econometrics

MDPI

Article

Important Issues in Statistical Testing and Recommended Improvements in Accounting Research

Thomas R. Dyckman [1,*] and Stephen A. Zeff [2]

[1] Accounting Department, Cornell University, Ithaca, NY 14850, USA
[2] Accounting Department, Rice University, Houston, TX 77005, USA; sazeff@rice.edu
* Correspondence: trd2@cornell.edu

Received: 17 December 2018; Accepted: 26 April 2019; Published: 8 May 2019

Abstract: A great deal of the accounting research published in recent years has involved statistical tests. Our paper proposes improvements to both the quality and execution of such research. We address the following limitations in current research that appear to us to be ignored or used inappropriately: (1) unaddressed situational effects resulting from model limitations and what has been referred to as "data carpentry," (2) limitations and alternatives to winsorizing, (3) necessary improvements to relying on a study's calculated "*p*-values" instead of on the economic or behavioral importance of the results, and (4) the information loss incurred by under-valuing what can and cannot be learned from replications.

Keywords: model specification; model testing; reporting results (*p*-values); replications

1. Introduction

As professors of accounting for nearly 60 years and past presidents of the American Accounting Association, we are concerned about the quality of statistical research in accounting. This article is a call to our accounting colleagues, and also perhaps to those in other fields, to invest substantial time and effort toward improving their requisite knowledge and skill when conducting the appropriate statistical analysis. Involving expert statisticians may be helpful, as we all need to recognize the limitations in our own knowledge in order to tap into this expertise. Our heightened interest in improvements to the quality of statistical analysis in accounting research was in response to attending research presentations and reading the current literature.

Several years ago, we suggested several improvements to statistical testing and reporting (Dyckman and Zeff 2014). In that paper, we reviewed the 66 articles involving statistical testing that accounted for 90 percent of the research papers published between September 2012 and May 2013 in *The Accounting Review* and the *Journal of Accounting Research*, two leading journals in the field of accounting. Of these 66 papers, 90 percent relied on regression analysis. Our paper examined ways of improving the statistical analysis and the need to report the economic importance of the results.

An extension of these concerns was included in a commissioned paper included in the 50th anniversary of Abacus (Dyckman and Zeff 2015). We acknowledge several accounting academics who are also concerned with these issues, including Ohlson (2018), Kim et al. (2018), and Stone (2018), whose works we cite.

Concerns about statistical testing led to exploring the advantages of a Bayesian approach and abandoning null hypothesis tests (NHST) in favor of reporting confidence intervals. We also suggested the advantages—and limitations—of meta-analysis that would allow for the inclusion of replication studies in the assessment of evidence. This approach would replace the typical NHST process and its reliance on *p*-values (Dyckman 2016).

A fourth article which reviewed the first 30 years' history of the research journal, Accounting Horizons, continued our concern with the current applications of statistical testing to accounting research. An additional aspect of this article was the attention we gave to accounting researchers' seeming lack of interest in communicating with an audience of professionals beyond other like researchers, as if their only role as researchers was to enrich the research literature and not to contribute to the stock of accounting knowledge. We submit that accounting academics, because of the academic reward structure in their universities, tend to write for their peers. Accounting standard setters and accounting professionals, as well as those who make business and policy decisions, are all too often relegated to the sidelines. We argued that accounting research should, in the end, be relevant to important issues faced by accounting professionals, regulators and management, and that the research findings should be readable by individuals in this broader user community (Zeff and Dyckman 2018).

In the current paper, we expand on the statistical testing issues raised in our earlier papers, and we identify limitations often overlooked or ignored. Our experience suggests that many accounting professors, and perhaps those in other fields, are not familiar with, or equipped to, address them. We take up the following major topics: Model Specification and Data Carpentry, Testing the Model, Reporting Results, and Replication Studies, followed by A Critical Evaluation and A Way Forward.

2. Model Specification and Data Carpentry

The choice of a topic and related theory established the basis for the hypotheses to be examined and the concepts that will constitute the independent variables. Accounting investigations often rest only on a story rather than on a theory. A major problem here is that a story, but not theory, can be changed or modified, which encourages data mining (Black 1993, p. 73). Establishing the appropriate relationships require an understanding of the actual decision-making environment. These ingredients, along with the research team's insights and abilities, are critical to designing the research testing program and the data collection and analysis process. Failure to take them into account in the data-selection decision process and analysis was discussed in detail in a recent paper by Gow et al. (2016). There, the authors provided a detailed example (pp. 502–14) of how the decision environment can reflect its own idiosyncratic differences that, in turn, influence the data. For example, even if the business context is essentially the same across companies, data limitations remain. First, the data will inevitably reflect different sets of decision makers and different organizations, different time periods, different information, and, at least, some differences in the definitions of the variables deemed to be relevant. The interactions between these variables, and with any relevant but excluded variables, will, as the authors showed, lead to questionable results. How the selected variables interact with each other—and with any excluded but relevant variables—depends on the nature of the contextual environment in which the relation arises. We note here that careful research designs up front can reduce interactions among the independent variables. Authors can and should describe the decision environment and differences, if any, that have a potential impact upon the analysis and conclusions. A thorough analysis and description of the decision environments is essential and endows additional credibility on the research.

Typically, a concept can be operationalized by more than one variable. For example, firm size may be proxied by the number of employees or by revenue. Furthermore, the choice of a measure is often made according to data availability. Even the topic selection may be determined by the availability of an interesting data set. Unfortunately, authors usually do not acknowledge the latter and may fail to justify the selected variable measure. Once the hypotheses have been modeled and the variables with their measure selected, the decisions must not be altered, expanded, or dropped without full disclosure. Yet, we have seldom seen these explicit limitations revealed, let alone discussed. Authors appear to ask readers to accept implicitly that such alterations have not occurred. Even a careful reading may not reveal the authors' reasons for their specific choices. Authors should not assume that their choices are transparent and elect not to address the choice process.

The choice of the data set for the variables included in the study is critical. We think of this as the data carpentry, during which the raw data are melded into the data set for analysis. This is

when data snooping, data mining, and related inappropriate activities must be avoided. Furthermore, researchers should not unquestioningly adopt a data set used by previous authors without verifying its accuracy and applicability to the current issue addressed. (For a discussion of what can occur, see Zeff 2016). Authors should also be alert to data sets reflecting different time periods, locations, or information processes. Conditions can be very different for the same variable across these dimensions. An assumption that data obtained under such circumstances will lead to valid conclusions cannot be sustained. Moreover, if the data source, timing, processing, or availability changes, the research team is obliged to bring these changes to the attention of the reader, together with the resulting limitations imposed on the findings.

2.1. Assumption of Randomness

The concept of hypothesis testing and its key elements, including test statistic, *p*-value, standard error, sampling distribution, significance level, rely on an implicit assumption of randomness. The investigation relies on the researchers obtaining a random sample from a well-defined population. Indeed, one of the purposes of hypothesis testing is to determine how big or small the random sampling error is with respect to the parameter value being tested under the null hypothesis. Accounting researchers, by their failure to address the issue, are taking this fundamental assumption for granted. This is unfortunate. Authors appear to be implicitly relying on Dunning (2012) assurance that randomness can be accepted if the reader can be assured that the researchers had no influence, intended or not, on the data process. Unfortunately, databases may be problematic in the context of random sampling. For example, these databases often cover the data for listed companies only, which can provide a biased sample if the research outcome is applied to non-listed companies. The decision to seek big data or even a large sample can lead to a similar problem (Harford 2014). Several examples with serious consequences are examined in this article.

A thorough defining of the population is essential, but is not easily accomplished, and often remains unspecified by the authors. An implicit assumption of randomness may be comforting, but it is not adequate. Authors are obliged to expend the necessary human capital to alert the reader to possible limitations in their data and how any such limitations could affect their results. An example is provided by big data (Boyd and Kate 2011). Unless the research design takes the sampling distribution into account, it becomes difficult to justify resampling and randomization. The authors recall no recent accounting papers, including those relying on big data, that have addressed this situation. The process of determining whether a subset of big data amenable to the theory of relevance could be identified is likely to doom any honest sampling process. Additionally, it would preclude replication.

2.2. Model Modifications

Once the hypotheses have been modeled and the variables have been selected and measured, any changes must be justified with full disclosure. Yet we have seldom seen such changes revealed. Authors apparently expect readers to accept implicitly that such alterations have either not occurred or are appropriate. A new approach to reduce this problem is being explored that requires authors to describe their choices in advance of executing the research project and to communicate to the editors any changes thereafter (Bloomfield et al. 2018; Kupferschmidt 2018). However, there is no assurance that this requirement will always be met, because the action may occur before initial submission.

2.3. Winsorizing

It is not uncommon to find accounting studies whose authors have winsorized their data and assumed their readers understand the process. By winsorizing, the authors are attempting to prevent what they regard as outliers in the data from unduly influencing the results. Authors using this approach apparently assume that the outliers do not belong to the set defined by the variable under consideration. Retaining these data points, if inappropriate, will bias the results. Each such data point also has a larger impact on the results. However, the adjustment process used is generally ad hoc.

We submit that data points omitted from the analysis require individual justification based on analysis. An omitted observation might even be the most interesting data point, were it to be investigated. Or it may be due to factors not associated with the other sample data, a possibility advanced by Belsley et al. (1980). There is no theory justifying winsorizing (or truncation). These methods also make replication decidedly more difficult.

Winsorizing is one example of inappropriate data manipulation practiced during data carpentry, together with data mining and data snooping. Other examples of inappropriate data activities involved in establishing the data set include omitting data obtained under different circumstances, such as from different companies, time periods, or locations. Data produced by different individuals operating under different procedures and dissimilar situations also should be assumed to be inappropriate. See Zeff (2016) for an example. Inclusion of any such data must be thoroughly vetted and disclosed, including its impact on the identified hypothesis.

Instead of winsorizing, we suggest the authors consider robust regression (RR) recommended by Leone et al. (2019). Using simulation, the authors find RR outperforms winsorization and truncation, which are largely ineffective. The authors suggest using approaches based on residuals for which RR is both theoretically appealing and easy to implement.

3. Testing the Model

An approach that some researchers are turning to—and which we encourage—is to apply the research model to alternative relevant data sets (Lindsay 1995). While this approach is time-consuming, the results, when thus confirmed, are more compelling. Additionally, tests might also be run on logical choices of subsets of the original data. An interesting alternative would be to test the model through one or more predictions, although we have not seen much enthusiasm for this option.

The data sets in accounting studies, other than experimental ones, tend to be large. Data sets over 10,000 are not uncommon. Small data sets, below 25, are unusual, except in behavioral accounting work, an area we are not explicitly targeting here. Accounting journals would not reject a small sample of, say, 25 if there were a compelling reason for the size and the results were clearly of interest. Multiple hypotheses based on a single data set are common, and the use of a data set to examine a different hypothesis by a different research team is not uncommon. The concern here, however, is that any data problem that is unresolved or miss-handled in the original research is likely to influence the new work. We believe that a borrowed data set must go through a thorough analysis before it can be presumed to be appropriate for testing a new hypothesis. This is appropriate for the original paper and even more so for a replication or the use of the sample by a new investigating team.

Our reading of the accounting literature indicates that some authors do rely on the same data set to test multiple hypotheses. Yet Floyd and List (2016, p. 454) observe, "When multiple hypotheses ... are considered together, the probability that at least some Type 1 errors are committed often increases dramatically with the number of hypotheses." See also Ohlson (2018). Fortunately, studies that perform multiple tests on a single data set are not difficult to identify. Authors should either alert readers to what amounts to over-testing or confine their analysis to the critical issue of the study. Additionally, it often happens that the ideal data set is impossible to obtain. When this is the case, the ideal data set should be acknowledged and its absence justified, including any change in the variables selected and their measures. Unfortunately, while applauding the changes in the reviewing process championed by Bloomfield et al. (2018) and others, we believe there is currently no way to assure that data tampering, including data mining, does not occur prior to the submission of a research paper to a journal.

Sample Size Concerns

Sample size has an important effect on statistical tests. Many accounting studies involve very large samples, while small samples are rare except in situations involving behavioral experiments. Indeed, researchers appear to believe that very large samples are somehow superior or more likely to generate statistical significance. Yet what the researcher observes in the sample may not be true at

the population level. This is particularly likely to happen when the sample data ultimately used in the test are substantially fewer than what are contained in the initial sample. This situation, while not common in accounting research, does occur. See Santanu et al. (2015). The authors reported a sample size of 11,262 after concluding that, for undisclosed reasons, 14,042 observations were rejected, a rejection level of 55% of the available data. This condition alone does not invalidate the research. Such a large reduction calls for an explanation, which was not given. Indeed, the authors were likely engaged in data snooping, which could lead to such a large reduction in sample size. Researchers should also not forget the Jeffery-Lindley paradox, which shows that, with a large enough sample size, a 0.05 significance result can correspond to assigning the null hypothesis a high probability (0.95). This result does not hold, however, for interval hypotheses. See also Ohlson (2015) on sample size.

4. Reporting Results

The most important point to be made here is not whether a reported significant *p*-value, say at the 0.01 level, has been obtained but rather the overall credibility of the work. Credibility depends on a myriad of factors. These factors include the accuracy and veracity of not only the model but also the variable choices and their measurement. For example, if the model were to omit an important explanatory variable, the effect of the omission may be subsumed under one or more of the other explanatory variables, causing it or them to appear more significant than would otherwise be the case. It remains the responsibility of the authors to consider the challenges that serious readers are likely to raise concerning the central model, the variable choices or omissions, and how they should be operationalized. Assuring that readers have an adequate description of the methodology, including the model, data set, and the computer protocol in order to permit, and indeed invite, a replication would be one template for ascertaining that the essential elements have been disclosed. Improperly executed research can, as pointed out most recently by Kim et al. (2018) and by Lindsay (1994, 1995), ultimately lead to poor decisions and may even inflict serious social harm.

4.1. Reporting p-Values

First, it is useful to define what a *p*-value is. A *p*-value is the probability of observing a value of the test statistic that is as extreme as, or more extreme, than the value resulting from the sample, given that the null hypothesis is true. It is both a conditional probability and a statistic with a sampling distribution. Our view of *p*-values' contribution to the research in accounting is best captured by the following quotation: "Misinterpretations and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific literature" (Greenland et al. 2016, p. 337). Authors of publications in accounting and related fields invariably report and rely on small *p*-values (0.01, 0.05, 0.10) as an indication of the importance of their work. Yet, there is no theoretical justification to guide researchers in selecting a specific *p*-value to justify the conclusion that significance has been in fact attained or, if so, whether it matters. The smaller the calculated *p*-value, the more comfortable the researcher may feel in rejecting the null hypothesis. However, this information alone considers neither the importance of the results nor the costs of an incorrect rejection. Johnstone and Lindley (1995) argue that significance at 0.05 is meaningless without knowing the sample size, the magnitude of the observed effect, and the operational importance of that effect. Alone, it fails to assure readers that the analysis has even uncovered a useful result. Indeed, the *p*-value, so endemic to much of accounting research, is useless by itself. Without some measure of the impact (size effect) or economic importance of the result, little if anything has been learned. Ziliak and McCloskey (2004), after reviewing the literature in several fields, found that nine out of 10 published articles make this mistake. If, in the accounting

literature for example, the reported results suggest that a specific behavior reduces audit delays, little is gained unless a reasonably accurate determination of the economic impact of the revealed delay to an identified clientele is determined. Authors, then, must identify in advance the user of the research result who would find the size or impact important. We note that several distinguished journals in other fields, including Basic and Applied Social Psychology, have banned the use of *p*-values, while others, including PLOS Medicine and the Journal of Allergy and Clinical Immunology, actively discourage its use. The American Statistical Association has recently urged caution in relying on statistical significance at the traditional 0.05 level as a basis for claims (Wasserstein et al. 2019). We continue to be dismayed that editors or reviewers in our field appear to require a reported *p*-value of 0.10 or less as a necessary condition for publishing the results of a study relying on statistical research in accounting. Perhaps the recent final appeal to renounce relying on *p*-values was sounded by a recent paper published in The American Statistician in which Wasserstein et al. (2019, p. 1) state: "Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof)." The same issue of the American Statistician includes 43 additional papers that addresses statistical Inference in the twenty-first century.

An improvement would be to report the Bayes factor, as suggested in the context of accounting research by Kim et al. (2018). This is the ratio of the observed value, or a lesser value under the null hypothesis, to the probability of the observed value or a lesser value under the alternate hypothesis. The approach is a Bayesian concept and reflects the ratio of the new knowledge to what was previously presumed. This concept is consistent with using new information to update one's prior beliefs, a common accounting objective. The calculation necessitates that the researcher initially specifies the denominator of the likelihood ratio. Furthermore, reporting a confidence interval is an improvement over reporting a *p*-value and provides more information.

Yet, it is wrong to conclude that *p*-values are useless. A *p*-value from a well-executed study could provide useful information. For example, such a value could indicate that there is something unusual or interesting in the analysis, justifying further study. Alternatively, a review of the process may indicate a flaw in the analysis. Perhaps the data or the data-carpentry activity was faulty. The model may be inappropriate. There could have been an error in the computer program. And well-done studies could suggest further potential regardless of the resulting *p*-value.

4.2. Effect Size (ES) or Economic Importance (EI)

Determining the effect size (Cohen 1990; Stone 2018) or the economic importance (Basu 2012; Dyckman 2016) of the results should be the sought-after objective of research. One way of presenting the result is to use a confidence interval measure to capture the impact on the specific costs or benefits suggested by the research. Yet we could locate very few articles in the accounting literature that have rcct size. Judd et al. (2017, p. 34) provide an example that does addresses this issue. "In terms of economic significance, we find on average, a one standard deviation increase in CEO narcissism [proxied by the CEO's picture size in the annual report and the CEO's relative cash and non-cash pay] is associated with a 2.4 percent to 3.3 percent increase in external audit fees, which equates to approximately $116,497 to $160,183 for our sample mean firms." We note here that the authors elect, not to emphasize the economic significance by reporting their findings in either the Synopsis or Conclusion. Their approach may be explained in part by the studies limitations described in their footnote two.

Irani et al. (2015, p. 847) provide a recent example of explicitly addressing the statistical significance/economic importance issue. They state, "We recognize the small magnitude of the univariate market reaction, which, although *statistically significant*, is arguably not *economically significant*" [emphasis added]. It is essential to identify the importance of a study's results and not just rely on whether one or more hypotheses are statistically significant at a specific reported *p*-value. Furthermore, as noted earlier, not finding a variable to be statistically significant does not necessarily mean it is unimportant. Eshleman and Lawson (2017, p. 75) report that they "find a positive association between audit market concentration and audit fees." Their main conclusion was, "As a whole, our findings

suggest that U.S. audit market concentration is associated with both higher audit fees and higher audit quality" (p. 76). Unfortunately, we are not informed of the importance of the impact that the concentration had on audit fees.

A more recent accounting study by Brown et al. (2018) recognizes the limitation of research that stops with the reporting of a significant *p*-value, yet the authors fail to deal with the economic importance.

5. Replication Studies

Researchers quite understandably seek to explore new questions in their research. Thus, it is not surprising that replication studies are rare. Yet, regardless of a study's results, whether it is an important finding or an unsuspected failure to support an expected finding, replication studies are relevant and important. Replications are, however, decidedly difficult to perform satisfactorily and are not welcomed by many journals across the accounting research landscape. Replications therefore must pass rigorous scrutiny. Several new journals have been launched, and a few existing outlets now do consider replication papers. New journals have been initiated recently that do consider replication papers. There are also existing journals that have published replications for some time. The *American Economic Review* and the *Journal of Applied Econometrics* are leaders in their field, and they publish about a 30 percent of the replications in economics (Reed 2018). The Replication Network is an excellent source for information on replication studies.

A few replication studies have begun to appear in accounting journals. The ability to fully replicate depends on an agreed theory. Stories do not provide ideal bases for replications. An early, well executed replication study in accounting that deserved and ultimately achieved publication was done by Bamber et al. (2000), who replicated Beaver (2000) Seminal Award-winning paper. It is interesting to note that the authors' work was published in *Accounting, Organizations and Society*, not in one of the journals noted for empirical/archival research. This is not where one would expect to find it, because it was rejected by the journal that published Beaver's article.

Mayo (2018, preface) sets a high bar when applied to any study, including replications. She advises that the results of any study need to be "severely tested." She writes, "The [severe] testing metaphor grows out of the idea that before we have evidence for a claim, it must have passed an analysis that could have found it flawed." In other words, as Mayo states, "a hypothesis must have passed an analysis that could have found it flawed" (Mayo 2018, preface). We are unable to locate an accounting paper that currently meets this or a similar rigorous standard. We would encourage researchers to consider applying this test.

The rewards for attempting replications are currently not enticing. Furthermore, precise replications have seldom been possible, in part because the necessary information to perform such studies is not ordinarily made available by authors. Nevertheless, we encourage replications because they provide the test that what has been found matters. This could be the case if a study were to identify a potential measurable and meaningful size effect. Unfortunately, in accounting, we are left with a sparse landscape of replication studies that provide confirmation of important results or which encourage the publication of synopses of important replications (particularly of effect size or economic importance) that are well-executed. Fortunately, journals do exist that consider replications. One relatively new journal is The International Journal for Re-Views in Empirical Economics (IREE).

A recent study by Peng (2015) concludes that a high proportion of published results across fields were not reproducible by replication. This does not reflect well on the academic community. Studies of reproducible results lend credence to the value of the exercise. On the other hand, failed replications cast doubt on the original results. Brodeur et al. (2018, abstract) report that in "Applying multiple methods to 13,440 hypothesis tests reported in 25 top economics journals in 2015, we show that selective publication and p-hacking is a substantial problem in research employing DID [differences-in-differences] and (in particular) IV [instrumental variables]." A large study reported in Science describes the results of 270 researchers replicating 100 experiments reported in papers published in 2008 in three high-ranking psychology journals (Aarts et al. 2015). The replications yielded the same results according to several

criteria. They showed that only 39% of the original findings could be replicated unambiguously. This information is not encouraging.

Recently, we came across an announcement of a new e-journal, SURE (for The Series of Unsurprising Results in Economics), which commits to publishing high-quality research even with unsurprising findings. The journal emphasizes scientifically important and carefully executed studies with statistically insignificant results or otherwise unsurprising findings. Studies from all fields of economics will be considered. As a bonus, there are no submission fees.

An additional process that can increase our confidence in results, and one that merits consideration, is meta-analysis (Dyckman 2016; Hay and Knechel 2017; Stone 2018). An advantage of meta-analysis is that it suggests the integration of current and future investigations of a given phenomenon. Using this technique reflects a cumulative approach to a specified hypothesis by which a triangulation on the topic can lead to a better understanding of a common research objective. This approach allows competing explanations of a given phenomenon to be merged to produce a result depicted, for example, by a confidence interval. Opting for a meta-analysis approach provides an opportunity for researchers to reexamine and perhaps reinforce important past results. The adoption of meta-analysis in accounting remains exceedingly rare, perhaps partially reflecting editor reluctance to publish replications.

6. A Critical Evaluation and a Way Forward

We believe that a healthy skepticism should abide from the beginning and then remain with the authors throughout the process. This attitude must extend to the apparent confirmation of any basic hypotheses. Asking why and how the authors could be wrong, or whether they have missed an important influence on the analysis, should accompany the entire investigation. We fear, however, that authors may not expend the time and human capital to critique their work sufficiently. Readers are often not sufficiently familiar with the authors' subject to discover a study's short-comings. Indeed, the authors themselves are the most likely to be conscious of their study's limitations and potential extensions. They should advise readers on where additional analysis would be most fruitful, including the known or suspected limitations to their own work. Essentially all studies have limitations, and that alone provides ample reason for encouraging disclosure, and in important situations, replications. Access to all that went into the original analysis should be available, on request if necessary.

The process begins with the selection of an important question or issue. A relevant and available data source will need to be identified or created in order to proceed. Once these are determined, we suggest that the investigation concentrate on operationalizing the dependent variable, identifying the independent variables, including their interrelationships, and how they can best be operationalized. This approach then allows the research team to craft the model. The research team should keep a record of all assumptions and decisions made in this process. Attention will need to be given to variable interactions, and whether the conditions affecting the observations are different in a way, or ways, that could have an impact upon the findings.

The primary objective is to reveal an important result, one that is based on an important economic or behavioral impact. If no such impact was revealed, the authors should take what has been learned, pack their bags and move on to a new project. If the process has been appropriate, the authors should take what has been learned and seek a new topic worthy of investigation. There is no reason to attempt to resurrect a deceased patient.

Authors should avoid placing reliance on *p*-values, concentrating instead on the economic or behavioral implications of the work. If the result is controversial and the analysis is well done, so much the better. Reporting confidence intervals instead of *p*-values should be the common practice. (Dyckman 2016; Stone 2018).

Our discipline, and others as well, will benefit from applying new approaches to establishing the importance of the phenomena being studied. Stone (2018, p. 113), has recently suggested exploring triangulation, a complementary approach that could, for example, combine a quantitative and a behavioral approach to a problem. Also see Jick 1979. Such studies can provide insights not otherwise

apparent. Combining methodologies has limitations, one being that replications are extremely difficult to execute. The adoption or reliance on new methodologies or pirating them from other disciplines is also to be encouraged.

Thus, we are in accord with Johnstone (1990) and with Kim et al. (2018, p. 14) that a Bayesian approach to statistical hypothesis testing, which recognizes the importance of the power of the test, offers a means of dealing with the inherent bias introduced by the conventional hypothesis testing currently prevalent in accounting. Furthermore, we encourage authors, as a few have done, to consider areas and methodologies from sister disciplines, including medicine and even philosophy. In this paper, we have relied on medicine (Ioannidis 2005), epidemiology (Greenland et al. 2016), and on philosophy (Mayo 2018). We maintain that there is much to be learned from these and other disciplines.

The ultimate importance of a study is the economic or behavioral consequences of the research findings, not the statistical significance as reflected in a calculated *p*-value. Investigators should be looking for a size or economic Importance measure. The *p*-value may provide some information, as described above. However, it should not be considered the study's goal or a measure of its contribution.

Finally, we would encourage accounting professors to improve their statistical knowledge. The resources are immediately available. In addition, universities hold excellent summer programs. One current example is an August program under the auspicious of Northwestern and Duke Universities.

7. Conclusions

Our purpose in this paper has been to encourage authors to consider ways to improve their research. Our paper began by emphasizing the importance of careful model building, data selection, and carpentry as well as the disclosures essential to assuring the integrity and reproducibility of the work. The choice of topic should be based on its relevance to practice or on improving research techniques.

Researchers must understand the limitation of relying on *p*-values as the sole or even primary support for their findings. Confidence intervals should replace reporting *p*-values. Until the reader is informed of the importance of the results, framed by the economic or behavioral consequences of the findings, the research objective has not been achieved. The issues addressed in this paper are advanced as a means of accomplishing this objective. To assist in the mission, we strongly encourage researchers to read the paper by Greenland et al. 2016, which offers a tutorial on statistical tests, *p*-values, confidence intervals, and power. (Greenland et al. 2016 is also available through Dave Giles' Blog of 2 May 2019 (Giles 2019) and Bob Jensen's Blog 2 September 2019 (Jensen 2018). The Jensen blog includes the article).

Author Contributions: Only the categories that apply to our paper are covered here. Unless indicated, the authors contributed equally. Data curation was mainly obtained from prior research papers of the authors and papers authored by others in and outside our field. The original draft was primarily undertaken by the first author (T.R.D.) while the second author (S.A.Z.) took the lead on project review and editing. The project was under our own supervision and there was no outside project administration. The project was entirely funded and written by the authors, T.R.D. and S.A.Z.

Funding: The project received no outside funding.

Conflicts of Interest: There were no conflicts of interest.

References

Aarts, Alexander A., Joanna E. Anderson, Christopher J. Anderson, Peter Attridge, Angela Attwood, Jordan Axt, Molly Babel, Štěpán Bahník, Erica Baranski, Michael Barnett-Cowan, and et al. 2015. Estimating the reproducibility of psychological science. *Science* 349: 6251.

Bamber, Linda Smith, Theodore E. Christensen, and Kenneth M. Gaver. 2000. Do we really "know" what we think we know? A case study of seminal research and its subsequent overgeneralization. *Accounting Organizations and Society* 25: 103–29. [CrossRef]

Basu, Sudipta. 2012. How can accounting researchers become more innovative? *Accounting Horizons* 26: 851–70. [CrossRef]

Beaver, William H. 1968. The information content of annual earnings announcements. *Empirical Research in Accounting, Selected Studies 1968. Supplement to Journal of Accounting Research* 6: 67–92. [CrossRef]

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley.

Black, Fischer. 1993. Beta and return. *Journal of Portfolio Management* 20: 8–18. [CrossRef]

Bloomfield, Robert, Kristina Rennekamp, and Blake Steenhoven. 2018. No system is perfect: Understanding how registration-based editorial processes affect reproducibility and investment in research quality. *Journal of Accounting Research* 56: 313–62. [CrossRef]

Boyd, Danah, and Crawford Kate. 2011. Six Provocations for Big Data. A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, September 21. Available online: https://ssrn.com/abstract=1926431 (accessed on 14 March 2019).

Brodeur, Abel, Nikolai Cook, and Anthony G. Heyes. 2018. Methods Matter: P-Hacking and Causal Influence in Economics. *Dated August 2018.* Available online: https://drive.google.com/file/d/10an9l3ndpjIfBVy1q5tC-9YGrVzPvmfg/view (accessed on 15 March 2019).

Brown, Jason P., Dayton M. Lambert, and Timothy R. Wojan. 2018. At the intersection of null findings and replication. *The Replication Network*. August 23. Available online: https://replicationnetwork.com/2018/08/23/brown-lambert-wojan-at-the-intersection-of-null-findings-and-replication/ (accessed on 23 August 2018).

Cohen, Jacob. 1990. Things I have learned (so far). *The American Psychologist* 45: 1304–12. [CrossRef]

Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press.

Dyckman, Thomas R. 2016. Significance testing: We can do better. *Abacus* 52: 319–42. [CrossRef]

Dyckman, Thomas R., and Stephen A. Zeff. 2014. Some methodological deficiencies in empirical research articles in accounting. *Accounting Horizons* 28: 695–712. [CrossRef]

Dyckman, Thomas R., and Stephen A. Zeff. 2015. Accounting research: Past, present and future. *Abacus* 51: 511–24. [CrossRef]

Eshleman, John Daniel, and B. P. Lawson. 2017. Audit market structure and audit pricing. *Accounting Horizons* 31: 57–81. [CrossRef]

Floyd, Eric, and John A. List. 2016. Using field experiments in accounting and finance. *Journal of Accounting Research* 54: 437–75. [CrossRef]

Giles, David. 2019. Blog. Available online: https://davegiles.blogspot.com/ (accessed on 5 February 2019).

Gow, Ian D., David F. Larcker, and Peter C. Reiss. 2016. Causal inference in accounting research. *Journal of Accounting Research* 54: 477–523. [CrossRef]

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. Statistical tests, *p*-values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology* 31: 337–50. [CrossRef]

Harford, Tim. 2014. Big data: A big mistake? *Significance* 11: 14–19. [CrossRef]

Hay, David C., and W. Robert Knechel. 2017. Meta-regression in auditing research: Evaluating the evidence on the Big N audit firm premium. *Auditing: A Journal of Practice & Theory* 36: 133–59.

Ioannidis, John P. A. 2005. Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association* 294: 218–28. [CrossRef]

Irani, Afshad J., Stefanie L. Tate, and Le Xu. 2015. Restatements: Do they affect auditor reputation for quality? *Accounting Horizons* 29: 829–51. [CrossRef]

Jensen, Robert. 2018. Blog. Available online: http://faculty.trinity.edu/rjensen/ (accessed on 2 September 2018).

Jick, Todd D. 1979. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly* 24: 602–11. [CrossRef]

Johnstone, David J. 1990. Sample size and the strength of Evidence: A bayesian interpretation of binomial tests of the information content of qualified audit reports. *Abacus* 26: 17–35. [CrossRef]

Johnstone, David J., and D. V. Lindley. 1995. Bayesian inference given data "significant at α": Tests of point hypotheses. *Theory and Decision* 38: 51–60. [CrossRef]

Judd, J. Scott, Kari Joseph Olsen, and James Stekelberg. 2017. How do auditors respond to CEO narcissism? Evidence from external audit fees. *Accounting Horizons* 31: 33–52. [CrossRef]

Kim, Jae H., Kamran Ahmed, and Philip Inyeob Ji. 2018. Significance testing in accounting research: A critical evaluation based on evidence. *Abacus: A Journal of Accounting, Finance and Business Studies* 54: 524–46. [CrossRef]

Kupferschmidt, Kai. 2018. A recipe for rigor. *Science* 361: 1192–93. [CrossRef] [PubMed]

Leone, Andrew J., Miguel Minutti-Meza, and Charles E. Wasley. 2019. Influential observations and inference in accounting research. *The Accounting Review*. forthcoming. [CrossRef]

Lindsay, R. Murray. 1994. Publication system biases associated with the statistical testing paradigm. *Contemporary Accounting Research* 11: 33–57. [CrossRef]

Lindsay, R. Murray. 1995. Reconsidering the status of tests of significance: an alternate criterion of Adequacy. *Accounting Organizations and Society* 20: 35–53. [CrossRef]

Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Ohlson, James A. 2015. Accounting research and common sense. *Abacus* 51: 525–35. [CrossRef]

Ohlson, James A. 2018. Researchers' Data Analysis Choices: An Excess of False Positives? Available online: https://ssrn.com/abstract=3089571 (accessed on 6 January 2019).

Peng, Roger. 2015. The reproducibility crisis in science: A statistical counterattack. *Significance* 12: 30–32. [CrossRef]

Reed, Robert. 2018. An Update on Progress of Replications in Economics. Available online: https://replicationnetwork.com/2018/10/31/reed-an-update-on-the-progress-of-replications-in-economics/ (accessed on 5 January 2018).

Santanu, Mitra, Hakjoon Song, and Joon Sun Yang. 2015. The effect of Auditing Standard No. 5 on audit report lags. *Accounting Horizons* 29: 507–27.

Stone, Dan N. 2018. The "new statistics" and nullifying the null: Twelve actions for improving quantitative accounting research quality and integrity. *Accounting Horizons* 32: 105–20. [CrossRef]

Wasserstein, Ronald L., Allen. L. Schirm, and Nicole. A. Lazar. 2019. Moving to a world beyond "p > 0.05". *The American Statistician* 73: 1–19. [CrossRef]

Zeff, Stephen A. 2016. "In the literature" but wrong: Switzerland and the adoption of IFRS. *Journal of Accounting and Public Policy* 35: 1–2. [CrossRef]

Zeff, Stephen A., and Thomas R. Dyckman. 2018. A historical study of the first 30 years of Accounting Horizons. *Accounting Historians Journal* 45: 115–31. [CrossRef]

Ziliak, Stephen T., and Deirdre N. McCloskey. 2004. Size matters: The standard error of regressions in the American Economic Review. *Journal of Socio-Economics* 33: 527–46. [CrossRef]