons are excellent tools for teasing out competing explanations of the data, but these sorts of model comparisons are outside the scope of this paper.

3. *Explicitly state where analyses fall on the exploratory/confirmatory continuum.* Exploratory data analysis (EDA) is the process whereby researchers analyze data without preconceived notions of what patterns they might find. While confirmatory data analysis (CDA) has been likened to placing a hypothesis on trial, EDA is like detective-work; data are searched for interesting patterns that might be worth pursuing for future investigation (including preregistered replications). Researchers might engage in either activity, or anything in between, such as "rough" CDA (Fife & Rodgers, 2019). Unfortunately, researchers have blamed EDA for the replication crisis . This is unfortunate and comes from misunderstanding the role, purpose, and tools associated with EDA. Prior to the replication crisis, researchers were *not* doing EDA. Rather, they had exploratory intentions that utilized confirmatory tools. Additionally, results discovered through exploration were presented as if they were confirmatory, which is a critical violation of one of EDA's most important rules: users of EDA must be explicit about which results were obtained through exploration versus confirmation. For a more thorough and complete treatment of EDA, CDA, and everything in between, see Fife and Rodgers (2019).

   These eight steps are designed to assist researchers participating in confirmatory or "rough" confirmatory research (Fife & Rodgers, 2019), or even exploratory data analysis. In confirmatory studies, researchers have a preconceived hypothesis for which

> they seek to evaluate the evidence. However, along the way of evaluating the evidence, data analysis plans often do not pan out as intended; interesting patterns present themselves which a researcher may want to investigate, unexpected complications arise that necessitate modifying the data analysis plan. In alignment with the rules and ethics of EDA, it is necessary that researchers be explicit about which analyses were uncovered through exploratory analysis, and which were obtained through confirmatory analysis. One might choose to blend the two into one paper, provided the researcher is again explicit about which results were confirmatory and which were exploratory. Returning to Figure 1, for example, the researcher may have decided before collecting data that the video would make participants more aggressive, yet did not anticipate the bimodality of scores in the treatment group. In the report, the researcher might say, "Upon graphing the results, a bimodal distribution was discovered among the participants in the treatment group. These results were unanticipated *a priori* so we decided to explore this relationship further. . . "

As mentioned previously, the eight steps can be used to guide confirmatory, rough confirmatory, or even exploratory data analysis. In the following section I will outline the eight steps and hope to illustrate how they might be used to direct the analyst's focus in such a way that improves understanding of data.

The Eight Steps to Data Analysis

In this section I will describe each of the eight steps. For simplicity, I have included a table (Table 1) that shows the eight steps and the function each step serves. For each of the steps that follow, I will address why I recommend performing said operation and what weaknesses it aims to overcome.

When evaluating the recommendations in Table 1, the reader might be inclined to say, "Why *these* steps instead of others?" For example, a Bayesian might insist that Bayes factors should accompany all analyses. Another might complain there's no mention of missing data or correcting for unreliability. Still others might say these steps fail to include preregistration.

To this question, I would offer a few reasons why I am advocating for these steps rather than others. First, these eights steps are not about tools/techniques, but about the approach one takes toward data analysis. This approach is designed to direct the analyst's attention toward the evidence in favor (or against) the chosen hypothesis. Additionally, these steps are fairly universal, regardless of what type of data one chooses to analyze. Techniques, on the other hand, are situation-specific. Bayes factors are excellent for deciding between two hypotheses, but not so excellent when one is concerned about estimation. Increasing reliability is necessary when one has unreliable measures, but unnecessary when one has no measurement error. Preregistration is critical for strong confirmatory studies, but unnecessary when one is doing exploratory research. The eight steps, on the other hand, are always appropriate and are critical in helping the analyst decide which tools are appropriate for a given situation. For example, Bayes factors could be used for making a

Table 1

A Summary of the eight Steps of Data Analysis

Step	Purpose
1. State the theoretical hypothesis	Helps to minimize "fishing" for statistical significance Provides a translational map from theory to data Allows users to specify their own decision criteria Invites researchers to consider preregistering hypotheses
2. Assess psychometric properties of variables	Invites researchers to think about the impact of measurement
3. Plot univariate distributions	Helps identify outliers Helps identify issues with non-normality Assists in identifying coding errors
4. Plot a graphic to match the theoretical hypothesis	Directs focus toward the size of effects Helps identify potential problems with non-linearity/heteroskedasticity Improves cognitive encoding of results Highlights uncertainty
5. Study residuals	Helps identify problems with normality (e.g., through histogram of residuals) Helps identify problems with non-linearity/homoskedasticity (e.g., through a residual dependence or SL plot)
6. Interpret parameter estimates/effect sizes	Encourages the researcher to focus on estimation before significance Put graphical information into concrete numbers
7. Set a decision criteria (if appropriate)	Assists in making a decision about significance
8. Replicate on a new dataset	Encourages cumulative and reproducible science

decision criteria that is also pre-registered (Step 1). Likewise, unreliability and missing data will reveal themselves in Steps 2/3.

A second reason I advocate for these steps is because they are relatively uncontroversial. There's a great deal of disagreement about the merits of Bayesian estimation (Simonsohn, 2014), confidence intervals (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016), preregistration (Szollosi et al., 2019), etc. Too often, I think methodologists are quick to argue over differences between approaches and give less air time to similarities. No methodologists that I know of would argue that graphics are bad, evaluating model assumptions is problematic, or interpreting parameter estimates is dangerous. Rather, they argue this is obvious, yet people are rarely doing these things; to me, that seems to indicate that maybe we ought to spend less time arguing and more time going back to the basics.

Third, I believe these steps are beneficial because they emphasize graphics. The visual processing system has enormous bandwidth and is able to encode large amounts of information with little effort (Bonneau et al., 2014). Graphics have a way of simultaneously highlighting uncertainty while also revealing problems with statistical models. Additionally, humans are really good at detecting visual patterns. If it is something we are good at, should we not be leveraging that when doing scientific research?

Finally, I advocate for these eight steps because I find them useful. I am not so dogmatic as to think these eight steps are *the* eight steps of data analysis. One might conjure a different set of eight, or ten, or twenty. Each step seems to provide a different "view" of the research question, each contributing a unique piece of evidence for the final evaluation. When I use them, and when I teach my students to use them, we gain a lot of information about the data. Additionally, when I have, through laziness, bypassed a step, I have been guilty of making short-sighted conclusions.

1. State the Theoretical Hypothesis and (Optionally) State a Decision Criteria

The first step to data analysis ought to be to state the theoretical hypothesis. Ideally, this would take place long before the researcher actually collects data (i.e., the hypothesis is preregistered). Doing so in advance will assist in preventing researchers from "bending" their original hypothesis to fit the actual analysis performed (e.g., "Oh yeah. I, uh, meant to include that as a covariate originally. I just forgot"). This step is entirely voluntary and preregistration strengthens the weight of the evidence in favor of the hypothesis. (It is, after all, much less impressive to "predict" an effect you already discovered than to predict an effect before you actually collect data).

If the analyst does decide a decision is needed, I recommend three strategies that will increase the utility of their decision criteria: (1) mapping hypotheses to specific statistical parameters, (2) stating strong hypotheses, and (3) developing decision criteria for clinical "significance."

Mapping hypotheses to parameters. In an ideal world, one's statistical hypothesis is closely tied to the theoretical hypothesis, such that support for the statistical hypothesis provides support for the theoretical hypothesis. However, there is always some simplification

that occurs when one translates from theory to statistical inference. This is inevitable, yet it is critical to ensure that the two (theory and model) are as closely aligned as possible. To do so, the theoretically-derived hypotheses ought to be tied to specific parameters in a model. Too often researchers write well-crafted introduction sections, providing strong theoretical rationale for their chosen verbal hypothesis, yet, there is a disconnect between the well-crafted introduction and the results section; the hypothesis points to a particular parameter (e.g., the interaction term in a model or the main effect of a predictor after controlling for another), yet the results report gobs of results and corresponding tests of significance. On other occasions, the statistical hypothesis is very weakly tied to the theoretical hypothesis. Not only does this dilute the message (because the parameter of interest is buried between other tests), but this constitutes fishing. Each reported p-value is the result of a tested hypothesis and, as such, each reported p-value ought to be clearly supported by strong theoretical rationale. If the introduction section only develops arguments for testing one parameter, then only one parameter ought to be interpreted (though all parameters ought to be reported).

Granted, some analyses require entering other parameters in the model. For example, if the researcher's hypothesis concerns an interaction effect between two independent variables, the main effects must be included in the model as well. However, these main effects need not be tested (or rather, interpreted since software packages tend to report significance for all parameters) for significance because, again, the researcher's hypothesis is not concerned with these parameters.

Stating Strong Hypotheses. Years ago, Meehl (1967) criticized the use of zero as a tested hypothesis. That's a rather low bar to pass. Instead, he advocated for "strong" hypotheses, where researchers specify numeric values for the hypothesized parameter. For example, rather than testing whether a correlation is different from zero, a researcher can test whether the correlation is different from +0.4. This amounts to reversing the role of the null and the alternative and can lead to some logistic problems (e.g., researchers might be inclined to collect small samples so they don't have power to reject their cherished hypotheses). With some modification, we might instead hypothesize the parameter of interest falls within a particular range (e.g., from $r = 0.2$ to 0.4), or at least that its direction is positive (or negative). Better yet, researchers might use the values of the parameters themselves to set their own decision criteria, which I will discuss next. For a more detailed treatment about various approaches for developing precise hypotheses see Edwards and Berry (2010).

Developing decision criteria. As mentioned previously, one of the purported advantages of NHST is that it provides a bridge from theory to conclusion via a p-value (Cortina & Landis, 2011). However, not all decisions require one to make a decision. Clearly, certain situations call for such judgments (e.g., Does this finding have scientific merit? Should this treatment be used? Are side-effects of medication small enough to merit implementation?) and alternative frameworks have no clear route from theory to judgment.

Unfortunately, using a universal criteria ($p < 0.05$) has, in a way, "hijacked" decision making from the scientific community. A p-value is a function of both the sample size and the effect size. In certain domains (e.g., neuroscience) a large N is simply not feasible, and yet the culture of NHST does not permit flexibility in considering other decision criteria

that allow for more lenient p-values. On the other hand, alternatives to NHST (e.g., effect sizes) have been criticized because they do not provide simple rules for deciding whether a finding has (or has not) scientific merit (Cortina & Landis, 2011).

I advocate, instead, the decision criteria be left in the hands of the researcher. Researchers may choose, if they wish, that a significant finding is one that reaches $p < 0.05$. Other researchers, on the other hand, may decide a clinically significant finding is one where $d > 0.83$, or a mean difference between treatment and control group is greater than 10 points, or expenditures are reduced by $10,000. In short, any metric may serve as the basis for making decisions; one is not limited to p-values.

One might criticize this approach by asking what is to stop researchers from setting more lenient criteria than $p < 0.05$. Researchers have a vested interest in a paper reaching "significance" (however it is defined) and if they can lower the threshold for reaching significance to even less than what is currently acceptable, they will abuse that. They might, for example, state in advance that any correlation greater than 0.0000001 is clinically significant.

Fortunately, researchers eventually have to defend their decision criteria when submitting their paper for review. If they set a low bar for their decision criteria at preregistration, they will have to answer to a skeptical community of reviewers at a later date. I suspect this knowledge will severely limit the degree to which researchers seek to abuse the practice of setting their own decision criteria. Rather, I suspect researchers will actually impose *more* stringent criteria on their hypotheses. In addition, by placing these sorts of decisions back in the hands of the researcher and the scientific community, the significance of results will not have to be qualified as "statistically but not clinically significant."

2. Assess psychometric properties.

Many have suggested the "replication crisis" is a result of attempting to make conclusive answers on noisy data (Loken & Gelman, 2017). The obvious solution to the noise is simply to increase the sample size. In some situations, however, this is not feasible, nor is it always the most practical approach. Gelman (2018) noted that doubling the reliability of a test will yield equivalent gains in precision as quadrupling one's sample size. In other words, we might get more "bang for our buck" by spending a bit more time with measurement.

By assessing the psychometric properties of our measures, it will invite deeper thinking on measurement issues and how they might affect data analysis. If our measures fail psychometrically, no amount of sophisticated modeling will yield any insights that have scientific merit. For more details about assessing psychometrics, see Furr (2014).

3. Plot the Univariate Distributions

The third step of data analysis is to plot the univariate distributions of the variables of interest. For quantitative variables, histograms are good candidates. (Quantile-quantile plots may also be beneficial, but they are less common and thus less interpretable). For

categorical variables, bar charts are appropriate. A visual of the distribution will inform the researcher of several potential pitfalls in the coming analysis, such as:

- Incorrectly coded values (e.g., with a second wave of data collection, the researcher accidentally changes the treatment group designation from "Treatment" to "TRT," then later aggregates the two data waves and accidentally treats those labeled as "Treatment" and "TRT" as separate groups)
- Improperly coded missing values (e.g., a -999 is treated as a value, rather than a missing variable)
- Non-normality (e.g., if there are excessive zeroes in the sample)
- Outliers
- Role reversals (e.g., if a bar chart shows many more women than men firefighters, the labels were likely flipped)

Beginning with such plots may prevent embarrassing retractions later. For example, Hofmann, Fang, and Brager (2015) wrote an article that suggested Oxytocin reduced psychiatric symptoms, but later had to retract the article. When entering the effect sizes for a meta analysis program, they assumed all effect sizes were positive. Because the program they used required specifying which effects were negative (and because they improperly assumed they were all positive), the aggregated effect size was inflated. This could be avoided by simply plotting the univariate distribution of effect sizes. (Read more at Chawla, 2016).

At this point it may not be necessary to address the outliers and/or non-normality of the data. Remember the assumption of linear models (e.g., regression, ANOVA, *t*-tests) are that the *residuals* of the model are normally distributed. The outcome variable itself need not be normal (though it usually helps). Often including one's predictors in a model will render the residuals normal, even if the variable itself was not normal. Likewise, an outlier in univariate space may not be an outlier in multivariate space. However, plotting the univariate distributions in advance will inform the researcher of potential problems that may occur later in the analysis. For a video playlist on evaluating univariate distributions, see https://yt.vu/p/PL8F480DgtpW-T_ySqIurOMIaChNlOr3Ka.

4. Plot a Graphic to Match the Theoretical Hypothesis

Once the univariate distributions are plotted (and any coding errors handled), the next step is to plot a graphic to match the theoretical hypothesis. If one were to perform simple linear regression, for example, a scatterplot would help the reader visualize the results. For ANOVAs/*t*-test, violin plots or bee swarm plots would be appropriate. Although some of the strategies used (and types of plots) may be new, all of them are easy to produce in either R or even point and click software, such as Jamovi (project, 2019) or JASP (JASP Team, 2019).[7] For a tutorial on determining which graphic is most appropriate and for instructions on creating each of these types of plots, see Fife (2019a), as well as a video playlist at https://yt.vu/p/PL8F480DgtpW8WFhHFRzos7iUK2r-MhmKw.

[7]SPSS can perform most of these plots, but unfortunately (as far as I know) does not allow the user to overlay the raw datapoints over boxplots, mean plots, multi-way dot plots, etc.

The advantage of utilizing *good* graphics is that they make it nearly impossible to deceive one's readers (or one's self, for that matter, especially if raw data are displayed). Sound graphics are essential for identifying two problems in particular: (1) outliers, and (2) non-linearity.

If one or more outliers are present, it is possible they are driving statistical significance. If so, this will be easy to determine from a graphic. Likewise, if the researcher attempts to fit a straight line to data that are clearly non-linear, usually a graphic will show the error of one's ways. If this is the case, one need not hang their head in defeat. It simply means the researcher has chosen the wrong model and must select another (such as polynomial regression, a transformed dependent variable, non-parametric procedures, or generalized linear models).

To better detect departures from linearity, I recommend adding loess lines to a graphic; the package Flexplot (Fife, 2019b), which is available in R, the point-and-click platform Jamovi, and will shortly be released in JASP, defaults to show loess lines in scatterplots. Loess lines are non-parametric curves that are allowed to "bend" with the data. They can assist in detecting non-linearities that a standard model (which forces a straight line) would not detect.

Recall our original goal: we are trying not to make a false positive/negative. At this point, it is impossible to make a false conclusion because we haven't made a conclusion at all. We have not yet even computed an estimate, let alone made a decision about statistical significance. In addition, the researcher may decide that computing the p-value on a dataset is not necessary. If the visual analysis shows a whopping effect, who really cares about whether it is statistically significant? In the words of Joseph Berkson, these sorts of images pass the "intraocular trauma test" (Berkson, 1942). (Though the converse is also true. I once had a graduate student produce scatterplots and find the predicted relationship was non-existent. She immediately quit the analysis and concluded, "If there is an effect, it's so small I don't even care about it." I promptly gave her an A on her assignment).

In short, plotting the data before computing the analysis will prevent researchers from deceiving themselves, render tests of statistical significance less necessary, and force the reader to think in terms of the size of the effect, rather than its existence.

Unfortunately, not all statistical analyses lend themselves nicely to graphical display. Structural equation models, factor analysis models, and hierarchical linear models, for example, are more difficult to map onto a single plot. These may require multiple plots and, unfortunately, the fragmented nature of these graphics may detach the visuals slightly from the actual analysis. Future research ought to attempt to bridge that gap and find intuitive, graphical representations of these more complex models.

5. Study the Residuals

After plotting the data, the researcher may not know if the chosen analysis (e.g., linear regression) is appropriate. The data may violate the assumption of linearity, normality of residuals, or homoskedasticity. Yet in order to properly diagnose the problem, we ought

to compute the residuals (and thus have to actually perform the analysis). Unfortunately, before extracting the residuals, the model has to actually be fit to the data. I would advise the reader to close one's eyes (metaphorically or otherwise) until after the residuals have been studied before studying the results of a statistic.

With residuals in hand, the researcher is now ready to properly assess the assumptions of the model. It is common (at least in the bio-medical literature) to perform statistical tests of normality or homoskedasticity, or independence. I would advise against it. Like other tests of significance, these tests of assumptions are sensitive to sample size. With a small enough N , even large departures from statistical assumptions are not detected, while with large N , even trivial differences are flagged. These statistical tests tell us whether our distributions depart from what is expected. They do not tell us whether they are different enough to muck up our analysis. The latter is better done through visual interpretation of results (as well as through a sensitivity analysis).

Histograms will inform the researcher whether the normality assumption has been approximately met. Residual dependence plots assist in determining whether linearity and homoskedasticity have been met. Either of these plots will assist in flagging outliers. For instructions in how to diagnose problems using these plots, see Kutner, Nachtsheim, Neter, and Li (2004), as well as the following YouTube playlist: https://yt.vu/p/PL8F480DgtpW8v-h-7s9Ih826Qi7aa3rBS

If the visual inspection of the residuals signals problems, one may have to iterate through steps 2-4 until the assumptions have been met, each time making a modification to the model (such as transforming the DV, removing outliers, utilizing weighted least squares, or using generalized linear models). Furthermore, problems at this early stage demonstrate the researcher is not yet ready for confirmatory data analysis. Again, there is nothing wrong with migrating to rough confirmatory or exploratory data analysis and if a researcher finds the intended model is not appropriate, the researcher ought to explicitly state their analysis has turned from confirmatory to something else (Fife & Rodgers, 2019). Regardless, the researcher may proceed to Step 6 once the assumptions have been met.

6. Interpret Effect Sizes/Parameter Estimates

At this point the researcher has far more information about the data than what is typically reported in psychological journals; the researcher knows outliers are not driving the analysis, knows the model chosen is appropriate, and has a visual that illustrates the strength of the relationship between the variables of interest. After Step 5, the researcher ought to be confident the model chosen is appropriate (i.e., the assumptions of the model have been met). Once again, I emphasize that it is impossible to commit a Type I (or Type II) error because statistical significance has not yet been evaluated. Likewise, no conclusions have been made, so it's impossible to make a false positive. Rather, I recommend the researcher study and interpret effect sizes and parameter estimates. We all have been cautioned against making mountains out of molehills, or emphasizing statistical significance at the expense of practical significance. This is why the APA recommended researchers report effect sizes in addition to statistical significance. I would argue practical significance is far more important than

statistical significance. Studying the effect size (and parameter estimates) before statistical significance is a conscious choice aimed at reminding the researcher of this preference for estimation rather than significance.

Most statistical packages offer readily available estimates of effect sizes, including f^2, part and partial correlations, r^2, and Cohen's d. To determine which measure of effect is appropriate, I recommend the concise and effective article by Cohen (1992). For software dedicated to effect size calculations and graphical data analysis, see Fife (2019b), as well as a tutorial on estimates at https://yt.vu/p/PL8F480DgtpW8jshu9vTyCf4HqSHYx05Fw.

Where possible, effect sizes in the original (unstandardized) metric should be interpreted (Baguley, 2009; Bond, Wiitala, & Richard, 2003; Tukey, 1986). In a regression, the parameters of interest are the slopes (and occasionally the intercept). For ANOVAs/t-test, the parameters of interest are the mean differences between groups. For structural equation modeling, the parameters of interest are the path coefficients. For logistic regression and other generalized linear models, the researchers may have to perform mental gymnastics as they attempt to interpret things in terms of log odds (or in terms of odds ratios).

Studying these parameters adds another layer of depth at which the researcher can make sense of the data. Not only will it inform the researcher about the direction of the effect (e.g., males scored higher in aggression than females, anxiety is positively predictive of depression, performance is inversely related to mood), but also offers a mathematical equation that maps predictors onto outcomes. For example, suppose a researcher performs a regression that assesses weight loss from experimental condition, controlling for motivation. Further suppose the regression equation is as follows:

$$\textit{weight change} = 1.2 - 0.8 \times \textit{motivation} - 4.5 \times \textit{treatment} - 1.2 \times \textit{motivation} \times \textit{treatment}$$

This regression equation would be an interesting result indeed. This suggests the following:

- Those in the control group who have no motivation will actually gain an average of 1.2 pounds (because control group is the reference group)
- Every time an individual increases their motivation by a point, they can expect to lose 0.8 pounds
- The treatment group averages 4.5 pounds more weight loss than the control group
- The relationship between motivation and weight loss is stronger for the treatment group than for the control group, such that for the treatment group, for every point increase in motivation, they lose an additional 2 pounds (i.e., 1.2+0.8)

Granted, much of this information could be gleaned from a graphic, but the estimates put the visual interpretation into concrete mathematical terms that are interesting in their own right. Furthermore, the effect sizes and parameter estimates reduce ambiguity inherent in visual interpretation. To further reduce ambiguity (and marry the ideas of significance testing with estimation), a researcher should pair confidence (or credible) intervals with these estimates. Doing so will further reduce ambiguity, while explicitly recognizing the degree of uncertainty.

7. Make a Decision (If Applicable)

Recall we have plotted univariate distributions to flag potential data recording errors, assess normality, and identify potential outliers. We have also created a visual representation of our analysis that shows both the size of the effects and the direction. We have also thoroughly assessed whether our model is appropriate through residual analysis, estimated effect sizes, and interpreted parameter estimates. In short, we have much more thoroughly familiarized ourselves with our own data (Tukey, 1986).

If, at this stage, the reader feels it rather pointless to assess statistical significance, I have successfully made my point. This second to last step is entirely optional and ideally makes it clear that our data have long been trying to tell us much more than we have allowed them. Simply computing statistical significance without doing the previous steps is akin to eating a single sprinkle off a large birthday cake. With so much richness remaining, it is a shame that we limit ourselves to a single test that is largely uninformative.

Earlier, I advocated that researchers instead set their own decision criteria. At this point, making a decision of significance is easy; one has already pre-specified what is clinically significant and now they simply compute the numbers and identify whether significance was reached.

8. Replicate With New Data

The decision made in the previous step is always provisional. Few single studies have the power (statistically or otherwise) to make conclusive statements about the truthfulness of a hypothesis. Rather, these findings are tentative and ought to invite closer scrutiny and replication. The discovery of the Higgs Boson, for example, was not considered settled until after hundreds of trillions of replications. The estimates obtained for the parameter of interest may serve as a prior in a Bayesian analysis, or could to be aggregated into the next study (and others like it) via meta-analysis. Such cumulative methods will invite a greater sense of humility about one's own role on the scientific process and consequently invite deeper attention to development of theory.

Reporting Results

I recommend every researcher perform the eight steps when doing data analysis. However, it may not be necessary to report every step in a journal article. Not only would this increase the length of most articles (a trivial problem as journals become more digitized), but it may detract from the purpose of the article (to evaluate the original hypothesis). However, at minimum, I strongly recommend a researcher's final report contain:

1. One or more graphical depictions of the analysis of interest (Step 4).
2. A comment on how the researcher determined the appropriateness of statistical assumptions.
3. Parameter and effect size estimates, with confidence or credible intervals.

4. A supplemental section containing all graphics and sensitivity analyses *or* a link to a website where these can be viewed.

In other words, I am not advocating for a complete revamping and replacing of how statistics are reported in journal articles. Rather, I suggest we add these few pieces of information so the richness of our data becomes more visible.

Example

In the section that follows, I decided *not* to re-analyze existing datasets of previously published papers for two reasons. First, it can be difficult to find studies where researchers have actually uploaded their data for public scrutiny (although the Open Science Framework is making data far more accessible). Second, I would hate to pick on researchers conscientious enough to actually offer their dataset by highlighting their data analysis mistakes. I want to encourage openness, and becoming a public data vigilante would be counterproductive. Consequently, I will analyze a publicly available dataset, the National Survey of Drug Use and Health (2014) and offer my own hypothesis that, when analyzed using traditional NHST methods, yields misleading results.

1. State the Theoretical Hypothesis of Interest and (Optionally) Set a Decision Criteria.

Suppose I am a drug counselor that has had only marginal success in assisting heroin addicts overcome their addictions; those who use heroin experience more psychological distress, which in turn motivates them to escape their distress via heroin. Now let us suppose I believe promoting healthy behaviors (e.g., exercise, nutritious eating) will reduce psychological distress and help break that negative feedback loop. Ideally, one would perform an experiment, but perhaps as a preliminary study, I decide to use an existing dataset to perform an observational analysis to assess the potential efficacy of promoting healthy behaviors in a full experiment.

However, I may consider controlling for mental illness. One may be experiencing psychological distress because of their mental illness, which may make them more likely to escape such distress through drug use. Stated differently:

Among heroin users, those who report having more healthy behaviors will report less psychological distress, after controlling for mental illness

Further suppose that this is the first attempt the researcher has made at evaluating this hypothesis. In other words, the researcher is in "rough" confirmatory mode, at best, and may even be leaning toward exploratory research. Given that, it makes little sense to set a decision criteria since any attempt to do so will be somewhat arbitrary.

For illustrative purposes, I will test this hypothesis using standard NHST methodology using an ANCOVA model. Based on these data, I could conclude:

> Self-reported health rating was significantly associated with psychological distress, after controlling for mental illness, $F(4, 189) = 10.84$, $MSE = 28.86$, $p < .001$, $\hat{\eta}^2_G = .187$.

As I will show, the above statement is both misleading as well as incomplete.

2. Psychometrics

The NSDUH utilizes the Kessler-6 distress scale to measure psychological distress (Kessler et al., 2003). This is comprised of six items that asks the degree to which participants experienced the following symptoms within the last thirty days: nervousness, hopelessness, restlessness, depression, feeling everything required effort, and worthlessness. Reliability estimates suggested that within this sample, the measure was quite consistent (α=0.94). Because of the high reliability estimates, I simply summed the scores to create my distress scale.

As for the other two variables, mental illness was derived using a logistic regression model that utilized various indicators of mental health (e.g., suicidal thoughts, suicidal attempts, major depression, psychological impairment). Health is a one-item, self-reported question that asks them to rate their overall health. Unfortunately, I cannot assess the reliability of either health or mental illness, since both were assessed using only one item.

3. Plot Univariate Distributions

I plotted the univariate distributions of the three variables of interest: probability of mental illness, health rating, and psychological distress. These distributions are shown in Figure 2.

The plots reveal potential issues with the data. The probability of mental illness is far from normally distributed. The mode of the distribution is near zero (i.e., the data are zero-inflated). Note that linear models make no assumptions of normality for the independent variables (or the dependent variables for that matter, rather the assumption is about the *residuals* of the dependent variable). However, in my experience, if both the IV and the DV are skewed, the assumption of linearity will almost certainly be violated. Given that the distress variable is also skewed, this could certainly be problematic for linear models (such as an ANCOVA).

At this point I am primed to look for serious issues with normality, linearity, and likely heteroskedasticity. As I mentioned earlier, if these assumptions are violated, it simply means we have chosen the wrong model to fit the data.

4. Plot a Graphic to Match the Analysis of Interest

Recall that the fictitious researcher has decided to perform an ANCOVA. An ANCOVA essentially performs a standard linear regression between the covariate and the DV, extracts the residuals, then performs an ANOVA on the residuals (though this is all done

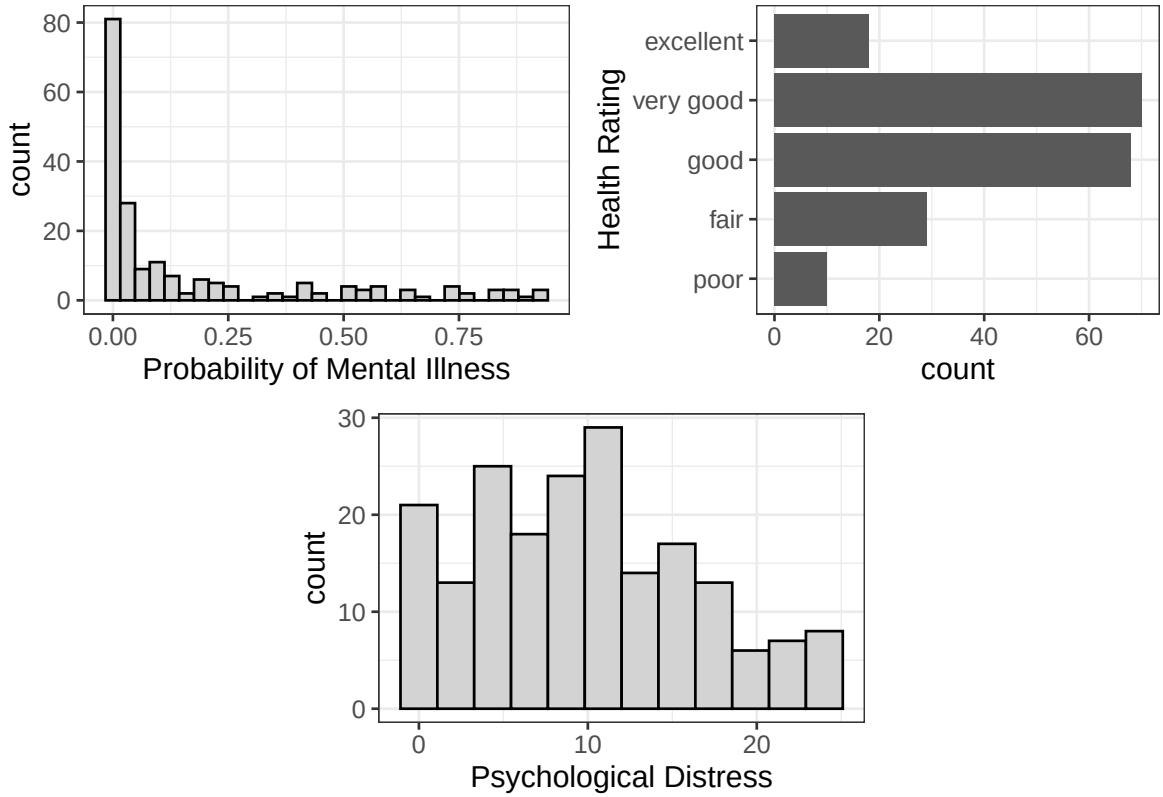

Figure 2. Univariate distributions of psychological distress, health rating, and probability of mental illness.

simultaneously with an ANCOVA). An "added variable plot" would be an appropriate graphic to match the ANCOVA, where the grouping variable is plotted against the residuals of the model where the covariate is removed. This image is shown in the left image of Figure 3. One shortcoming of this sort of plot is that it masks any violations of the assumption of homogeneity of regression. (Homogeneity of regression states that the regression lines for each group are parallel). As such, I have plotted a paneled graphic in Figure 3, which shows a different scatterplot for each level of health rating, with quadratic lines overlaying the data. I used quadratic lines to see if nonlinearity might be an issue. Once again, there are a few things worth noting:

1. These fitted lines are not "parallel," meaning that the assumption of homogeneity of regression has been violated (indicating that ANCOVA is not appropriate). This also means that the left image in Figure 3 is misleading.
2. The fitted lines are not linear, indicating that linear models will not be appropriate.
3. There are very few people who report poor or excellent health, suggesting that I might combine the poor/fair groups as well as the excellent/very good.

Also at this point, I shouldn't interpret the plots shown in Figure 3; the model clearly does not fit. As a result, I may have to iterate through this step (and probably Step 4) until

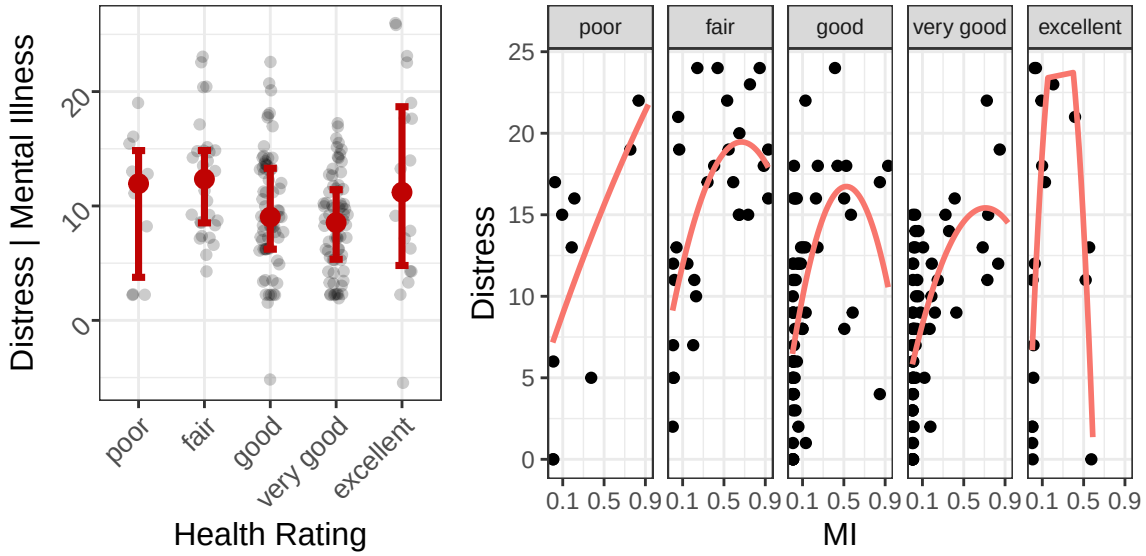

Figure 3. Visual displays of the relationship between mental illness, health, and distress. The left image is an added variable plot of the relationship between health rating and distress (controlling for mental illness). The right image shows the relationship between mental illness and distress for each level of health rating. The lines are quadratic lines mapping the relationship between mental illness and distress, conditional on self-reported health rating.

I find a model that appropriately fit the data. It is also clear that any attempt at strictly confirmatory analysis must take a backseat as I have some decisions to make that were not anticipated a priori. I have a few options:

1. I can transform the dependent variable and attempt to "linearize" the relationships. This generally fails when the DV is zero-inflated (as it is in this case).
2. I can attempt to use non-parametric procedures, such as rank transformations of the dependent variable. Unfortunately, rank transformations fail to preserve interaction effects (which are clearly happening, as seen in Figure 3).
3. I can perform more "modern" robust methods (Erceg-Hurn & Mirosevich, 2008). These methods essentially replace mean-based estimates (including conditional means) with trimmed means, standard variances with winsorized variances, and standard confidence intervals (CIs) with bootstrapped CIs that are computed from the trimmed/winsorized estimates. Unfortunately, these methods would not work since more than 10% (the "traditional" degree of trimming from either tail) of the tail of the distribution is contained at zero. (In general, modern robust methods do not work well for zero-inflated data).
4. I can attempt to fit a nonlinear model. Unfortunately, these sorts of models can be more difficult to interpret (because coefficients may not have intuitive interpretations).

Unfortunately, *all* of these choices are complicated, and require some sophisticated modeling. I did not choose a sophisticated model to show off my statistics skills. (In fact, it

took me an embarrassing amount of time to find something that would actually model the data well)[8]. Rather, I chose a particularly illustrative example that, unfortunately, required a relatively complex model. (In full disclosure, I intentionally "fished" for an example that would be particularly illustrative).

Of the strategies listed above, I favor the fourth strategy. Once again, I prefer a visual approach to modeling these data, so I will overlay the fit of the model atop the raw datapoints. Before I do so, however, I will aggregate the poor and fair health categories, as well as the very good and excellent categories, otherwise, the poor and excellent categories may be unduly influenced by outlying datapoints.[9]

I attempted to use multiple nonlinear models, including a gamma generalized linear model (GLM), a gamma GLM with a polynomial term, a random forest model (Breiman, 2001), and an ordered logistic regression. Most failed to model the data adequately, as the fit of the model failed to pass through the more concentrated parts of the dataset. Finally, I settled on a splined model, as well as a nonlinear model that utilizes the Michaelis-Menten equation (MME). The MME was designed to model enzyme reactions and has the following form:

$$\mathrm{Y} = \frac{V_{max} \cdot X}{k_M + \mathrm{X}}$$

where V_{max} represent the maximum fitted Y value (distress in this case) over the entire range of X (mental illness) and k_M is loosely interpreted as the rate of increase in distress. (Technically, k_M is the point at which V_{max} reaches its halfway point, which increases if it has a steeper slope). To allow the model to generate predictions for each health rating, I modified the model as follows:

$$\text{Distress} = \frac{\text{MI} \cdot (V_{max} + \beta_1 \text{Good Health} + \beta_2 \text{Fair Health})}{k_M + \beta_3 \text{Good Health} + \beta_4 \text{Fair Health} + \text{MI}}$$

where β_1/β_2 indicate how V_{max} differs for those in good/fair health relative to the "very good" health individuals, and β_3/β_4 indicate deviations in slopes. The model's parameters were estimated using a Bayesian approach with diffuse priors. A Bayesian approach will make it more seamless when I replicate these findings in Step 8 (because the posterior estimates can serve as priors for the replication).

Why did I use the Michaelis-Menten equation? Well, because it fit the data, as shown in Figure 4. The top rows are for the model where I combined fair/poor and very good/excellent, and the bottom plot is for the non-combined data. Both models fit fairly well, except for the spline model under the excellent condition. For that reason, I am going

[8] To see a YouTube video explaining my thought process during my statistical modeling, visit https://youtu.be/5BpmktmvgIA.

[9] I did perform a sensitivity analysis on this decision as well (i.e., performing the analysis both before and after aggregating those categories). Combining the categories makes the picture more clear. Whether that clarity is spurious, I leave it to the reader to decide. Occasionally, I will present the results from both approaches.

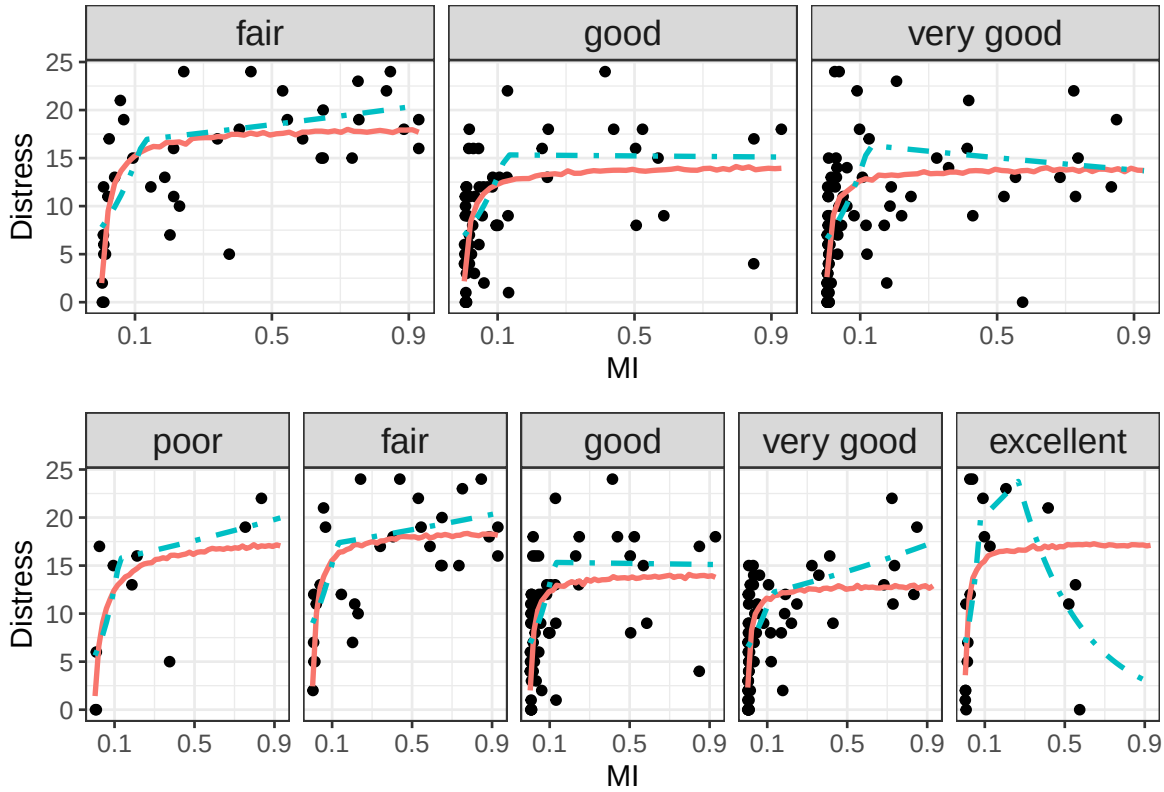

Figure 4. Predictions for the spline (blue lines) and Michaelis-Menten equation (MME; red lines) models of the NSDUH data. The top rows combine the poor/fair and very good/excellent conditions, while the bottom plots are for the uncombined analysis.

to choose the MME model for the remainder of my analyses. I am also inclined to choose the model that combines fair/poor and very good/excellent; the extreme categories have so few people I don't trust the predictions.

At this point, I am also going to temporarily refrain from interpreting the graphics until I have assessed the viability of the assumptions, which I will do in the following section.

5. Study the Residuals

To ensure that the MME model adequately fits the data, I generated residual plots (Figure 5), including a histogram of the residuals, as well as residual dependence and scale location (SL) plots. Again, these are the results for the combined categories (though the uncombined looked similar as well). There are potential deviations from homoskedasticity; the residuals have a slight "megaphone" shape. However, this is likely because the DV is a likert scale, which limits variability at the upper and lower ranges of distress. However, inferences tend to be fairly robust to modest deviations (Maxwell & Delaney, 2004). As such, the model appears to at least approximately meet assumptions.

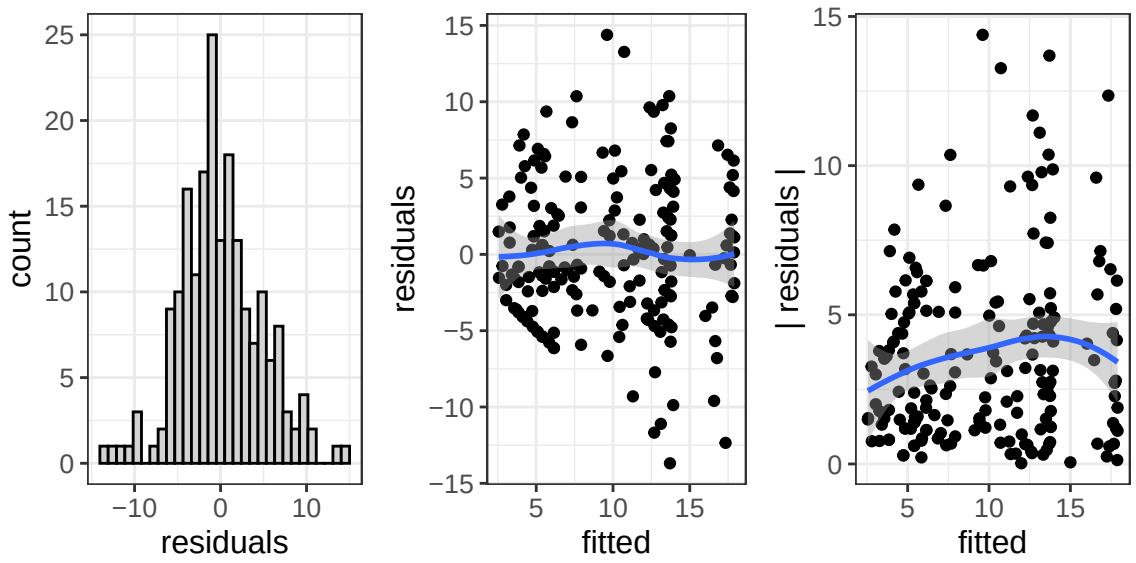

Figure 5. Residual plots of the MME analysis: a histogram of the residuals, a residual dependence plot, and an SL plot. Plotted lines are loess lines.

6. Study Parameter Estimates/Effect sizes

As mentioned previously, β_1 and β_2 indicate how the maximum predicted value in the good/fair health individuals (respectively) differ from the maximum of those in very good health. The difference between those in good/very good is 0.18 distress points, with a 95% credible interval ranging from -2.83 to 3.29, while the difference between those in fair versus very good health is 4.31 distress points, with a 95% credible interval of 1.37 to 7.25.

7. Determine Clinical Significance Based on a Decision Criteria

As I mentioned previously, this analysis was not a replication of a previous study. As such, there was little statistical information from which I could derive a decision criteria. However, in the next step, replication, I now have empirical information from which to generate a decision criteria.

In the mean time, however, I will spend some time summarizing the insights I have gained from my first analysis. Both mental illness and distress were skewed and the fit of the model was not linear. Interestingly, as individuals in this sample increased in mental illness, there are very dramatic increases in distress, though only to a point (approximately 0.10). From that point, it did not seem to matter whether an individual had a really high probability of mental illness (e.g., 0.9) or relatively low (e.g., 0.1), their distress level is approximately the same. Additionally, those in fair/poor health do seem to experience more distress than those in very good health, or at least the estimated maximum distress (V_{max}) for the two groups are different by about 4.31. In other words, having even moderately good health limits the maximum distress one might experience. On the other hand, increasing

one's health from good to very good/excellent doesn't seem to make much of a difference.

Revisiting my original hypothesis, I sought to estimate the effect of health after controlling for mental illness. However, I'm not entirely sure it makes sense to control for mental illness; the relationship between mental illness and distress is both complicated and nuanced and it doesn't really make sense to me to remove its effect from my interpretation. Put differently, the effect of health on distress depends *highly* on one's mental illness, and the effect of health can only be interpreted in that context.

8. Repeat With a New Dataset

Fortunately, the NSDUH routinely reports results of their survey of drug and alcohol use every year. As such, I decided to do a strictly confirmatory test of my model on the 2018 survey. To do so, I used the posterior distribution from the previous MME model as the prior for the Bayesian model. Aside from inputting the priors, I used identical syntax to run the analysis. For my decision criteria, I chose the lower limit of my 95% credible interval for β_2, 1.37, as my decision-criteria. In other words, if the difference between those who self-report as fair versus very good is more than 1.37 points, I consider that difference practically significant.

Figure 6 shows the results of the replication. As before, each health category is displayed as a separate panel and the fits of the MME model are shown as red lines. The black lines are "ghost lines" (Fife, 2019b), which simply repeat the pattern from one panel ("good" in this case) across the other panels to make it easier to compare fits across panels. For the top graphic, the results are remarkably similar to those in Figure 4. Also, the average difference in distress between those of fair versus very good health is 5.42, which is well above the decision criteria. Further, the 95% credible interval has narrowed from the initial study (1.37, 3.29) to the replication (4, 6.87).

Fortunately, the uncombined results also suggest that improved health mitigates distress. Both the poor and fair conditions max out at much higher levels of distress than those in good/very good/excellent distress. Although the excellent condition does not follow the same pattern, likely because there is so much noise.

Results Section

Earlier I stated that, at a minimum, a researcher should report (1) a graphical depiction of the analysis, (2) a comment on the appropriateness of statistical assumptions, (3) parameter/effect size estimates with confidence intervals, and (4) a supplemental section or link where the reader can view all graphics/sensitivity analyses. For this example, the results section may read as follows:

> Upon visual inspection of the residuals, it was determined that linear models were not appropriate. This discovery forced a change from confirmatory to exploratory analysis. Consequently, we utilized a nonlinear model, using the Michaelis-Menten equation*:

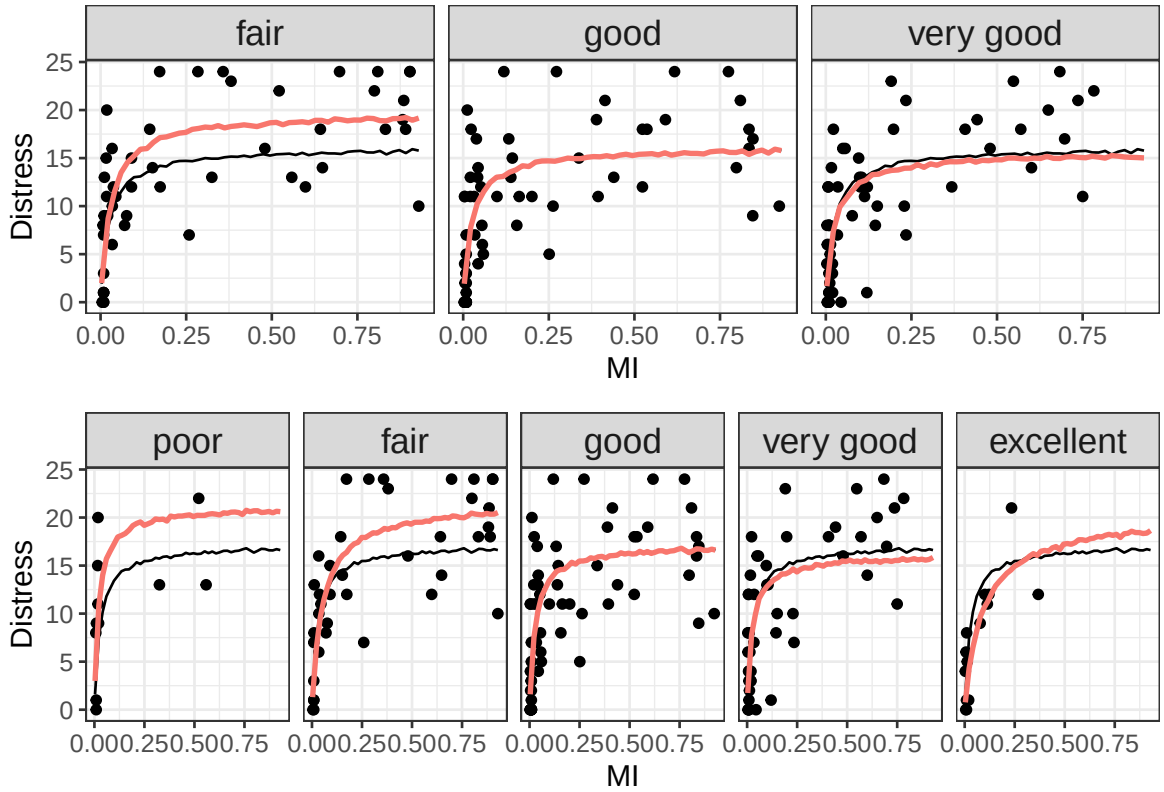

Figure 6. Results of a strict confirmatory replication of the MME model, which models the relationship between mental illness/health and distress. The top plots show the combined results while the bottom plots show the uncombined results. The posterior of the parameter estimates from the original study served as priors in a Bayesian analysis. This graphic shows the fit. Black lines are ghost lines (Fife, 2019b), which repeat the pattern from the good health condition to the other conditions.

$$\text{distress} = \frac{\text{MI} \cdot (V_{max} + \beta_1 \text{Good Health} + \beta_2 \text{Fair Health})}{k_M + \beta_3 \text{Good Health} + \beta_4 \text{Fair Health} + \text{MI}}$$

We also aggregated the poor/fair, as well as the excellent/very good groups because data were quite sparse at the extremes. Sensitivity analyses revealed that the final conclusions were relatively insensitive to whether the data were combined or not. Further, we utilized a Bayesian analysis with diffuse priors. This model was fit to the 2014 NSDUH dataset, the hypotheses preregistered, then replicated on the 2018 dataset, but using the posterior of the original analysis as the priors in the replication. The results of the replication are presented in Figure 6. The predicted difference between those with fair versus very good health was 5.42, with a 95% credible interval of 4 to 6.87. More details (such as graphics of distributions and sensitivity analyses) can be viewed at

http://www.examplesite.com/article

* We also analyzed the data using a various other models, including a spline model. These other models fit the data poorly. For full details and source code, see the supplemental section.

Summary

My first naive NHST analysis revealed a significant effect of health on psychological distress among heroin users, with a respectable effect size. However, by following the eight steps of data analysis we have learned many things we otherwise would have missed, such as:

- Mental illness is highly skewed, and distress is moderately skewed
- The relationship between distress and mental illness is highly curvilinear; as the probability of mental illness increases, one's distress increases rapidly, then levels off with higher levels of mental illness.
- Those with a 10% probability of mental illness have about equivalent distress as those with a probability of 90% or higher.
- Regardless of one's health, the predicted rate at which distress increases is the same, though the predicted maximum distress is different.
- Relative to those of poor/fair health, having good health (or very good) reduces the maximal distress someone might experience by approximately 5.42 points.
- Improving one's health from good to very good makes little difference in distress.

It is important to note that the issues noted above led to *substantially deeper insights* about the data which were entirely missed by the standard approach.

Discussion

The recent "replication crisis" suggests there are statistical practices within the field of psychology that inflate false positives and negatives. These practices include "p-hacking," failing to meet statistical assumptions, and a narrow focus on statistical significance rather than interpreting what the data are actually telling us. In this paper, I have suggested a framework under which researchers might perform data analysis that easily fits within current data analysis practices, while inviting a greater focus on estimation and data visualization. In addition, this framework provides step-by-step guidance for researchers that aims to empower analysts to focus on what the data are actually saying.

I have also highlighted these principals and practices with the use of an actual dataset. My analysis revealed that performing the NHST ritual, even when effect sizes were reported, yielded a misleading and incomplete picture. My re-investigation of the same hypothesis revealed patterns more nuanced than a single NHST p-value (or effect size) captured.

It is my hope the example I provided was illustrative. I suspect most researchers do not have the interest to utilize advanced nonlinear models. Allow me to offer some hope. First, I intentionally chose a dataset I knew would severely violate the assumption of linearity. I would hope most researchers would not encounter such "zero-inflated" models where

non-linear relationships are almost inevitable. On other hand, some constructs psychologists investigate (e.g., frequency of rape occurrences, number of times one has attempted suicide) should not be investigated with standard linear models. Those researchers who study these types of data perhaps ought to learn more sophisticated models to best interpret what their data are saying. However, it may be optimistic to think researchers will learn these complex modeling techniques themselves. In these situations, it might be best to collaborate with those who are familiar with these techniques. On the other hand, visuals are both intuitive to interpret and easy to produce. As such, these ought to *always* be employed.

Finally, despite my best efforts to emphasize this framework easily fits within current statistical practices, I suspect there may be some resistance. For example, editors may lament that performing these eight steps will double the length of the average article. I agree, though this is less of a concern as journals become more digital. However, I do not think it necessary every plot and sensitivity analysis make it to the final version of the paper. Researchers already frequently omit details about data cleanup and how models were decided. However, I do strongly suggest this information be publicly available, either through supplemental material or through the author's website. Doing so will allow future consumers of the research to understand what decisions were made, why they were made, and how these decisions may (or may not) have affected the analysis. In addition, it provides additional tools to consumers that allows them to judge the verisimilitude of the research themselves.

Another obstacle to incorporating these suggestions may be a lack of training. Many researchers may not know how to graph loess lines, jitter categorical variables, or create paneled plots. Because of this, I have created a step-by-step tutorial that shows researchers how to visually represent the most common analyses, including regression, multiple regression, factorial ANOVAs, and t-tests. This tutorial demonstrates how to perform these in both the point-and-click software Jamovi. This tutorial can be found at http://rpubs.com/dustinfife/528244. Additional resources can be found at https://yt.vu/p/PL8F480DgtpW_v1fmBauNMPF9Gqdoaa8zJ.

I also invite flexibility among reviewers and editors. I recognize the approach I introduce is different than what is traditionally taught in textbooks and what is traditionally performed in applied settings. While traditional statistics is taught as a mechanical sequence that yields unambiguous answers to research questions, my approach encourages flexibility in reporting results. The natural reaction among editors and reviewers might be to reject anything unfamiliar. However, as illustrated in the example, this would be unwise. This approach to data analysis invites ambiguity and forces analysts, reviewers, and editors to confront the uncertainty inherent in data analysis. This is a very good thing.

Under such flexibility, I hope editors and reviewers might be open to very different approaches to analyzing data and interpreting results. These might include:

- *Analysis sections that do not report p-values.* Analysts utilizing the eight steps approach may feel that a p-value is misleading and/or inadequate at summarizing one's results. Rather, they may choose to determine statistical significance using other values (e.g., Bayes factors, mean differences, slopes).

- *Reports that are purely exploratory.* As users practice this approach to data analysis, they will quickly learn the richness intrinsic in data and recognize a largely untapped resource. Indeed, this is what I *hope* will happen. This will inevitably lead to an influx in EDA. While some have suggested *all* published analyses ought to be confirmatory (Lindsay, 2015), I feel otherwise. Granted, EDA offers weaker evidence than CDA, and all EDA ought to be followed by CDA. However, if the choice is between failing to publish EDA because one does not have the resources to CDA and publishing only the EDA, I choose the later. Published EDA results can then be replicated by someone who does have the resources.

- *Substantive conclusions that are based on non-traditional criteria (e.g., Bayes factors, slopes, reaction time differences, correlation coefficients, AICs).* The first step encourages researchers to set their own criteria for what is deemed "significant" or important. The metrics of choice will vary from discipline to discipline and editors and reviewers ought to evaluate the choice of criteria in light of the substantive questions being asked.

- *Substantive conclusions that are based on graphical interpretation alone.* This approach to data analysis recognizes the critical role graphics play in encoding important information, evaluating model fit, conveying uncertainty, etc. If researchers utilize this approach, I *hope* they will use graphical depictions to aid decisions of scientific relevance. If they do, they ought not to be penalized simply because it is different.

- *Results expositions that seem disorganized.* If these eight steps are followed, analyses will rarely be linear; rather, analysts may have to stumble through various modeling strategies quite iteratively as they search for the best representation of the data. Although some might be inclined to relegate the stops and starts to a supplemental section, others might feel ethically inclined to convey how they arrived at their conclusions in a way that reflects the messiness inherent in their analysis. Granted, the exposition should be clear, but I invite editors and reviewers to be forgiving of interpretations that take some time to unfold.

- *Conclusions that are hedged with uncertainty and/or ambiguity.* While p-values invite a false sense of certainty, the approach I advocate often invites *un*certainty. This is a good thing. In the past, authors writing eye-catching titles and conclusions have damaged the field's reputation and perpetuated fallacious beliefs in the media. Cautious language, though less flashy than bold claims, is probably a better way of conveying results.

- *Extremely lengthy supplemental sections located on outside repositories.* Once again, this eight step approach is all but guaranteed to increase the number of starts and stumbles during the data analysis process. In line with the second guiding principle of data analysis (utilizing sensitivity analyses for ad-hoc decisions), these starts and stumbles should be communicated to the audience. This may require a rather long supplemental section and editors and reviewers should be accepting about these sorts of submissions.

- *Detailed analyses reported via blogs, YouTube videos, podcasts, etc.* Sometimes the

optimal way to convey statistical results may not be in journal format, or even a traditional supplemental section. Writing for journals often invokes a need for precision and clarity, and writers may anticipate they'll have to spend a great deal of time preemptively defending every statement and decision. This need for clarity and precision is a good thing, but may be a handicap during the more ambiguous stages of data analysis. Additionally, some results are best interpreted interactively (e.g., through rotating three dimensional plots or multiple static plots), while others may be best understood during stream-of-consciousness narration. Other avenues may be more amenable to expounding on the process whereby the researcher arrived at their conclusions, and editors and reviewers may need to be flexible in allowing this. In this paper, for example, I uploaded a screencast on my YouTube channel that more fully explains my thought-process (see https://youtu.be/5BpmktmvgIA), which allowed me to give more detailed information than I might have if I were required to write about it.

- *Alternative modeling strategies (random forest, SVM, Bayesian methods).* The eight steps inevitably will shift an analyst's perspective away from testing hypotheses and toward building a statistical model that adequately represents the data. Sometimes the best model to use will not be a *t*-test, ANOVA, regression, etc. I anticipate that as more people adopt this approach, diverse modeling approaches will become more common.
- *Become comfortable with subjectivity.* As researchers begin to make personalized decision criteria, they may choose to set their decision criteria based more on their subjective judgment of a meaningful effect. For example, a researcher might decide that a smoking cessation program is effective if greater than 20% of participants quit without remission. In this situation, it may be tempting to reject an author's decision criteria for being arbitrary (why not 30%?). However, there is nothing inherently wrong with a subjective decision criteria. A threshold of 0.05 is an arbitrary threshold for *p*-values. A researcher's decision criteria should be judged not on how objective it is, but by its stringency.

In summary, my recommendation for editors and reviewers is to never reject an article simply because it utilizes a non-standard approach for arriving at substantive conclusions. Rather, the approach should be evaluated in terms of how well it fits the needs of the situation and whether the assumptions of the model are reasonably met. If the model is appropriate and reasonable, there's no reason an author's attempt at ingenuity should count against them.

In conclusion, the discipline of psychology is at a crossroads. We can continue to participate in NHST-based psychology and the problems we have recently encountered will persist. Or we can revolutionize the way we think about analysis, listen to the messages the data are trying to tell us, and uncover truths previously buried behind ANOVA summary tables and *p*-values.

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. doi:10.1348/000712608X377117

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association*, *37*(219), 325–335. doi:10.1080/01621459.1942.10501760

Bond, C. F., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*(4), 406–18. doi:10.1037/1082-989X.8.4.406

Bonneau, G. P., Hege, H. C., Johnson, C. R., Oliveira, M. M., Potter, K., Rheingans, P., & Schultz, T. (2014). Overview and state-of-the-art of uncertainty visualization. *Mathematics and Visualization*, *37*, 3–27. doi:10.1007/978-1-4471-6497-5_1

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Chawla, D. S. (2016, October). Oh, well - "love hormone" doesn't reduce psychiatric symptoms, say researchers in request to retract. Retrieved from http://retractionwatch.com/2016/10/04/oh-well-love-hormone-doesnt-reduce-psychiatric-symptoms-says-researchers-in-request-to-retract/

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. doi:http://dx.doi.org/10.1037/0033-2909.112.1.155

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

Cook, T. D., & Campbell, D. T. (1976). The design and conduct of quasi-experiments and true experiments in field settings. *Research in Organizations: Issues and Controversies*, 223–326.

Cortina, J. M., & Landis, R. S. (2011). The earth is not round ($p = .00$). *Organizational Research Methods*, *14*(2), 332–349. doi:10.1177/1094428110391542

Counsell, A., & Harlow, L. L. (2017). Reporting practices and use of quantitative methods in canadian journal articles in psychology. *Canadian Psychology/Psychologie Canadienne*, *58*(2), 140–147. doi:10.1037/cap0000074

Coyne, J. C. (2016). Replication initiatives will not salvage the trustworthiness of psychology. *BMC Psychology*. doi:10.1186/s40359-016-0134-3

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. doi:10.1177/0956797613504966

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., ... Wilson, S. (2007). Statistical reform in psychology is anything changing ? *Psychological Science*, *18*(3), 1–4. doi:10.1111/j.1467-9280.2007.01881.x

Cumming, G., Fidler, F., & Thomason, N. (2001). The statistical re-education of psychology. In *The 6th international conference on teaching statistics.* Hawthorn, Victorial: Swinburne Press.

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. doi:10.1177/1094428110380467

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *The American Psychologist*, *63*(7), 591–601. doi:10.1037/0003-066X.63.7.591

Fife, D. A. (2019a). A graphic is worth a thousand test statistics: Mapping visuals onto common analyses. Retrieved from http://rpubs.com/dustinfife/528244

Fife, D. A. (2019b). Flexplot: Graphical-based data analysis [r and jamovi]. Available at www.Jamovi.com; www.github.com/dustinfife/flexplot. doi:10.31234/osf.io/kh9c3

Fife, D. A., & Rodgers, J. L. (2019). Exonerating eda: Addressing the replication crisis by expanding the eda/cda continuum. *Unpublished Manuscript.* Retrieved from http://quantpsych.net/fife-exonerating-eda-draft-oct2019-df-edits/

Furr, R. M. (2014). *Scale construction and psychometrics for social and personality psychology.* Thousand Oaks, CA: SAGE. doi:10.4135/9781446287866

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (2016). *What if there were no significance tests?* (2nd ed.). New York, NY: Routledge.

Healy, K., & Moody, J. (2014). Data visualization in sociology. *Annual Review of Sociology*, *40*(1), 105–128. doi:10.1146/ANNUREV-SOC-071312-145551

Hoekstra, R., Kiers, H., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why not? *Frontiers in Psychology*, *3*(137). doi:10.3389/fpsyg.2012.00137

Hofmann, S. G., Fang, A., & Brager, D. N. (2015). Effect of intranasal oxytocin administration on psychiatric symptoms: A meta-analysis of placebo-controlled studies. *Psychiatry Research*, *228*(3), 708. doi:10.1016/j.psychres.2015.05.039

JASP Team. (2019). JASP (version 0.10.2)[Computer software]. Retrieved from https://jasp-stats.org/

Jones, L. V. (1952). Test of hypotheses: One-sided vs. Two-sided alternatives. *Psychological Bulletin*, *49*(1), 43.

Kessler, R. C., Barker, P. R., Colpe, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., ... Zaslavsky, A. M. (2003). Screening for serious mental illness in the general population. *Archives of General Psychiatry*, *60*(2), 184–189. doi:10.1001/archpsyc.60.2.184

Kruschke, J. K., & Liddell, T. M. (2018). The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. *Psychonomic Bulletin and Review*, *25*(1). doi:10.3758/s13423-016-1221-4

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). Applied linear statistical models. In *Applied linear statistical models*. New York, NY: McGraw-Hill/Irwin.

Levine, S. S. (2018). Show us your data: Connect the dots, improve science. *Management and Organization Review*, *14*(2), 433–437. doi:10.1017/mor.2018.19

Lindsay, D. S. (2015). Replication in psychological science. *Psychological Science*, *26*(12), 1827–1832. doi:10.1177/0956797615616374

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325). doi:10.1126/science.aal3618

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective*. New York, NY: Taylor & Francis.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*(1), 156–166. Retrieved from https://pdfs.semanticscholar.org/2903/180261ee0d99a27cfe85cde9cf4af74923c6.pdf

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin and Review*, *23*(1). doi:10.3758/s13423-015-0947-8

National Survey on Drug Use and Health. (2014). *National survey on drug use and health 2014*. Substance Abuse; Mental Health Services Administration, Center for Behavioral Health Statistics; Quality. Retrieved from https://www.datafiles.samhsa.gov/study/national-survey-drug-use-and-health-nsduh-2014-nid13618

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–545. doi:10.1146/annurev-psych-122216

Osborne, J. W. (2013). Is data cleaning and the testing of assumptions relevant in the 21st century? *Frontiers in Psychology*, *4*. doi:10.3389/fpsyg.2013.00370

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. doi:10.1177/1745691612465253

project, T. jamovi. (2019). Jamovi (version 0.9) [computer software]. Retrieved from https://www.jamovi.org

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *The American Psychologist*, *65*(1), 1–12. doi:10.1037/a0018326

Rothman, K. J. (2010). Curbing type i and type ii errors. *European Journal of Epidemiology*, *25*(4), 223–224. doi:10.1007/s10654-010-9437-5

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115–129. doi:10.1037/1082-989X.1.2.115

Shadish, W., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632

Simonsohn, U. (2014). Posterior-hacking: Selective reporting invalidates bayesian results also. *SSRN Electronic Journal*. doi:10.2139/ssrn.2374040

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. doi:10.1177/1745691616658637

Szollosi, A., Kellen, D., Navarro, D., Shiffrin, R., Rooij, I. van, Zandt, T. V., & Donkin, C. (2019). Preregistration is redundant, at best. doi:10.31234/OSF.IO/X36PZ

Trafimow, D. (2017). Using the coefficient of confidence to make the philosophical switch from a posteriori to a priori inferential statistics. *Educational and Psychological Measurement*, *77*(5), 831–854. doi:10.1177/0013164416667977

Tukey, J. W. (1986). Analyzing data: Sanctification or detective work? *The Collected Works of John W. Tukey*, 721–737.

Valentine, J. C., Aloe, A. M., & Lau, T. S. (2015). Life after nhst: How to describe your data without p-ing everywhere. *Basic and Applied Social Psychology*, *37*(5), 260–273. doi:10.1080/01973533.2015.1060240

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Maas, H. L. J. van der, & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. doi:10.1177/1745691612463078

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Aert, R. C. M. van, & Assen, M. A. L. M. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*(NOV). doi:10.3389/fpsyg.2016.01832

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–601.