**[Building an Empirical Body of Evidence:
Developing Rapport with Reviewers and Overcoming Skepticism in Strategic Management Research]**

[Timothy J. Quigley]

PLEASE PUT THE CORRESPONDING AUTHOR IN BOLD

[Chapter Author 1: Timothy J. Quigley, University of Georgia]

tquigley@uga.edu

Author/s Biography/ies:

Timothy J. Quigley is the Georgia Athletic Associate Professor of Management in the Terry College of Business at the University of Georgia and is current an associate editor at Strategic Management Journal.

ABSTRACT

As the field of strategic management has evolved, expectations for the empirical evidence presented in manuscripts have risen substantially. Rather than a single model testing a hypothesis with a p-value below a standard threshold being sufficient, reviewers, editors, and eventual readers now demand additional evidence including multiple tests, advanced statistical models, alternative specifications, interpretation of practical rather than just statistical significance, and more. Reviewers appear to be increasingly skeptical and often raise a seemingly endless number of questions. In this chapter, I outline the idea of a body of evidence and suggests ways authors can build their evidence by anticipating reviewer questions and structuring manuscripts accordingly. Doing so allows authors to overcome skepticism by building positive rapport and trust with reviewers and the ultimate readers of their work. I conclude by discussing the review process where I offer suggestions about how reviewers and editors might adapt to this changing landscape. I specifically argue that all studies are flawed. Rather than asking for a single study to do more to address small inconsistencies or puzzling results, I suggest gatekeepers in the review process should consider the possibility that publishing and allowing research conversations to flourish might result in greater knowledge generation over time.

KEYWORDS:

1. Building a body of evidence
2. Empirical research
3. Trust
4. Robustness
5. Authoring
6. Reviewing

INTRODUCTION

When you submit an empirical manuscript to a journal, you enter an asynchronous dialogue with a review team composed of an action editor and a set of anonymous reviewers. In doing so, you have two primary tasks. First, you must convince the review team that your ideas have merit. There is some debate about what this entails. The oft cited Davis treatise "That's Interesting" argues that papers should be interesting, novel, or counterintuitive and that they should create a "movement of the mind" by confronting taken-for-granted assumptions (Davis, 1971). Others stress the importance of addressing problems that are relevant to society while noting that an extreme focus on novelty might predispose our research toward chance findings and encourage troubling behavior like p-hacking and HARKING (hypothesizing after results are known) (Bettis *et al.*, 2016; Tihanyi, 2020). As Bettis et al. note, it seems odd that the world would only be arranged so "that all phenomena of research importance are counterintuitive (p. 260)". Rather, they argue, we should "return the word interesting to its standard English language meaning of something that you want to learn more about (p. 260)." To this end, Huff (1999:3) conveys the importance of positioning research within one or more ongoing scholarly "conversations," while Colquitt & George (2011) proffer that research should address big problems, change conversations, encourage new discussions, and provide insights for practice.

I'm of the opinion that generating interest is part of the creative process of research. It's hard to fully define, up front, all the ways something can be interesting, but we all know it when we see it. In a classic case of equifinality, there exist numerous creative ways to generate interest from readers (and, before that, reviewers) but, without it, your chances of success fade quickly as reviewers find conceptual cause to recommend rejection before even considering your empirical

efforts. Nevertheless, this first imperative is conceptual in nature, focused mostly on the topic, framing, and motivation for a study.

Should you succeed in the first task of generating interest—a topic which receives ongoing attention from editors and seminal scholars alike—consideration moves to the second task. Here, your aim is to convince reviewers that your empirical results are robust, that they support your conceptual arguments and hypotheses, that the related inferences reasonably reflect what can be concluded from your data and analyses, and that these likely reflect the state of things in the population of interest. Frank (2000), the progenitor of the increasingly ubiquitous impact threshold of a confounding variable (ITCV) test, often refers to this as engaging in conversation with skeptics of empirical work. Frank surmises that it is incumbent on authors to assume readers are skeptical of the study's methodological procedures and to conduct (and explain) sufficient empirical techniques to assuage that trepidation.

The purpose of this chapter is to further address this second task, which has received far less attention that its initial counterpart. In doing so, I will offer the concept of "*building a body of evidence"* as your objective in this dialogue with methodological skeptics. No research project is perfect. No set of conclusions can be supported uniformly by every empirical test. Rather, your goal is to present a convincing body of evidence to persuade readers that your results are robust, and that you questioned your empirical inferences by subjecting supportive findings to the same scrutiny as you would if the results didn't work out. Doing this conveys your findings are trustworthy and likely reflective of something happening in the world. Here, the creative process continues, and the savvy scholar can use descriptive statistics, multiple tests, alternative specifications, and other means to craft a robust story that builds further interest in a paper

through the strength and comprehensiveness of the empirical evidence. Done well, it creates a dialogue with reviewers and, once published, the readers as well.

Those who have been through the review process might quibble with the characterization of a paper submission as the "beginning of a dialogue." As the "gatekeepers of science" (Crane, 1967:195), reviewers clearly have an active voice in the process. They are chosen from among a pool of experts in the field and, as Frank (2000) points out, they generally assume the role of a skeptic who must be convinced through a collection of persuasive evidence that the arguments merit publication and empirical findings are legitimate. Reviewers exercise a certain privilege where they get to tell authors what they like and what they don't like without ever having to directly face the author(s) they are assessing. Unless the manuscript is given a revise-and-resubmit, authors don't have much of a chance to engage in a "dialogue" by responding to criticism. If a paper is accepted, authors become known while reviewers enjoy anonymity. The potential problems within this process are well documented (Miller, 2006; Starbuck, 2003).

At first glance, I can see doubters of my view instinctively making the case that there's no "dialogue" in this process, but hear me out. My argument is that as reviewers digest a paper, they are stimulated by the writing to ask a series of seemingly rhetorical questions. While they turn the pages of your manuscript, reviewers repeatedly ask "but what about…", "why didn't you….", "but did you consider…", or "what happens if …". I say "rhetorical" because, while reviewers (and later, readers) would like an answer, the authors are not physically present to offer one. Knowing that reviewers engage in such self-dialogue, an astute scholar can try to anticipate these questions and use them to assemble a set of logical arguments, alternative specifications, and supplemental analyses that effectively addresses those rhetorical questions *as the reviewer raises them silently in their own mind.* Done successfully, an author can have an

asynchronous dialogue across time and space with reviewers. Yet, doing so requires that authors have some foresight created largely through a willingness to solicit critical feedback and be critical of one's own work.

To be clear, however, this chapter is not about simply piling up multiple statistical tests to overwhelm reviewers. They will see through that and doing so at best proves superfluous and cumbersome for all parties involved. Further, this is not about extensive use of jargon, technical terms, or seemingly advanced methodological wizardry to make a point. Of course, many papers call for complex statistical methods and they should be used (and clearly explained) when needed. However, their use can, at times, create more problems than they solve (Certo *et al.*, 2016; Semadeni, Withers, & Trevis Certo, 2014). Rather, building a body of evidence is about telling a story through descriptive statistics, formal empirical tests, alternative specifications, supplemental tests, and examples that flow logically from your conceptual arguments.

These ideas were primarily developed through my own journey trying to publish my research, sparring with reviewers (who are, more often than not, correct), through work as an associate editor, and by collaborating with and teaching applied econometrics to doctoral students. In developing these ideas, some colleagues insisted much of it should already be well understood by most strategic management scholars. Perhaps that's true. On the other hand, considering how many scholars misunderstand or have difficulty explaining what a p-value is (Aschwanden, 2015) or fail to grasp the differences between fixed- and random-effects models (Certo, Withers, & Semadeni, 2017), among many other seemingly innocuous and ostensibly well-grasped empirical procedures, I suspect there is a considerable gap between what should be commonly understood and what actually is. In my experience, I find many scholars, both new and seasoned, fail to take advantage of many of the ideas outlined here when submitting their

work. Too many accept statistical significance as "proof" of support for their ideas and hastily push toward submission never considering if that result was luck or the work of statistical anomalies traced to outliers or coding errors. Thus, while the primary audience of this chapter is likely to be those who are relatively new to the field of strategic management (doctoral students and newer faculty), it is my hope that many of the ideas will resonate with and be useful to more seasoned scholars as well.

Moreover, and perhaps even more crucially for our field, I hope these ideas can inform all of us as reviewers. No study is perfect. Rather, every study has lingering weaknesses or slight inconsistencies. Most papers conflict, to some degree, with prior theory or empirical findings. Often there's a result that doesn't quite line up with the rest of the empirical evidence. Research should be judged on the collective merit of the arguments and evidence offered, with some level of acceptance that flaws and inconsistencies will always exist. A body of evidence can paint a broad picture about what might be. Within that body of evidence, inconsistencies or weaknesses create avenues for continued dialogue (Huff, 1999), new research, and eventual breakthroughs (Hollenbeck & Wright, 2017). From this perspective, then, a secondary audience of this chapter might be the more seasoned scholars serving as reviewers or editors and in the privileged position of determining what gets published. Here my hope is that these ideas serve as a counterbalance against our natural tendency to nitpick every small flaw while failing to see how a given study builds on past conversations while opening the door to new ones yet to come.

A BODY OF EVIDENCE

What is a "body of evidence"? Borrowing from criminal justice, when pursuing a criminal case, a prosecutor must present evidence that proves, beyond a reasonable doubt, that a defendant is guilty of a crime. In pursuing this work, the astute prosecutor must identify and lay

to rest any alternative and reasonably plausible theory about the events that occurred. This can be accomplished by presenting substantial physical evidence establishing critical facts of the case, the use of witness testimony, and the opinions of experts that might speak to the plausibility of alternative explanations or the chances for error when using scientific techniques to analyze evidence.

As an example, imagine a case of theft where a suspect's fingerprints were found at the scene of a crime where valuables were stolen. Simply demonstrating the existence of the defendant's fingerprints is probably not sufficient to convict. One must show that items were actually stolen and not lost, that it was likely the defendant who stole them, and that it is unlikely the crime was committed by someone else. An astute defense attorney, knowing many other fingerprints were present at the scene, will raise this issue to cast doubt. To counteract this, a skilled prosecutor might anticipate the defense will raise this issue. Rather than wait, the prosecutor might proactively demonstrate they belong to individuals known to be routinely at the scene while also documenting that each has an alibi that reasonably eliminates them from consideration. In doing so, the prosecutor is showing that they questioned their own case.

The parallels aren't perfect. In court, testimony happens in real time and witnesses from each side can be immediately cross examined. In attempting to publish a manuscript, the process is asynchronous, and authors get to respond formally only if a revise and resubmit is offered. Still, an author is akin to the prosecutor presenting a case to the jury (action editor) while a reviewer is the defense attorney for science – a guardian for truth aiming to keep bad science from entering the hallowed pages of our respected journals. A reviewer's job, then, is to identify pockets of reasonable doubt in a manuscript. An author, in turn, must present a body of compelling logical arguments and empirical evidence that provides ample support for the claims

made. If an author can effectively foresee the lines of inquiry that are likely to come from reviewers (like the astute prosecutor above), authors can also participate in the cross examination by proactively answering reviewer questions as they turn the pages of a manuscript.

It is not entirely clear to me what the standard of evidence should be. Depending on the maturity of the research topic under consideration and how the findings might affect the health, welfare, or safety of the people, businesses, society or other relevant entity that is the focus of the study, one could easily argue for a standard of "more likely than not" a hypothesis is true, "beyond a reasonable doubt", or some other standard. Of course, setting that standard is under the purview of reviewers and editors. But, what seems clear to me is that one cannot simply offer a hypothesis, report the result of a single regression model with stars in the appropriate place documenting a p-value of less than 0.05, and then declare victory. More is needed, and this can be accomplished by building a body of evidence.

Below I will discuss some of the elements that can encompass a body of evidence. As summarized in Table 1, these include descriptions of the sample and measures, descriptive statistics, formal hypothesis tests, alternative specifications, supplemental analyses, and examples. Notably, papers do not need to include all of these nor is this an exhaustive list of possible forms of evidence. A body of evidence should be assembled to tell a story, build from existing research conversations (Huff, 1999), flow logically from the theoretical arguments and formal results, and can include an array of creative approaches not covered here. Before getting into the various forms of evidence, I address why formal tests, alone, are often insufficient evidence to fully support theoretical arguments. I then discuss the various forms of evidence. To conclude, I discuss the value of having this dialogue with reviewers and offer suggestions for

how an author might gain the foresight needed to anticipate review questions before submitting a paper for review. I also briefly address how reviewers might think about this process as well.

==========
Insert Table 1 about here
==========

INSUFFICIENCY OF FORMAL TESTS AND THE FOLLY OF OVERINTERPRETING P-VALUES

Recall that, for a null-hypothesis test, a p-value is simply the odds of finding a relationship as large as was found (or larger) in a given sample when there is no corresponding relationship in the underlying population (Bettis *et al.*, 2016; Kennedy, 2008). Many misinterpret p-values as one or all of the following: the odds a finding is wrong; or one minus the p-value as the odds the finding is true; or an indication of the relative strength of a finding (p=0.01 is "stronger" than p=0.05). These are all commonly applied and wildly incorrect. To further emphasize this, run the Stata code available in Appendix 1. The code generates 100 random "y" variables and 100 random "x" variables and then generates 100 regressions where a single x (x1, x2, x3…x100) is used to predict the corresponding y (y1, y2, y3…y100). Imagine each x-y pair is a random draw from some large population of interest. Each randomly drawn sample includes 100 observations. In the final step of the Stata code, the betas and p-values for x variable in the 100 regressions are captured and tallied. The final line outputs a value tabulating the number of times out of the 100 models when the p-value for the x beta was less than or equal to 0.05.

Given the data are randomly generated, the "true" relationship in the underlying population is zero. Yet, by definition, the expected outcome here is five. That is, we'd expect five cases out of 100 where the p-value is less than or equal to 0.05 even if our data were purely random. If you run this code repeatedly (say 100 times), the average number of cases with p-

value less than or equal to five will be approximately 5 but individual runs will return values that range well above and below 5. Why does this happen? When randomly generating a sample from the population, sometimes through dumb luck we get a sample that shows a relationship that doesn't exist in the population – a Type 1 error. If our standard is a p-value that is less than or equal to 0.05, then, even with completely random data, we will see statistically significant coefficients approximately five percent of the time. With a critical value of 0.10, the expectation would be 10. In the version I ran as I am typing this, there were 8 with p less than or equal to 0.05 and 12 less than or equal to 0.10 (see Table 2 for a list of these cases). The two lowest p-values were 0.000 and 0.005 corresponding to betas of -0.389 and +0.297 respectively. If the two lowest p-values provide "significant" estimates for the population that are similarly sized in magnitude but in the opposite direction, it should be clear why the p-value cannot speak to the "strength" of a result.

==========
Insert Table 2 about here
==========

Understanding how this simple example applies to research highlights a critical reason why we need to build a body of evidence in our papers. With each study we complete, we take a scoop of randomly selected data (we hope it is randomly selected, and for sake of argument here, let's assume it is) from an underlying population and compute statistical tests from that sample. Using the results of these tests on the sample, we estimate or infer—why it's called inferential statistics—what the value of that parameter is in the population. No empirical estimator is perfect. Each has numerous assumptions, and we almost always violate at least some of them. Even if we didn't, it's important to acknowledge that our fancy statistical tools only provide estimates of what is occurring in the population along with a confidence interval of that estimate.

But this is not 100% certainty. That is, there's no guarantee the real value lies within the specified range because estimates will offer false positives a non-trivial portion of the time.

As illustrated with my sampling distribution example via the code in Appendix 1, if the 100 pairs of x and y represented 100 scholars simultaneously but independently pursuing similar research questions with a randomly chosen sample from the same underlying population, we would expect about 5 of them to have supported results with a p-value less than or equal to 0.05 even if the relationship didn't exist. If one "significant" test was enough, we would have to concede that our journals are disproportionately filled with the lucky few who found the spurious or random results supportive of hypothesized relationships that do not exist in the population (Goldfarb & King, 2016). We can begin to overcome this problem by building a body of evidence. Authors must accept that reviewers are rightly asking, at least in part, "was this paper one of the 5-in-100 that found a result from luck or is this result real?" Of course, in a cruel bit of irony regarding what it takes to be interesting, with increasingly counterintuitive or novel hypotheses (Davis, 1971), we would expect a reviewer to be increasingly skeptical.

Sadly, p-values have been given "almost mythical properties far removed from the mundane probabilistic definition" (Bettis *et al.*, 2016: 259) that underlies their correct meaning. A p-value of 0.049 is taken to mean that a corresponding hypothesis is "true" yet one that is 0.051 means it is most certainly "false." Of course, in most studies of modest sample sizes, these two p-values are certainly not practically or even statistically different from each other. Yet, in many cases, one gains the scientist a publication while the other goes into the file drawer. While beyond the scope of this chapter, alternative approaches do exist. Replications allow us to assess the validity and boundary conditions of prior work, and Bayesian analyses allow us to calculate the degree of belief in an outcome based on our prior knowledge of the topic. Further, the

random paper that gets published when the results were spurious should not doom science to the fate desired by skeptics. As Bettis and colleagues (2016) note, "one study proves little or nothing…Instead, it establishes initial confirming evidence" (p. 260). In assessing the true nature of the world, we should assess the collective body of scientific evidence. For this, meta-analytical approaches allow scientists to combine multiple studies which then provide estimates of relationships that have narrower confidence intervals.

One final point on p-values. The role of theory here is critical. Absent theory, a finding with a p-value of 0.05 should be interpreted as I have described above. However, given plausible theory and a p-value of 0.05, we now have a conditional probability, an idea on which Bayesian analyses are fundamentally built. That is, given a theory is plausible and a low p-value (e.g., less than or equal to 0.05), one can place more confidence on the research finding than would be the case without the theory. Thus, offering compelling theory, though not addressed below, can be considered a part of the body of evidence, and should spur greater confidence in our results.

FORMS OF EVIDENCE

In the research context, we build a body of evidence by presenting a compelling case that demonstrates a robust link between our conceptual arguments or hypotheses and the corresponding empirical tests. As discussed above, this evidence should tell a story and flow logically from the arguments of the manuscript. Below I discuss several possible forms of evidence and how they might be used. I start with the sampling process, measures, and descriptive statistics. These are critical as they form the foundation of a study and provide an initial lens into the reliability and transparency of a researcher's efforts. From there, I expand to various empirical approaches that can be included in a body of evidence. This should not be seen

as an exhaustive approach, as scholars can and should find creative ways to showcase their ideas and support their work.

Sample and Measures

Most "Empirical Methodology" sections begin with a description of the sample used in the study followed by variable descriptions. Given that the inferences for a population are based on the sample, it is critical that researchers clearly document how the sample was formed, what data sources were used, inclusion or exclusion criteria to make an initial sample, and why any data were removed from an initial sample on the way to a final sample (e.g., random selection, missing data, or exclusion based on theoretical grounds). For example, studies in strategic management focus on specific years and include some industries and not others. Financial services firms report financials in fundamentally different ways and are often excluded from studies as a result. Some studies exclude new firms or firms smaller than a certain size. Still others only focus on firms in a particular stock index (e.g., S&P 500 or 1500). It's critical that scholars clearly articulate the relevant population of interest. Then, they must document sampling criteria the generated the final sample for a given study. This should include specific counts of cases at the start, how many are dropped in each step, and the final count. Counts should also be provided for any relevant groupings within data. If the study is on CEOs, then counts of unique CEOs, unique firms, and total rows of data should be included. Adding additional sample descriptions such as number of industries, firms by industry, and similar items can be useful as well. All of this allows a reader to understand the nature of the sample relative to a broader population.

In accounting for the sample, it is especially important to address missing data. If your sample is the S&P 500 for a 10-year period, it would be expected that you would have roughly

500 x 10 or 5,000 firm-years. If your final sample includes only 4,000 firm-years and no explanation is offered for the "missing" 1,000 records, it shouldn't be surprising when reviewers react with skepticism. If the base of your sample is Compustat or other commonly available dataset, it is reasonable that a reviewer might even download the dataset and see for themselves what sort of sample they can generate following your description. Any unexplained large discrepancies in sample size will make reviewers increasingly skeptical of everything else.

Related to this, each measure used in a study should be clearly described. If using existing measures, share enough to demonstrate it was done the same way as the original. If there are deviations from prior work, this and the rationale for deviating must be clearly documented. Finally, it is critical that measures convincingly capture the underlying construct they claim to represent. Notably, if a measure is relatively new, and sometimes even if it isn't, construct validity can still be called into question (Boyd *et al.*, 2013; Gove & Junkunc, 2013), so the most prudent authors should offer their own arguments or evidence demonstrating validity.

Descriptive Statistics

In a typical manuscript of 40-pages, the correlation table might take up a full page or roughly 2% of the overall length, yet it is often almost entirely ignored in the text of a paper. That is, the typical "Table 1" of a study is generally mentioned just once at the start of the results section with a sentence such as: "Table 1 displays correlations and descriptive statistics of our sample." But, with this table, authors have a chance to demonstrate the reliability of their data and offer the first bit of evidence in support of their work (or address the first line of potential criticism). First, to generate confidence, researchers can demonstrate that certain values are consistent with prior work. For example, if there is a firm performance measure, its mean and

standard deviation (as well as minimum and maximum if reported) should be consistent with prior work with comparable samples.

A similar approach can be taken with correlations. If a study including measures of firm performance and an indicator for CEO termination provides a correlation that is positive (e.g., suggesting good performance leads to termination), reviewers can and should be skeptical. Moving to the hypothesized relationships, if there is a proposed negative link between x and y, one would generally expect there to be at least a modest negative bivariate correlation between these variables as well, assuming unexplained heterogeneity wasn't so pernicious as to invert the direction of the estimated association. As such, one might start the results section of a paper by highlighting such a correlation. If it doesn't exist (or if the relationship is inverted), the author is well-advised to begin addressing the complex relations that create a scenario where the general relationship is positive, but the relationship of interest is still negative. For example, Simpson's paradox (Simpson, 1951) highlights how it is possible for a relationship to exist within subgroup but not exist (or exist in the opposite direction) across groups. Not directly discussing this might doom your paper immediately in the eyes of an observant and skeptical reviewer. Of course, this creates a link to the formal test where, if this paradox existed, one would also have to account for grouping or nesting of data through, for example, multi-level modeling.

Next, one might present additional descriptive statistics relevant to the relationship at hand. Perhaps the hypothesis entails a binary x variable predicting a continuous outcome such that the x=1 group is expected to have higher values of y than the x=0 group. One can demonstrate this by reporting simple group means for y under the two conditions of x, a common practice in many other disciplines. Alternatively, one could graph the values of y to visually show how one group has consistently higher values than the other.

As an anecdote from my own work, my co-authors and I hypothesized that CEOs hired from outside the firm would deliver more extreme performance than insiders (Quigley *et al.*, 2019). Our first bit of "evidence" demonstrating this was a pair of histograms showing a wider and flatter distribution of outcomes for outsider CEOs and a narrower and taller distribution of outcomes for insiders. If the hypothesized x-variable is continuous, this same sort of descriptive statistics or visual depiction can be created by dividing the x-values into low, medium, and high groups and reporting means or displaying graphical distributions for these groups, or even by producing binned scatterplots (Starr & Goldfarb, 2020). This approach allows a reader to visibly see the claimed relationships in the underlying data before going into more complex analyses. Notably, one could also use this approach to address the issue of Simpson's Paradox. If the relationship is believed to be negative within group but the bivariate correlation in the sample is positive, then graphing by group or reporting the proportion of times the within group correlation was negative could begin to address the issue.

Formal Tests and Effect Sizes

After these initial descriptive results, one might move to reporting results of formal hypothesis testing using the analytical techniques described in the methods section of a manuscript. While substantial effort might be needed to demonstrate that the model assumptions hold (e.g., reporting instrument validity tests for two-stage modeling or demonstrating reasonable evidence of parallel trends for difference-in-difference modeling, among others), the actual process of reporting results is straightforward. If your hypothesis claimed a negative relationship between x and y, this is demonstrated with a negative coefficient and some level of statistical significance. However, authors can build additional evidence by reporting confidence intervals and, as is now required in *Strategic Management Journal*, by demonstrating the practical and/or

economic impact of the reported relationship. If the claim is that increases in your x variable results in a reduction in market valuation in firms, for instance, you can build your body of evidence within the formal econometric test by documenting the average amount of decline that is associated with a one standard deviation change in your independent variable.

You should also consider reporting marginal effects—which represent the differential relationships between the independent and dependent variables over different values of relevant variables—numerically and via graphs, especially when hypothesizing interactions (Busenbark *et al.*, 2022a). Simply noting that an interaction was "significant" does not provide a clear picture of the actual relationship, but a graph can. Further, calculating and reporting marginal effects allows an author to describe the size of impact at various levels of the x-variable and moderator while noting if the differences are statistically significant and/or practically meaningful. One pitfall that can be easily avoided with this approach is an interaction that is significant but not meaningfully impactful in a reasonable range of the data. For example, perhaps CEO tenure moderates the link between some CEO trait and firm performance. But if the difference only becomes significant or economically important in year 37 of a CEO's tenure, it becomes hard to claim any sort of practical impact given how few CEOs serve that long.

Alternative Specifications

Next, you can consider alternative specifications. That is, in justifying the use of a particular estimator, it is likely that a reasonable case could be made for one or more alternative analytic procedures. For example, while it is common to use random-effects estimators when there is no within-group variance on a key independent variable (for example, when CEO prior experience, education, or gender is a variable of interest), some reviewers might prefer generalized estimating (GEE) equations or multi-level modeling (MLM). Executing these tests

and stating, “Our results were robust to the use of GEE and MLM as well (results available upon request)” generates considerable confidence, especially for a reviewer who might prefer those approaches. Just be sure to properly save those alternative models in the case someone asks for them (and in the review process, you might include them in a response letter or as an appendix).

In some cases, however, you might choose a particular estimator over another commonly used approach because the assumptions of the first are more completely satisfied while there are compelling violations with the other. In these cases, you should rightfully reject the less appropriate tool and explain the choice. But it might also be useful to consider its use to study boundary conditions where changes in the underlying assumptions might yield nuanced findings that prove useful to the field.

Authors can also consider different measures, changes to the sampling approach (including smaller firms, different industries, or additional time periods), shifts in lags for calculating variables across time, or other alternatives. Each of these can demonstrate the robustness and external validity of your results, showing the skeptic that you are questioning your own choices and thus providing some assurances that you didn’t just pick the one model that supported hypotheses out of myriad that didn’t. For example, if the initial sampling frame included removing firms with less than $100 million in revenues, then an alternative model might probe if the results are robust to a cutoff of $500 million or $50 million. Or, if a primary variable is measured as the number of acquisitions in a two-year period, one might test if the results are robust to measurement across one-year or three-years. So long as these more exploratory analyses are conducted openly (rather than repackaged as *a priori* hypotheses after finding the result), they should be encouraged in the spirit of “THARKING” (transparently

hypothesizing after the results are known) rather than SHARKING (secretly hypothesizing after the results are known), as discussed by Hollenbeck and Wright (2017).

Alternative Explanations and Supplemental Tests

You should also step back to consider alternative explanations for your findings and offer empirical tests that attempt to rule these out. Imagine a study linking the personality trait of extroversion in CEOs with a particular firm outcome. Rather than being randomly assigned, it is conceivable that these results were driven by CEOs attracted to a certain opportunity or that they were recruited by the board because their personality was believed to be ideal for navigating the firm a specific way in the current task environment. Or, as research has shown, a prior CEO might be inclined to advocate for the hiring of a new CEO that is similar to themselves (Zajac & Westphal, 1996). The primary test of this relationship will likely include some sort of model to deal with this possible endogeneity. But reviewers may remain skeptical even in the presence of compelling findings from this primary model. Here, creative thinking might allow you to test alterative explanations in hopes of assuaging these reviewer concerns. For example, you might be able to demonstrate variance in the personality of CEOs within a firm (specifically showing that the personality of the outgoing CEOs is different than incoming CEOs), that CEOs coming to firms facing similar internal and external environments exhibit different personality traits and diverging paths once in office, and that the personality traits of CEOs, as measured through archival sources, remain consistent to repeated measurement over time.

None of these tests can "prove" that reviewer concerns are invalid. Still, offering robustness tests, supplemental analysis, and tests of boundary conditions builds the body of evidence and can go a long way toward assuaging reviewer concerns. In some cases, these efforts might yield novel findings that spur further conversations and research. The key here is to

step back and consider what a reasonable reviewer might be asking, rhetorically, as they read your work. And rather than let the question go unanswered, you offer an answer to them as they turn the pages of your manuscript.

Examples and Anecdotes

Some scholars take the approach of using mixed-methods to study a phenomenon (e.g., Eisenhardt, 1989). This entails linking two formal studies in a single paper—often one qualitative and the other quantitative. While laudable, doing so is a challenge for some topics and the associated complexity can be limiting for some scholars. At the same time, one can borrow from this approach as a means to generate some realism and support for empirical claims. As a reviewer and associate editor, I have often asked authors to "show me how this looks in the data with a real-world example." What I am requesting is a case where a firm experienced the relationship proposed in the study. Imagine a scenario where one is hypothesizing a within-firm negative relationship between a form of strategic action and future performance. You might document this in a few firms by showing actual data highlighting the expected pattern of strategic choices and then showing the measures for the outcome. If this can be done with anecdotal accounts from news coverage, that would be even more compelling.

Similarly, you might consider using quotes to show plausibility. That is, imagine a study that argues for a particular CEO mindset leading to a certain type of strategic decision. The study might use content analysis of earnings calls to assess CEO mindset and archival accounting performance data for the outcome. To add further evidence, you might find quotes from an interview of a CEO clearly demonstrating the mindset prior to a decision you expect is associated with that mindset. Done well, a researcher can demonstrate the existence and plausibility of a phenomenon using concrete examples. Though I mention this item last, I believe

this approach can be used throughout a paper to demonstrate theory in action which can help readers more fully and clearly understand conceptual arguments.

Other Issues and Creative Approaches

Endogeneity is among the most common concerns for strategic management studies (Hamilton & Nickerson, 2003). While instrumental variables and various forms of two-stage models allow authors to deflect some of this criticism, authors can further address this through one or more techniques recently introduced to the management literature. One approach is the impact threshold of a confounding variable test (ITCV) (Busenbark *et al.*, 2022b; Frank, 2000). With endogeneity, there is unmeasurable confounding variance related to both the independent and dependent variables which, if included in a model, might negate a result. The ITCV allows one to quantify the size of the correlations needed between potentially missing (and unmeasurable) variables and both the key variables to invalidate the results. By comparing the size of this correlation with the size of known correlations between variables that are in the model, it is often possible to make the case that an omitted variable with the needed correlations is quite unlikely (Busenbark *et al.*, 2022b).

There are other approaches that offer sensitivity analyses that are similar to the ITCV. Oster (2019), for example, offers a test that calculates a bias-adjusted estimate given the impact of control variables and level of unexplained variance. Similarly, Cinelli and Hazlett (2020) offer a tool that "shows how strongly confounders explaining all the residual outcome variation would have to be associated with the treatment to eliminate the estimated effect" (p. 39). The point is, even after addressing endogeneity (or perhaps in cases where instruments for addressing endogeneity simply aren't available), these tools allow you to go a step further and quantify how large the problem would have to be to invalidate results. While you can never eliminate the

possibility that confounding variance would invalidate an inference, these tests can at least help quantify the hazard.

Outliers are also problematic for strategic management research, especially in studies using archival financial data. Specifically, reviewers often wonder if a small number of influential observations might be driving results. To demonstrate this, consider the code in Appendix 2. It generates 1000 random values of x and y and then regresses y on x. Like the earlier example, since this is purely random data, we would not expect a significant coefficient. But, due to chance in the sampling process, roughly 1 in 20 attempts will generate a p-value for x that is less than or equal to 0.05. But, if just a few outliers exist in this otherwise random data (in this case I define just four additional cases), the coefficients for x have p-values less than or equal to 0.05 almost every time. While Winsorizing and other transformations remain common but fallible remedies, authors might consider building on their body of evidence through others means. If a relationship is expected across the range of x and y values, consider dropping a small number of cases at the extremes of key variables to see if reported results remain robust. If a relationship only exists when the four most extreme cases are included, it is probably warranted to reconsider the strength of the empirical evidence.

ANTICIPATING QUESTIONS AND OVERCOMING SKEPTICISM

The Benefits of Building a Body of Evidence

If you successfully anticipate questions asked by reviewers, it is useful to then consider what this achieves. You, as an author, have spent months or years painstakingly developing theory, crafting hypotheses, building your dataset, and completing statistical analyses. If this research project is truly a new contribution to the literature, it is likely no one else in the world knows as much about the nuances of the topic and the underlying data as you (this should

especially be the case if this effort is a product of your dissertation). While reviewers might be experts in the general domain, it's unlikely they have the same knowledge as you, the author, about this particular topic, setting, or the key concepts at the center of your work.

Imagine a reviewer gets deep into your manuscript and asks the question "but did you consider <fill in the topic>?" The reviewer then flips the page only to find a passage that says, "…of course a reasonable criticism of our work is that <fill in the topic> might be affecting our results. To address this concern, we…" As the author, you might go on to say, "Building on this issue, it's also important to consider these three related issues, A, B, and C which we also addressed as follows…" Imagine the response of a reviewer at that point. They probably smile and even relax a little bit. They feel some pride in catching an issue that you quickly address and nod in agreement when you note the three additional issues related to the first. Why? While you, the author, have spent months or years on this project, the reviewer is maybe 30 minutes or an hour into it. Yet, by anticipating and answering the reviewer's question, you are engaging them in a powerful way even though the communication is asynchronous. In answering the reviewer question as it was asked, you are tacitly acknowledging the reviewer's intellect. You are also assuring them that they understand where you are going, thus leveraging the reviewer's confirmation bias to your benefit. This engagement creates positive rapport between you and the reviewer at a critical moment of the review process. In the moment skepticism was going to creep in, you stunted it by effectively anticipating and then answering the questions. The affective, explicit, and subconscious benefits of this asynchronous conversation cannot be overstated.

Anticipating Reviewer Questions

But how do you anticipate the multitude of possible questions reviewers might ask so you can address them proactively? You might even think, “sure, I can think of 3 or 4 things, but reviewers seemingly invent a thousand reasons to hate my work.” As an author, I can assure you I’ve felt that exact emotion before. But, as an associate editor, one of the more compelling realizations I’ve had is how often reviewers have considerable overlap on the major concerns that result in a paper getting rejected. My argument is that you can anticipate these, perhaps with some help, before submitting your paper.

The first step here is to critically evaluate your own work to identify potential weaknesses or alternative explanations and then addressing them as best you can. To aid in this process, scholars should marshal assistance from a variety of sources. For example, simply talking with colleagues about your work over coffee or lunch can often generate multiple lines of potential inquiry. Doing this while you are developing your research plan allows you to incorporate the feedback into the study design. Data needed for this can then be gathered up front rather than afterwards when it is more costly to do so.

As you develop the manuscript, you can gain insights from peers through presentations in seminars or informal “brown bag” sessions in your own department or during visits to other universities. The value here can be multifaceted. First, committing your project to a presentation forces you to fully explain your ideas in ways you have yet to do, and this might allow you to uncover previously unseen weaknesses or omissions in your conceptual arguments or empirics. Second, sharing these ideas to a new group might uncover novel ideas to enhance your study, such as additional moderators, novel measures, or important controls. These audiences might also raise important questions that can be addressed using some of the types of evidence discussed above (particularly alternative specifications and supplemental analyses). Addressing

these now adds to the likelihood that you might answer one of the eventual reviewers' questions thereby eliminating a potential line of criticism.

Perhaps you are at a point in your career where it is unlikely you will receive an invitation to present or in a department where internal presentations aren't common. If so, propose to create such a forum and offer to be the first presenter. If you are still a student, work within your cohort at your university and present informally to one and other. You might even seek out those in adjoining fields to join you. Students and junior faculty can also create informal groups of colleagues from other universities and meet virtually a few times a year via Zoom. I have personally benefitted from arrangements such as these long before I was ever invited to present at another university.

Once a manuscript is fully developed, submission to conferences creates an opportunity for even more feedback. Moreover, in this setting it is likely that one or more future reviewers of your work might be in attendance when you present. Imagine getting specific feedback at a conference, addressing it in your manuscript, and then that person becomes Reviewer 2 on your submitted paper. Imagine this reviewer recalls the paper, remembers giving the comment, and then seeing this issue addressed based on their ideas. It seems this could only help the odds of getting a revision opportunity. Even the prototypically-menacing "Reviewer 2" is likely to be gratified by seeing their own idea in place. It could also be the case that this reviewer doesn't remember the presentation or giving the comment but, given how our brains work, that reviewer is likely to have the same criticism reading your work as they did in your presentation. Addressing their comment will still have a compelling positive affect. Of course, this does not mean you should implement every idea you are given. The discerning scholar needs to be the expert within their domain and pick and choose which feedback to address accordingly.

Another step of gaining insights to lingering weaknesses in your manuscript is the friendly review. Once a paper is, in your mind, ready for submission to a journal, you then send it out to peers for friendly evaluation. This would preferably be scholars familiar with the broad theories you employ, but not familiar with your specific study. At this point you might think "I've spent 500 hours on this project, shared it with peers at brown bags, presented it at conferences, and poured over the results and writing for dozens of revisions…I need to get it under review, and can't afford another month of delay." Yet, the value of gaining feedback from respected scholars is tremendous. Imagine if you could send your paper to a journal once, get feedback, address those issues, and then send it back to that journal again with a fresh start. How much might this improve your prospects? This is essentially a friendly review, especially if you can lean on friends who routinely review for the journal you are targeting.

Let's say a friendly review spots a few issues you can address with some additional empirical tests and by simply clarifying your writing in a few places. Imagine this adds a month to the process but makes the difference between a reject and an R&R. Is it worth it? Of course, and this is why all the most successful senior scholars I know, including those with dozens of A-level publications, always share papers with friends before sending out to a journal. Just make sure you pick friends who agree to give you quick feedback and who aren't afraid of telling you the truth. The more effective you can be at anticipating questions from reviewers and providing answers to those questions in the manuscript as part of your body of evidence, the better your chances are with the review team.

One final thought that can really enhance your body of evidence. Consider sharing some or all your data and/or code. In the coming years, some form of this will probably become the norm in many of our journals. Scholars who get into the habit of conducting their research such

that their code and data are properly organized and ready to be shared will have an advantage. Until then, those who willingly share even portions of code or data demonstrate a commitment to open science and transparency that generates more confidence in the body of evidence than any other example offered here.

JOURNAL AND REVIEWER CONSIDERATIONS

Online Appendices

While incorporating any one of these suggestions will add negligible length to a manuscript, adding several to create a substantial body of evidence is likely to create challenges adhering the page length limits at most journals. Traditionally, authors were simply forced to cut content to maintain page lengths at something that could be published. Today, however, journals, editors, and reviewers are more welcoming of appendices, intended to be published online, that provide supplemental materials. Such appendices can contain additional descriptions of a complex aspect of your methods, sample details, robustness tests, or other materials. While it would be inaccurate to say that space is now unlimited (for example, reviewers still need to assess the veracity of supplemental materials), it is possible to briefly note something in a manuscript while providing more complete details shifted to an online appendix.

Thoughts for the Reviewer

It is also useful to consider the how this plays out from the perspective of reviewers and editors. I started out this chapter by arguing that authors should create a body of evidence while also noting that not every test works out perfectly. Some supplemental tests might return p-values that are above common thresholds of statistical significance. Imagine an author reporting a significant correlation, a primary test with a p-value less than or equal to 0.05, meaningful effect sizes, and a few supplemental tests that are significant at common levels, but also one

alternative test of the primary model where the coefficient p-value is 0.08. Authors should not shy away from reporting these results and, if the estimates remain reasonable, reviewers should take them for what they are: additional support offered as part of a body of evidence. Rather than recommending a rejection based on one test above a the commonly used, but arbitrary, threshold of 0.05, reviewers and editors should consider the full body of evidence. After all, as noted above, it is most likely the case that a pair of p-values, one slightly above and another slightly below common levels of statistical significance, are not meaningfully different from each other.

More broadly, reviewers and action editors should be open to the inconsistencies and messiness of research as a means to extend conversations (and the related research) into areas that help us further understand the phenomenon of interest. If a paper presents results that are otherwise trustworthy and interesting but, in some way, at odds with existing work—or if a paper has strong findings but some unanswered questions that can be addressed by future work—it seems reasonable to support publication so the discussion can continue, rather than reject because of an inconsistency or unanswered question. Reviewers should frequently ask themselves "is it reasonable for a single paper to do yet one more thing" or is it better to allow the conversation to continue by publishing a study knowing that a community of scholars can then wrestle with an inconsistency, conduct additional studies, and, over time, address concerns or open questions that remain. It's a balancing act but, when reviewing, it is a question we should all ask ourselves as an effective counterbalance against the ever-growing expectation that authors do more and more in a single manuscript. Rather than asking for more or rejecting over small inconsistencies or lingering questions, it may be better to publish which allows conversations to flourish through later studies conducted by a wider array of scholars. In my view, so long as the underlying scholarship is sound and trustworthy, this will ultimately lead to the generation of

more knowledge over time. However, if an inconsistency or open question is the result of sloppy scholarship, or if a glaring weakness goes completely unaddressed, it is reasonable to expect criticism and a resulting rejection.

CONCLUSION

In the early days of the field of strategic management, manuscripts were often published that tested hypotheses with a single model that relied heavily on reporting statistical significance via p-values. Effect sizes were often not discussed, and few, if any, supplemental tests or alternative specifications offered. Today, reviewers are demanding greater evidence that a study's empirical findings are robust. In this chapter, I outlined the idea of a "body of evidence" and offered some guidance for how scholars might develop more compelling compendium of results in support of their ideas. The process entails anticipating critical questions from reviewers and answering them as part of the initial submission. In doing so, authors can engage in an asynchronous dialogue that builds rapport with reviewers (and eventual readers) while the resulting transparency generates trust in the research process. Authors can build their body of evidence by focusing a critical lens on their own work and by gaining feedback through discussions with colleagues, presentations, and friendly reviews. It is my belief that pursuing this approach will result in greater chances of paper acceptance for authors, better experiences for reviewers, and stronger science for the field.

REFERENCES:

Aschwanden C. 2015. Not Even Scientists Can Easily Explain P-values. https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/ (2/15/2023 2023).

Bettis RA, Ethiraj S, Gambardella A, Helfat C, Mitchell W. 2016. Creating repeatable cumulative knowledge in strategic management: A call for a broad and deep conversation among authors, referees, and editors. *Strategic Management Journal*: 257-261.

Boyd BK, Bergh DD, Ireland RD, Ketchen Jr DJ. 2013. Constructs in strategic management. *Organizational Research Methods* **16**(1): 3-14.

Busenbark JR, Graffin SD, Campbell RJ, Lee EY. 2022a. A marginal effects approach to interpreting main effects and moderation. *Organizational Research Methods* **25**(1): 147-169.

Busenbark JR, Yoon H, Gamache DL, Withers MC. 2022b. Omitted variable bias: Examining management research with the impact threshold of a confounding variable (ITCV). *Journal of Management* **48**(1): 17-48.

Certo ST, Busenbark JR, Woo Hs, Semadeni M. 2016. Sample selection bias and Heckman models in strategic management research. *Strategic Management Journal* **37**(13): 2639-2657.

Certo ST, Withers MC, Semadeni M. 2017. A tale of two effects: Using longitudinal data to compare within-and between-firm effects. *Strategic Management Journal* **38**(7): 1536-1556.

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

Colquitt JA, George G. 2011. Publishing in AMJ—part 1: topic choice. *Academy of Management Journal* **54**(3): 432-435.

Crane D. 1967. The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *The American Sociologist*: 195-201.

Davis MS. 1971. That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology. *Philosophy of the social sciences* **1**(2): 309-344.

Eisenhardt K. 1989. Making fast strategic decisions in high-velocity environments. *Academy of Management Journal* **32**(3): 543-576.

Frank KA. 2000. Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research* **29**(2): 147-194.

Goldfarb B, King AA. 2016. Scientific apophenia in strategic management research: Significance tests & mistaken inference. *Strategic Management Journal* **37**(1): 167-176.

Gove S, Junkunc M. 2013. Dummy Constructs? Binomial Categorical Variables as Representations of Constructs: CEO Duality Through Time. *Organizational Research Methods* **16**(1): 100-126.

Hamilton BH, Nickerson JA. 2003. Correcting for endogeneity in strategic management research. *Strategic Organization* **1**(1): 51-78.

Hollenbeck JR, Wright PM. 2017. Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management* **43**(1): 5-18.

Huff AS. 1999. *Writing for Scholarly Publication*. SAGE Publications: Thousand Oaks, CA.

Kennedy P. 2008. *A Guide to Econometrics* (6 ed.). Wiley: Malden, MA.

Miller CC. 2006. Peer review in the organizational and management sciences: Prevalence and effects of reviewer hostility, bias, and dissensus. *Academy of Management Journal* **49**(3): 425-431.

Oster E. 2019. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics* **37**(2): 187-204.

Quigley TJ, Hambrick DC, Misangyi VF, Rizzi GA. 2019. CEO selection as risk-taking: A new vantage on the debate about the consequences of insiders versus outsiders. *Strategic Management Journal* **40**(9): 1453-1470.

Semadeni M, Withers MC, Trevis Certo S. 2014. The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations. *Strategic Management Journal* **35**(7): 1070-1079.

Simpson EH. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **13**(2): 238-241.

Starbuck WH. 2003. Turning lemons into lemonade: Where is the value in peer reviews? *Journal of Management Inquiry* **12**(4): 344-351.

Starr E, Goldfarb B. 2020. Binned scatterplots: A simple tool to make research easier and better. *Strategic Management Journal* **41**(12): 2261-2274.

Tihanyi L. 2020. From “that’s interesting” to “that’s important”, Academy of Management Briarcliff Manor, NY.

Zajac EJ, Westphal JD. 1996. Who shall succeed? How CEO board preferences and power affect the choice of new CEOs. *Academy of Management Journal* **39**(1): 64-90.

Table 1: Components of a Body of Evidence

Category	Examples	Purpose
Sample and Measures	• Complete description of sample including accounting for lost observations • Clear description of measures • Support for construct validity even for existing measures	• Transparency, future replication, clear understanding that sample reflects population of interest. • Ensuring measures meaningfully capture constructs
Descriptive Statistics	• Correlations of key relationships • Means with group comparisons • Distributions	• Generate a sense that the data are behaving as expected without fancy, multivariate analyses
Formal Tests and Effect Sizes	• Traditional regression results • Related beta coefficients • Confidence intervals • Marginal effects • Practical or Economic impact	• Provide formal statistical support for claimed relationships and demonstration of magnitude
Alternative Specifications	• Alternative measures • Varied sampling techniques • Alternative estimators	• Provide robustness to demonstrate relationships were not found due to luck or that the relationship is only found in a very particular sample with certain measures that was picked “because it worked”
Alternative Explanations and Supplemental Tests	• Alternative logical tests that must be true (or not) if your results are true • Tests of the limits of your theoretical arguments	• Test boundary conditions and attempt to rule out alternative arguments and causal mechanisms. • Explore novel findings that might generate new conversations
Examples	• Anecdotes • Case studies • Quotes	• Capture the phenomenon of study in the real world • Demonstrate relationship exists and is recognizable
Other creative approaches	• Limited only by creativity of author(s)	• Further tell the story of the paper

Table 2: "Significant" (p<=0.10) Betas and P-values in 100 Regressions with Random Data

Case	p-value	Beta	Confidence Interval	
1.	0.000	-0.389	-0.583	-0.195
2.	0.005	0.297	0.090	0.504
3.	0.014	-0.238	-0.426	-0.050
4.	0.020	-0.237	-0.436	-0.039
5.	0.028	0.205	0.023	0.386
6.	0.031	0.218	0.020	0.416
7.	0.032	-0.193	-0.368	-0.017
8.	0.040	0.184	0.008	0.360
9.	0.068	0.173	-0.013	0.359
10.	0.072	0.162	-0.015	0.340
11.	0.073	0.196	-0.019	0.412
12.	0.078	-0.192	-0.406	0.022

Appendix 1: Stata Code for Random Data Regressions

```
clear
//set the number of observations
set obs 100
//create a variable in dataset to capture betas, p-values, and confidence interval
gen b = .
gen p = .
gen ci_l = .
gen ci_h = .
//loop from 1 to 100
forvalues i = 1/100 {
    gen y`i' = rnormal() //generate y1 through y100
    gen x`i' = rnormal() //generate x1 through x100
    reg y`i' x`i' //run regression predicting y_i with x_i
    mat b=r(table) //save results table
    replace b = b[1,1] if _n==`i' //capture the b-value of the i-th model and save to i-th row
    replace p = b[4,1] if _n==`i' // p-value
    replace ci_l = b[5,1] if _n==`i' //low side of confidence interval
    replace ci_h = b[6,1] if _n==`i' // high side of confidence interval
 }
//count how many p-values are <= 0.05
count if p<=.05
```

Appendix 2: State Code for Outlier Example

```
clear
set obs 100
gen y = rnormal()
gen x = rnormal()
reg y x // significant at p<=0.05 ~ 1 in 20

set obs 104  //add 4 more observations to the data

//Generate the 4 outliers
replace y = 2.5 in 101
replace x = 2.5 in 101
replace y = -3.0 in 102
replace x = -3.5 in 102
replace y = 3.5 in 103
replace x = 3.0 in 103
replace y = -2.5 in 104
replace x = -2.5 in 104

reg y x //significant at p<=0.05 nearly every time
scatter y x
```