Researchers' Data Analysis Choices: An Excess of False Positives?

James A. Ohlson

Professor, Hong Kong Polytechnic University

December 2017

Abstract

The paper assesses the consequences of current ways of implementing empirical data analysis. It focuses on a likely failure: a pervasiveness of erroneous null-rejections. The argument goes beyond criticizing the (generally accepted) practice of researchers' publishing only those results that support the desired findings. It is also the case that this convention interacts negatively with methods of statistical analysis. Specifically, much research relies on powerful statistical tests which by themselves have the drawback of embedding a relatively high probability of a false positive. Supplementary analyses can easily counter this problem via use of statistical methods making null-rejection much harder when the null is true. Yet few papers report on such tests, and their absence implies (at least potentially) a lack of information that would bear on the findings intrinsic robustness. Research conventions can push against these shortcomings, but only if producers and consumers of research are willing to accept equivocal findings: in an inherently complex world, comprehensive data analysis should oblige researchers to often acknowledge that "the evidence supporting our core hypothesis must be qualified because…". The paper explains why researchers instead promote unequivocal findings which enhance chances of false positives.

I. Introduction: The Problem of Getting the "Right Results"

Empirical research centers on the idea that a paper's conclusions can be – and should be— supported by the data and related analysis.[1] Key statistics and how these substantiate narratives must be spelled out as clearly as possible. Researchers also tend to set the stage by posing a research question (RQ) where a null- rejection confirms the conclusion expected. That is, the expectation is that "no relation" will be rejected, like "EM increases cost-of-capital" and thus the two variables relate due to null- rejection. Most researchers apply this broad framework by identifying a key variable in a regression; a null rejection takes place if the related estimated coefficient has the correct sign and a small p-value. Given such a finding, researchers can then write the paper aiming for journal approval. Of course, this goal may not be achieved for a variety of reasons, such as "the RQ is not important enough", "inadequate recognition of endogeneity issues ", etc... But, despite the existence of an endless number potential objections, the imperative remains unchanged: the submitted paper must point to at least one table which substantiates the RQ's answer. And because the paper's findings cannot be guaranteed in advance, it is implied that readers will learn something new about accounting and its relevance. Aspiring researchers internalize this bird's eye view of empirical research quickly.

Though the big picture of null rejection may seem innocuous, researchers with a clear RQ answer in mind run into a prickly question: "What do I do if the analysis does not support the RQ as expected?" The question is far from incidental -- experienced researchers come to realize that for most projects the word "if" should be replaced by "when". Though this practical question is of epistemological significance, the accounting field has not addressed the question (as far as I can

[1] This paper extends some prior work of mine, "Accounting Research and Common Sense", Abacus (2016). As in that paper, the current paper makes points familiar to many readers. My hope is that the reader finds the totality interesting. As to how accounting academia works, I base it on personal observations influenced by discussions with colleagues. It goes almost without saying that most of what I am saying is very subjective. But I like to think that I am by no means alone. My preliminary take of my audience is that it is (i) easy to find people who agree, but (ii) it is not viewed as self-serving to talk about the deficiencies of accepted research conventions. The latter makes sense: there are presumably many successful academics who believe that the overall output due to our research is of social value and that it is all too easy to critique research practice.

tell). It becomes a subtle matter because an answer to the question interacts with other aspects of research, especially how to interpret regression statistics. Thus, the question is only stage setting. Starting with the question posed, a more expansive analysis leads me to conclude the following: *much of the published empirical research falls short on credibility; it ends up being of little interest to all but those researchers who later rely on the work to pursue their own publications* (the process becomes much like a so-called chain letter where citations serve as the game's reward). Stated in narrower technical terms, too many accounting empirical research papers tend to engage in statistical analysis without recognizing the potential for *false positives (FP or Type I error)*.[2,3]

[2] The claim that the academic literature publishes paper with a very high rate of FP became prominent in 2005 when the Stanford statistician Ioannides published his well-known paper "Why Most Published Research Findings are False". In this context the word "most" means what it says: more than half. What makes the paper particularly interesting is that it highlights the importance of prior probabilities of a hypothesis being true and the role of bias in the research process. A best-selling book by the academic Silver (2012) discusses Ioannides and others work related to the suspected high incidence of false research findings. He adds to the discussion by explaining how "big data" will make things even worse,

The paper by Ioannides discusses points which bear on accounting research:

" Several methodologists have pointed out the high rate of non-replication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded, strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically less than 0.05. Research is not most appropriately represented and summarized by p-values, but unfortunately…."

And truly striking:

"The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true."

If one takes these statements seriously, combined with Ioannides claim that more than half of all studies in the sciences are false, then one can reasonably claim that most articles in accounting A- journals fall into the category of "a branch of creative writing somewhat connected to to real world data" as opposed to "science". In other words, it is a rather frivolous activity which can be dismissed as irrelevant to non-academics. But there are arguments against this dystopian perspective. Social science methods can never formalize true vs. false hypothesis in a rigorous sense; it is not meaningful to compare it to the traditional sciences. Social science empirics can only provide evidence in the spirit of "it is not totally unreasonable to claim that the evidence supports Y relates to X positively". It is now *implicitly understood* that "there may well be all sorts of data analyses that run counter to this claim, but it is beyond the scope of the current research to consider such possibilities."

[3] Many papers in the Finance field suggest that its literature publishes an excess of false findings. For example, see Powell et.al. (2009), Harvey et.al. (2016), Harvey (2017, a presidential address), Hou et.al. (2017), Chordia, Goyal, Saretto (2017). All of these papers underscore the extreme incidence of false positives. As to Economics, it, too, has long history addressing the validity (or lack thereof) of empirical research. A recent paper by Brodeur et.al. (2016) shows how empirical analysis of statistics produced in published articles support that the literature's null-rejections depend on screen-picking: specifically, there is a material shortage of p-values below 0.05 but more than .010. This paper also provides extensive references.

Well informed accounting academics can of course disagree about the pervasiveness of false positives or, more generally, questionable findings. A close examination of available facts would probably produce a foggy picture since an objective truth is unavailable. Even co-authors may not agree on whether a paper's claims properly represent the totality of the data analysis. Nonetheless, if one grants for the sake of argument a pervasive presence of false positives in the literature, then one can still ask: what factors are relevant and how do these interact? In broad terms, the answer depends on how researchers perceive what constitutes acceptable methods per conventions and their predispositions when deciding on how to address a RQ. More specific aspects of the false positives incidence relate to the role of researchers' incentives, reviewers' perceptions about their task, and, foremost, researchers' choice of data analysis methods.

In sum, this paper connects a sequence of common sense observations to argue that, overall, published papers favor implausible null-rejections. While each piece in the chain of reasoning is neither new nor surprising, the subtleties relate to interactive effects. These worsen the false positives. And the discussion goes beyond that: the end part of the paper contends that the research conventions examined constitute a stable equilibrium. Any attempt at reforming the research environment would accordingly run into considerable hurdles.

II. Unsupportive Table(s): Researchers' Responses.

Getting the "wrong results" does not necessarily mean the researcher terminates a project, for a variety of reasons. Though the publication imperative is ever present and "wrong results" is a no go in the final version of the paper, writing off a project might well seem rash and premature. Researchers are prone to get attached to their story as to how the world works, so they look for ways the investment can pay off. More important, it is common knowledge that a large portion of papers have been published though the researchers initially got the "wrong results". Experienced researchers are also aware that almost no empirical research project follow a short straight line, starting with RQ, followed by data collection/analysis, and a final validation of the RQ. There are all sorts of complexities to be confronted along the way and these force the issue how to proceed when results that were hoped for failed to materialize.

Barring any ethical constraints, a researcher can simply make up the tables using whatever numbers he/she wants. Though such conduct is probably very rare (I think), this observation nonetheless reinforces the practical issue that successful research centers on the need to produce “supportive numbers” without getting dragged into behavior that violate research-conventions. Even inexperienced researchers quickly come to realize that, indeed, there are openings to address the problem at hand – conventions are generous. A succinct prescription how to proceed runs as follows: try out something else and rest assured that past disappointments can be forgotten.

(i) If preliminary empirical findings do not support to the story to be told, then researchers typically proceed by trying out alternatives: constructs, controlling variables, and statistical methods. Papers do not provide any record as to what was done and why. Abstracts do not provide the slightest hint that a paper’s take-away could be equivocal due to insights produced by preliminary data analyses.

The established practice of engaging in trial and errors until the results align with prior expectations is less than ideal. Though the research community tends to accept it as an unavoidable convention, many also view the practice as bordering on the unethical. In the popular parlance amongst academics, uninhibited trial-and-error searches for the “right” tables are often referred to as “screen-picking”. Of course, it comes with a distinctly negative connotation.[4] Readers interested in some specific paper often discuss the extent to which key tables depend critically on screen-picking. A conclusion along these lines questions the paper’s validity; it submits that the results would not be available if the researchers had followed a disinterested approach. But since the skeptical readers cannot be confident about the detailed history of the paper, the implied disapproval must be muted and not publicly proclaimed. Individuals who critique the paper may not be squeaky clean either – they, too, may have engaged in screen-picking. In this context, it is well to note that, more generally, people do not feel that they are obliged to provide information that reflects poorly on their behavior. It does not surprise that researchers’ stay away from

[4] From what I can judge, the profession at large become aware of screen-picking as an explicit “tool” during the 1980s. (I recall having had an extensive discussion of a paper published in JAE, 1989: could the paper be replicated at all? To what extent would that be feasible because of the screen-picking that assuredly had taken place?).

discussing publicly the extent to which they or others engage in screen-picking. Still, insofar researchers' behavior ought to meet the highest of ethical standards, it is an unhappy situation. And this aspect is magnified by the fact that most researchers rarely consider the use of methods that offer a serious probability of rejecting the null. The latter is a somber observation because the option is always available (the point is discussed soon). Seminar rooms are habitually stalked by an elephant: given that the RQ is rather farfetched, how come the paper did not report on any evidence whatsoever that raised some doubts about the conclusion?

A more appreciative aspect of the common use of trial and errors recognizes that researchers find it difficult to specify the "right model" which translates a vision of how the world works into measurements and equations. Attempts at implementation might well turn out to be misguided with hindsight, and options how to proceed tend to be far from clear-cut. Even skilled and experienced researchers can initially fall short. Insofar there are many balls that must be juggled, mental errors occur, and these can usefully be straightened out. One can accordingly argue that a researcher should be permitted to modify the data analysis in a pseudo-experimental manner. If procedure X1 yields "strange" results, then trying out X2 in the next step may work better because the researcher can refine his/her thinking based on what was learned from X1. This would seem to be justifiable if the researcher proceeds in a rational way relying on constructs that bear on the RQ rather than simply engaging in a random chase for "something that works". As the researcher writes the paper he/she must then decide on how to report on the research effort in its totality. It would seem to be both reasonable and practical that the *ex-post* misdirected steps are not mentioned due to publishing constraints and information overload. A paper's history of past efforts before the final version is often of little interest to readers; the issues and procedures used are too intertwined to lay out in an orderly fashion.

To have numerous options how to frame and implement the RQ opens a Pandora's Box: there are no limits as to what can be tried out, and *something* ought to lead up to the desired outcome. It becomes impossible to obfuscate that the over-exploited data confounds the statistical significance assessments, and the direction of the bias is always in the same direction: in favor of the desired outcome. More subtly, much of the research ends up recounting only the tip of the iceberg; the below water specifics can only be guessed at.

(ii) Though researchers try various approaches to get "the right results", concerns about statistical over-fitting are rarely expressed. Nor will the reader be informed about the preliminary results and the extent to which these run counter to the claimed findings. If in fact the researcher knows that the findings claimed are dubious, he/she need not worry about subsequent refutations of the paper: these carry few if any professional penalties (unless the numbers in the tables turn out to be bogus).

As to the second part of the above point, in principle it could be the case that a "Professor Smith is in professional trouble because her research claimed that 'CEOs engage in EM to increase their pay', even though recent empirics suggest otherwise. Smith failed to recognize that her t-statistics were materially overstated – which straightforward bootstrapping would have shown". As far as I can tell, these kinds of stories do not float around. Few people in the profession seem to worry much about what went into a paper and whether its conclusions are solid. The act of an author (or authors) retracting a well-known paper is unheard of (in accounting, as far as I can tell)). This lack is quite noteworthy since it would seem reasonable that at least some authors, with the advantage of hindsight, would change their minds as to the validity of a paper's conclusion. And the propensity to disown one's research becomes even less if the paper has been highly cited, even though the imperative to rectify now becomes truly important (People, including academics, do not like to part with their wealth.)

The possibility of false conclusions can be firmed up and tied to data analysis methods. Basic courses in statistics teach that powerful tests have the virtue of increasing chances of rejecting the null. It follows that researchers who want to reject the null gravitate toward picking powerful tests as well as the full exploitation of the data (that is, no hold out sample). It comes at a price, namely, a tilt towards FP. This price from a researcher's career perspective, however, becomes trivial since the he/she will not be penalized if the finding happens to be an FP.[5,6]

(iii)Classical statistics posits that the power of a test trades off against the probability of an FP: if one of the two attributes is relatively high, then the other tends to be relatively high. Most researchers focus single-mindedly on the former. The effective dismissal of FP as a concern aligns with points (i) and (ii) above.

The trade-off, the power of the test (a good thing) and a FP (a bad thing) means they move together. Thus, to reduce the FP possibility the research design makes the acceptance of the null very likely when the null is true – which appeals -- but there is a drawback because of the now relatively high probability of falsely accepting the null.

The celebrated Jeffreys-Lindley Paradox illustrates the trade-off: loosely speaking, as N increases so does the probability of a FP.[7] In other words, large N introduces challenges; the researcher who wants to reject the null does not mind – but the cost, whether articulated or not, is the potentially large probability of a FP. In principle, this can be handled by adjusting the significance level as N increases, a point made by no lesser individual than Pearson (more than a hundred years ago). But this way of looking at large N leaves no trace when the papers present findings. In fact, researchers proceed as if a t-statistic with three stars should satisfy everyone – no need to add some comments because a t-statistic depends on N and that a very large N has consequences. Few people get fooled: everyone knows that a t-statistic of 3.2, say, does not impress the least if N exceeds, say, 50,000. And, to make matters worse, under such circumstances the academic community has

[5] How serious are the accounting academics about producing research of integrity? Not very if one considers the following. Many colleagues have suggested it is viewed as inappropriate to challenge a presenter by suggesting that certain simple data analysis most likely would negate the claimed positive answer to the RQ. Why worry about the validity of conclusions presented? Will life not go on, no matter?

[6] To be sure, there are good reasons why the academic community is disinclined to reprove authors who have published articles with false findings. Most important, research is an intrinsically difficult activity, and researchers should be encouraged to take risks.

[7] The so-called Jeffreys -Lindley Paradox effectively says that, under relatively mild conditions, as N becomes large a rejection of the null becomes a sure thing even though a rational Bayesian would conclude otherwise. See Spanos (2012) and references in that paper.

learned to live with unqualified rejections of the null which are arguably unethical if the research also engaged in extensive screen picking. Nobody seems to know how to deal with this kind of research – except, perhaps, just quietly neglect it.

III. The Role of the Literature: Powerful Tests and Alternatives.

Given the last point, it becomes obvious what the answer will be when a researcher asks the next practical question: "How do I go about finding tests that optimize my chances of rejecting null?"

(iv) Because an overwhelming portion of A-journal published papers rejects the null hypothesis, researchers tend to gravitate toward commonly used statistical methodologies when testing hypotheses. Such choices are particularly likely when a hypothesis has a low prior of holding up. The possibility of the literature comprising a large proportion of FP has no apparent impact on researchers' data analysis choices.

By way of example, consider the use of fixed effects in regressions. This generic regression has to a considerable extent replaced less powerful techniques like a firm (or industry) specific regressions or Fama-McBeth annual regressions (in case panel data across years and firms). A researcher with a plausible hypothesis may be satisfied using an old-fashioned Fama-McBeth test -- a strong prior and the latter test is unlikely to upset the story to be told; no need to pool all data into one regression with fixed effects for years.[8] Conversely, a paper which does not apply a Fama-McBeth type of analysis when it could have been done broadcasts a red flag: either the researcher tried it and "it did not work out" or the author is suspicious that it would not work out if it was applied. (Astute researchers stay away from a Fama-McBeth type of test when the prior is low that the RQ holds up: why take an unnecessary risk.)

[8] A Fama-McBeth test does not have to be implemented with its details. On a basic level, if a regression is run for say 25 years and the sign of the estimated coefficient of interest is correct in 23 out of the 25 years it would seem reasonable that one rejects the null. But, of, course, the readers can, and should, make their own judgements. Compare this approach to the pooled over years fixed effects approach when N equals, say, 20,000. It is now hard to tell at what minimum magnitude of the t-statistic will convince the reader that the null can comfortably be rejected. (For what it is worth, the author uses sqrt(0.4%*N) which I picked up way back. It derives from the idea that a t-sat=2 and N=1,000 is border-line accept/reject. The formula reflects the necessary adjustment to the t-stat as N becomes larger.

Researchers who have posed a hypothesis with a low prior probability must be particularly careful when deciding on the data analysis methods. The method picked should be as a powerful as possible -- and the prior literature guides the choice. It is also clear that it is essential to stay away from straightforward data analysis which would allow for a serious possibility of null-acceptance. Even a complementary analysis with the "wrong results" can leave the reader with a skeptical attitude. So researchers recognize that these detours are best avoided. After all, low priors are low priors and it is hard to produce confirming evidence unless one uses a powerful test with a high FP probability.

It makes a lot of sense that the careful, and honest, researcher wants to avoid faulty conclusion and thus he/she engages in extensive data analysis prior to selecting the "final" tables. As noted earlier, such analysis can often be justified; it would now be inappropriate to use the derogatory term screen-picking. But a reader will in general find it close to impossible to evaluate whether the preliminary data analysis has been ethical or not. One can take this observation a step further: *robust* testing mitigates concerns related to point (i) *because in spite of the prior data analysis it can still be hard to reject the null*. To illustrate, consider a research project using panel data of say 25 years and an average of 2,000 firms in annual regressions. Suppose next that the sign of the estimated key coefficient is correct in 23 out of the 25 years. One can now reasonably conclude that screen-picking is not the likely reason for such a significant result rejecting the null. In contrast, if the regression had been based on a N=50,000 by pooling the data (with or without fixed effects) a critical reader would worry about screen-picking as a driving force. *Accordingly, powerful tests mix only with untainted statistical experiments -- which effectively requires hold-out samples.*

Certain kinds of data analysis are noticeably absent in the literature *because* they can effectively address whether a null hypothesis rejection most likely falls into the category of FP. To illustrate, consider the popular regression model framework:

$$y = a.X + b^*Z + \text{error term} \qquad (1).$$

In (1) X identifies a variable of particular interest due to the RQ, and Z represents a vector of controls (it allows and intercepts and FE if so required); (a,b) stand for the coefficients to be estimated. Researchers typically are satisfied as long as the estimate of the a-coefficient has the right sign and a small p-value.[9] Not-so-pleasant problems now potentially arise because, first, t-statistics are, arguably, almost always overstated and, second, the t-statistic may look too modest given a large N. Few (maybe even very few) papers express concerns. This makes sense insofar that the two points are hard to nail down objectively. However, one can now of course argue that it raises the need for additional tests to check whether the null still will be rejected. Should not the researcher be sensitive to the reader who feels unsettled about a modest t which is also likely to be overstated? If the will is in place, researchers would address the question constructively as a routine matter.

The FP issue can be dealt with head-on if one checks whether X, in fact, helps (beyond Z) to explain the dependent variable. Consider a competing model which has removed the key variable of interest (X) and where the estimated coefficient vector, c, has replaced b.

$$y = c*Z + \text{error term} \qquad (2).$$

In the next step, the researcher can evaluate the extent to which the accuracy declines going from (1) to (2). To check on this one can apply a relative accuracy scoring that compares the two models' inferred values of the dependent variable relative to actual values. The regression model that most often generates inferred values closer to the actual y values can be declared a "winner". It becomes an empirical question whether (2) does worse than (1); one needs to keep in mind that the t-statistic

[9] OLS-regressions are central in accounting research conventions. The method achieved such status originally due to the limits of computers. Nowadays, of course, there are no such reasons why OLS should be applied. From a strict intellectual perspective, there are no particular merits associated with OLS. In fact, its application seems bizarre insofar that OLS assigns a material role for outliers and at the same time most research winsorizes the outliers; in other words outliers are replaced by less abrasive outliers that are now real world disconnected. OLS's staying power, I would argue, stems from the overstated t-statistics. This aspect helps when researchers try to get the "right result" via screen picking. Injecting some irony (or humor, perhaps), by convention the published tables express the t-statistics (or standard errors) using no less than four digits. To add insult to injury, a critique of OLS t-stats includes the observation that t-stats become even more upward biased if one considers, realistically, that the independent variables are random rather than "predetermined" (I am never quite sure if "preordained is a better word in this context). And then there is the issue that the t-stats depend on N whereas the real world does not. For a succinct exposition which critiques classical statistics, see Silver's book, pp. 251-61. In finance Harvey (2017) discusses the issues; he provides an extensive discussion as to the intrinsic problems related to classical regressions.

related to X cannot forecast the outcome for sure. And it is well to note that it is perfectly possible that *(1) may do worse than (2*).[10]

The above can be modified so that the benchmark model includes only, say, 4-6 variables on the RHS, that is, those that one can reasonably guess to be the most important. Again, to improve the explanatory power of such a basic parsimonious model might well be a challenge, in which case the idea of X taking on a material role may be implausible. But it is of course by no means a foregone conclusion since any model may potentially exclude some helpful explanatory variables.

Data analysis in the spirit of the above poses no problems no matter the setting, and the above merely illustrates. The core issue centers on a researcher's willingness to evaluate the robustness of a null rejection as the power of tests decline. Though the degree of power in relative terms as one move from one test to another is bound to be highly subjective, the real purpose is to help the reader to understand the data and to allow the reader to make his/her own judgement about the extent to which the hypothesis at hand is compelling, or not. The fact that the null is accepted in a relative accuracy test should thus not be viewed as the "preferable" overall conclusion; the test is only one out of many (or at least two). A researcher might well think, and argue, that on balance the evidence suggests otherwise though the difference between (1) is no more accurate than (2). And a reader may or may not be persuaded, which merely recognizes that in the social sciences ambiguities are part and parcel of the business.

[10] Many papers posit regressions where the dependent variable is a ratio, such as the market-to-book (M/B) ratio. (These M/B settings are often referred to as Tobin Q-ratios and they purport to explain market values even though they do not.) Now it is often the case that the variable of substantive interest on the LHS of the regression is M; B serves the purpose of scaling M to avoid severe cross-sectional heteroscedasticity etc. In such case one obtains the inferred variable by multiplying both sides of equations (1) and (2) with B. Thus the inferred value using (1) now equals BV*[a*X + b*Z] where (a,b) are determined by their estimated values. Yet again, it becomes a trivial matter to ask whether, as an empirical matter, X helps beyond Z to explain the dependent variable of real interest, namely, M.

The procedure can also be applied in case of logit-regressions.

The procedure suggested yields binomial test-statistic.

I do not recall ever having seen a paper implement this procedure, or something similar. Yet, in some private conversation researchers have indicated that in most cases the procedure most likely would yield "undesirable results" – and this may be the reason for its complete absence in the literature. (As an aside, I learned about the procedure in graduate school. It was taken for granted that this step in the analysis could not be skipped.)

Many papers fall into the category where the outcome of a more demanding test of the null (like the one proposed) is far from a foregone conclusion, one way or another. Researchers can guess as to the likelihood on the basis of the t-statistic and N; a relative small t-statistic in light of a large N may tempt the reader that the null most likely will not get rejected. Thus, the researcher now has a real opportunity to convince the reader otherwise -- and in such case the language of "compelling evidence in favor of the hypothesis" becomes appropriate even in the paper's abstract. And the steadfast researcher may take the analysis a step further and compare (1) and (2) using a holdout sample of (y,X,Z) and estimated coefficients estimated from prior data. To claim the potential of a FP would now seem to be almost impossible as long as the explanatory power of (1) exceeds (2); even the most critical reader ought to be impressed.

The last paragraph raises the issue of the role of citations and related prior research. Researchers are expected to cite papers even if a non-trivial proportion of the papers are not viewed as credible or have been read; the imperative tends to be "when in the slightest doubt whether to cite or not, just cite" (of course the rule reduces chances of hurting colleagues' feelings, with its potential negative professional consequences). This convention of all-encompassing citation thus signals, at least as a first cut, that the author sticks to the literature and appreciates that the state of the art should generally be adhered to. And in fact deviations tend to be exceptions rather than the rule.

To use non-standard methodologies causes considerable problems insofar that the researcher would have to explain reasons for deviating. In other words, a researcher who wants to maximize chances of getting the paper accepted would not entertain this option. These more narrow professional aspects combined with the research imperative to use powerful tests lead to an equilibrium which discourages trying new approaches of data analysis. Researchers thus move in the direction of risk-aversion, an arguably negative outcome for the field as a whole. One should thus not be surprised that papers to not discuss and implement extensive tests to avoid FP and related faulty conclusions: it is generally viewed as a not-so-self-serving activity. The focus concerns cutting-edge conventions in the A-journals, no more no less. We discuss these issues in more detail next.

IV. Reviewers: Do They Mitigate False Positives?

Before a paper gets published, the standard process includes a so-called peer review that must be passed. One can then ask whether this process falls short because reviewers lack the capability or incentives to properly tackle issues related to FP. It may be the case that the current culture of producing reviewers' reports makes matters worse, that is, it increases the pervasiveness of FP. My overall impression is that such is the case. But the problem involves many aspects, so it by no means a foregone conclusion. My summary argument runs as follows: reviewers rarely take authors to task for *not* using tests that are unlikely to reject the null unless the null is true. In other words, reviewers do not care about the possibility of false positives as long as the paper uses cutting-edge methods. Reviewers and researchers thus reinforce the tilt toward false positives.

In many cases reviewer's will have a low prior concerning the plausibility of the hypotheses promoted, and he/she may then at least contemplate the suggestion that the researcher should take a closer look at the data. Some reviewers presumably proceed along these lines with some success; that is, the additional work performed does indeed suggest that a likely FP has been avoided. But to ask the author to produce such additional analysis may seem excessive and cause to much extra work for everyone, especially if the reviewer feels the effort is unlikely to yield the outcome that the author maintains. As a make-do alternative, the reviewer may decide to reject the paper on some other ground (like "the paper does not adequately connect its RQ with well-known papers that deal with closely connected topics, e.g, Smith et.al („„), Chang(…) et.al,…..."; "the exposition is confusing". This option also has the advantage that it avoids potential disputes with the author about how the real world works. Moreover, the rejection may be quite solid because the author in fact did not provided the exceptional evidence needed to support the perceived farfetched hypothesis.

Most reviewers do not like to second guess the extent to which an RQ is plausible or not, for good reasons. It is clearly on the subjective side. There is also often a sense that research ought to be agnostic as to the outcome of data analyses. Instead reviewers prefer to focus on the following question to resolve the acceptance/rejection issue: given the RQ (assuming it is new and logically sensible and refers to a solid number of citations to back up how relevant and important it is), did the researcher use methods of data analysis that are in the spirit of what I (the reviewer) would use? The question makes practical sense. It takes little imagination to surmise that reviewers often find that papers fall short on how to deal with a full range of implementation issues, like how to identify and specify confounding variables, the use of constructs markedly different from the literature and, to really annoy the author(s), the so-called endogeneity problem. It can go on and on, and it may include statistical issues like heteroscedasticity, choice of deflators, the lack of interactive effects, and of course failure to consider the latest state of the art such as corrections for "clustering" when calculating t-statistics. And the more assiduous the reviewer, the longer the list of things to do (and the longer the response memo) if the paper achieves R&R. This process would seem quite reasonable—and more often than not end up being quite constructive from the researcher's perspective– provided that the researcher and reviewer tend to look at the nature of research and how it should best be done through a similar lens. Conversely, it would seem reasonable to expect that serious conflicts surface if the reviewer and the author differ in their sensibilities as to what constitutes worthwhile research. Now the two parties involved most likely end up wasting time arguing without any resolution – and the ultimate outcome will be a rejection.

The discussion above suggests that the essence of the review centers on the extent to which he paper conforms with what may be called "generally accepted research procedures"; regressions and OLS estimation, controlling variables handed down from previously research (like ROA, size, and M/B; but not generally growth in sales, PM or ATO), winsorization or trimming at commonly accepted levels, the use of fixed effects, and ordinal measurement constructs that are commonly accepted (like measuring the M/B effect by putting the ratio into ten decile bins). To suggest that the author should modify his paper using procedures that are rare in the literature would seem to be, at best, as eccentric. For example, it is unlikely that a reviewer's report tells the author that he or she should use some estimation technique other than OLS because OLS is known to be inefficient (cites to that effect provided by the reviewer). And the notion that a reviewer tells the researcher that he/she should have use hold-out sample techniques is also exceedingly unlikely unless the research area has a history to that effect. As another example, more down to earth, a review is exceedingly unlikely to state something like "The paper uses the absolute value of (A – F)/P to measure analyst forecast errors. This measure correlates positively with F/P, which in turn means that it correlates negatively with firms' growth prospects. This feature can distort findings.

But it can be avoided using some alternative metric such as …….. (details and cites provided) ". A researcher subject to such criticisms would naturally react with annoyance: why should I have to discard my approach when it seems to have been perfectly fine in dozens of recently published papers? Thus, I expect these kinds of critiques to be rare because most reviewers prefer to act equitably. The only exception seems to occur if the referee tries put pressure on the reviewer so that he/she recognizes research promoted by the reviewer (often papers published by the reviewer). But this would be viewed as in rather poor taste, and most reviewers tend to focus on the rules of the game: the paper should rely on methods that are consistent with the reviewer's perception of how it should be done per currently accepted procedures -- no more, no less.[11]

It would not be accurate to say that reviewers always look negatively on papers that do new or unusual kinds of data analysis. Some reviewers actually look on attempts by authors' to be original with favor, and he/she many give the authors the benefit of doubt. However, my guess is that such occurrences are less likely than one might think. The reason is simple enough. Researchers do not to want gamble on reviewers having open minds about innovations in methodology; in other words, it is viewed as "too risky" by the great majority of researchers.[12] It leads to an equilibrium which would seem to be stable. Pre-existing research methods maximize chances of getting the "right results", and rarely if ever do reviewers require serious robustness tests that jeopardize the validity of findings -- as long as the methods used fall into the acceptable category. The notion of a reviewer suggesting that generally accepted principles of doing research are falling short, and a consequent rejection of the paper -- would not seem to be in the cards. Only well-established

[11] To reduce the incidence of false positives one could expect that if the reviewers fall short members of the community at large step in and make an attempt to rectify erroneous papers that passed the publication hurdles. That does not seem to be the case, however. Editorial policies effectively rule out the publishing of a "commentary" that claims to show that some prior paper is in error, or the conclusions are at best dubious. (Interestingly enough, the policy was not in place 4 decades ago). The incentives to write such a commentary, let alone an entire paper, appears to miniscule since it is likely to lead to serious professional hazards. Nonetheless, there are constant rumors floating around that this or that paper cannot be replicated. It is an unhealthy situation because the policy in place projects an impression that it protects the more important people in the system.

[12] Most people would agree, I think, that researchers spend considerable time on guessing what the potential reviewers' judgements will be. An overwhelming majority of authors sidestep any action that might increase chances of a negative report -- no matter how small the effect is on the perceived change in the probability. And, again as far as I can tell, in second rounds the whole idea of not being 100% compliant is not entertained.

researcher can experiment with new methods that fall outside the general principles of how to do empirical research. Individuals hard at work trying to get tenure, by contrast, stick to what they were taught in graduate school – they find it hard enough to get published and to start to experiment is not viewed as eslf-serving.

Changes in acceptable methods occur over time. It pretty much has to come from the top in the research community hierarchy. Thus the folks lower down on the pyramid know that the new procedures can be followed without risk. And these new procedures are unlikely to make it more difficult to reject the null. It is precisely what one should expect since people at the top do not promote changes that can undermine the credibility of their own (highly cited) publications. When everything is said and done, the importance of this aspect can hardly be overemphasized.

V. Potential Remedies: Changing the Research Mentality.

So what can be done to reduce the proportion of papers with dubious conclusions? From an individual researcher's perspective, it hinges on changing the mindset as to what good research is all about. The current approach tend to follow two steps, (i), pose a novel story about the world – completely unheard of in many cases – and, (ii), then proceed to validate the story – avoid qualifications. On reflection it would seem that (i) and (ii) as a practical matter contradict each other; it should serve as an indicator that the mindset must change. Some suggestions follow:

- Extraordinary claims require extraordinary evidence. In other words, a researcher should be sensitive to the notion that a proposed hypothesis may seem unlikely to the typical reader and, in such cases, the evidence ought to be more extensive and compelling than otherwise.
- Implausible hypotheses embedded in RQs all too often reflect a researcher's wishful thinking. In turn, it sets the stage for the dysfunctional imperative of going to extremes to verify the hypotheses; the statistical overfitting aspect is simply thrown out of the window. To get out of this mode of thinking, more constructive approaches rely on one of two alternatives. First, the question at hand is sufficiently interesting so it makes no real difference if the answer is 'Y' or 'N'. (To a surprising degree, very few papers posit hypotheses in an agnostic, neutral, tone.) Second, what the data will show in broad terms

is pretty much known by everyone, but the data analysis tries to bring out aspects that make the underlying issues more interesting.

- The conclusion should be primarily driven by the robustness of alternatives across the various ways a question can be framed – not p-values. As a direct consequence, researchers should feel comfortable with ambiguous conclusions.
- Researchers need to recognize that the idea of a never-previously-heard of hypothesis holding up robustly is simply not credible. In such case the researchers' needs to gather extensive evidence from financial media that the prior is really not that low. In the absence of such evidence a researcher might as well terminate the project rather than chase a sufficiently small p-value -- which will be of dubious quality. Most readers will not get fooled.
- To motivate research issues and related questions a researcher should go beyond just citing prior papers. A relative extensive review of relevant arguments should aid readers: they need to judge in their own terms the plausibility of the hypotheses.
- An analysis of basic facts that go beyond the usual descriptive statistics helps diligent readers. From a Bayesian perspective, the natural starting point considers the distribution of the core independent variable (x, "the signal") conditioned on the dependent variable (y, "the outcome"). Thus consider the following table. Put y in (say) ten equaled sized bins, and for each bin calculate the median of x. The examination of such a table will provide useful information, often very sobering as to the likelihood of obtaining a solid rejection of the null.
- To convey relevant information, it is hard to overemphasize that it makes little sense to focus squarely on p-values. A researcher who wishes to say something interesting may start out thinking about the following question: how can I say something worthwhile that goes beyond the sign and p-value (t-statistic) of an estimated coefficient?

To allow for real-world complexities as a matter of course leads to a way of thinking that centers on issues in broad terms, rather than questions that are rather precise. The latter approach runs counter to many researchers habits. It modifies current ways of doing research. A common prescription applied, taught in every doctoral program of some standing, underscores the importance of spelling out the so-called RQ – typically a story about how the rea world works --

without ambiguity. It tempts answers in terms of a ‘Y’ or ‘N’ without qualifications since the absence of qualifications enhances the clarity of the paper’s takeaway. To avoid imposing such a rigid framework, the researcher may consider posing a RQ that does not yield a Y or N answer but instead an answer that depends on various methodological aspects. The message of the paper may get diluted, or more difficult to internalize, but that should be less of the problem if the researcher can convince the reader that the real world can be complicated and interesting at the same time.

As indicated earlier, to really change the tone of research, the ultimate issue pertains to the willingness of researchers to apply tests that convinces the reader that the rejection of the null would not occur unless the null ought to be rejected. Judgements will be necessary as to the ultimate conclusion in light of the totality of evidence – for and against --, but this aspect should not act as an obstacle. After all, researchers ought to be more interested in informing the reader rather than acting as advocates of certain claims about how the real world works. With this perspective in place, the method related to equations (1) and (2) serve as an object example: the testing procedure, or something similar, ought to be implemented in all research projects. If it were to happen, it would change the research environment radically for the better.

VI. A Few Final Remarks: Nudging the Research in a Preferable Direction

With some over-simplification, to move current research practice to accept more constructive conventions the following can be said: researchers must abandon the idea of getting the “right results” and replace it with getting the “results right”. To an amazing degree, the literature has embraced the former as a guiding principle without recognizing its natural consequence, a trampling on the latter. Some people may even argue that the current state of affairs is bizarre: the abundance of “validated” stories has been growing exponentially over the years. While the RQ have become increasingly esoteric, as far as I can tell no evidence suggests that researchers have found it increasingly difficult to validate the RQs. Rejection of the null hypotheses, when desired, generally pose no recognized problems, and, as on cue, robustness tests always work out per want. From what I can tell, robustness tests never overturn the paper’s basic conclusion and researchers can so announce with forthright satisfaction. Even more amazing, I have never heard of any paper finding that some effect is *not* economically material. However, when something seems to be too good to be true, it probably is not. Few members of the community get fooled: to a considerable extent papers are not taken all that seriously. Individuals who attend seminars on a regular basis end up being less and less impressed by the totality of the research effort. It would not be easy to argue convincingly that the great bulk of paper published generate much interest (setting aside captive audiences).

The academic community become seriously interested in research output when it affects individuals' careers. But such assessments primarily deal with counting entries on CVs and citation counting, not the papers' validity, creativity, etc. An informed cynic may suggest that the research community relies on an "implicit contract" stipulating that individuals who aspire to be accepted as researchers in respectable standing will have a reasonable shot at making it if they follow established ways of data analysis, combined with citations of academic papers to motivate the RQ. Conversely, a researcher cannot expect any bonus points by applying straightforward data analysis methods with the conscious purpose to get the results right – rather than providing the empirical support for the pre-set results. And to explicitly suggest in a promotion case, or in a paper, that some highly cited papers are seriously flawed would stand out as, at best, eccentric. With some imagination, one can argue that the crafting of such sentences would be viewed as inviting ex-communication (or, at the very least, it would rule out future A+ journal publications).

A more positive view recognizes that research is intrinsically complicated and ought not to be idealized any more than any other activity. But some changes in research conventions can be called for precisely because research complexity stems from the intrinsic complexities of the real world. Put differently, the goal of getting the right result without any qualifiers is in the spirit of wishful thinking; it should be viewed as a true exception in the absence of ubiquitous prior evidence to the contrary. Now it would seem to follow that to achieve effective characterization of results a researcher must include a discussion of the inherent ambiguities. And, indeed, with a different mindset in place researchers may entertain so-called robustness tests which no longer always "work out". But these contrarian findings may inform the inquisitive reader, and the researcher may engage the reader by discussing such aspects. A clause like "the interpretation of this result depends on …." would be used often. After all, researchers typically apply a range of alternatives: how to measure variables, how to deal with controls, trying out tests with relatively low statistical power as well as those with high power, sample splitting, bootstrapping etc. The novel feature would be the researchers' lack of need to focus squarely on providing evidence that backs a narrow view of how the world works -- and thus he/she would feel totally comfortable with a head-on trying to get the results right.

References

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. 2016. Star wars: The empirics strike back. *Applied Economics* 8(1): 1-32.

Chordia, T., Goyal, A. and Saretto, A. 2017. P-hacking: Evidence from two million trading strategies. Working paper. Emory University, University of Lausanne, and University of Texas at Dallas.

Harvey, C. R. 2017. Presidential address: The scientific outlook in financial economics. *The Journal of Finance* 72 (4): 1399-1440.

Harvey, C.R., Liu, Y. and Zhu, H. 2016. … and the cross-section of expected returns. *The Review of Financial Studies* 29 (1): 5-68.

Hou, K., Xue, C. and Zhang, L. 2017. Replicating Anomalies. Working paper. Ohio State University and University of Cincinnati.

Ioannidis, J. P. 2005. Why most published research findings are false. *PLoS medicine* 2 (8): 124

Ohlson, J. A. 2015. Accounting research and common sense. *Abacus* 51 (4): 525-535.

Powell, J., Shi, J., Smith, T. and Whaley, R. 2009. Common divisors, payout persistence, and return predictability. *International Review of Finance* 9 (4):335-357.

Silver, N. 2012. The signal and the noise: why so many predictions fail-but some don't. Penguin.

Spanos, A. 2013. Who should be afraid of the Jeffreys-Lindley paradox? *Philosophy of Science* 80 (1): 73-93.