Is There a Confidence Interval for That? A Critical Examination of Null Outcome Reporting in Accounting Research

November, 2018

By

William M. Cready
The University of Texas at Dallas
cready@utdallas.edu

Jiapeng He
The University of Texas at Dallas
Jiapeng.He@utdallas.edu

Wenwei Lin
Xiamen University
Wenwei.lin@foxmail.com

Chengdao Shao
Xiamen University
cdshao@foxmail.com

Di Wang
Xiamen University
Wangdi.wendy@foxmail.com

Yang Zhang
Xiamen University
zhangyang_xmu@foxmail.com

Keywords: Methodology, Null Hypotheses, Accounting (Research) Quality

Is There a Confidence Interval for That? A Critical Examination of Null Outcome Reporting in Accounting Research

ABSTRACT

This study evaluates how null outcomes are analyzed and reported by accounting researchers based on an examination of two years of publications in *The Accounting Review*. As null outcomes reflect an inability to reject a null they, unlike rejections, do not lend themselves to specifically conclusive interpretations. Rather, substantive descriptive analyses are needed to draw useful inferences from such outcomes. In the 35 articles we identify as presenting substantive null outcomes, however, inappropriately conclusive interpretations of these outcomes are widespread while scant attention is given to providing the descriptive analyses needed to draw useful insights from them. Moreover, these deficiencies span articles published across all of the major accounting research areas (i.e., financial, managerial, audit, and tax) and encompass both archival and experimental designs. The analysis also proposes the use of descriptive techniques, particularly interval based analyses (e.g., Dyckman and Zeff, 2014; Dyckman, 2016)), as a desirable alternative for interpreting null outcomes.

I. INTRODUCTION

This article presents an analysis of how the academic accounting literature reports null outcomes. We define a null outcome as an instance where the statistical significance of two-tailed test of a null hypothesis results in p-values that are not small enough to be deemed statistically significant at conventional levels. From a classical hypothesis testing perspective such an outcome is taken as uninformative. It, in particular, does not indicate that the underlying null is true, only that there is an insufficient basis for reliably concluding that the examined evidence is inconsistent with it.[1] Alternatively, from more descriptive perspectives, an inability to reject the null is broadly consistent with conjectures that any underlying effect is either absent or possibly small. However, the academic accounting literature, as will become very apparent in the analysis that follows, generally frames its analyses within the structure of classical hypothesis testing (see Dyckman and Zeff, 2014), and classical null hypothesis testing is a rather limited vehicle (i.e., it is very identification of what is not consistent with the evidence oriented) for conducting meaningful descriptive analyses. Consequently, the analyses accompanying null outcomes in this literature are, in the articles we examine at least, uniformly materially deficient and the interpretations provided to them are commonly misleading.

The analysis is based on a comprehensive examination of all articles published in *The Accounting Review* over the 2016-2017 time period. We identify 35 papers reporting null outcomes that were central to the paper's analysis (evidenced by being connected to a formal hypothesis statement or being discussed in the article's abstract or its introductory sections). The analysis evaluates these articles on three distinct dimensions: (1) relevant descriptive information (not)

[1] See principles 1 and 2 of the *ASA Statement on Statistical Significance and P-Values"* (Wasserstein and Lazar, 2016).

examined regarding the null outcome; (2) the degree to which the paper interprets the null outcome in a conclusive manner; (3) and, for a selected subset of outcomes, alternative confidence interval based interpretations of the reported evidence. Collectively, these examinations indicate that in the accounting literature null outcomes: (1) are rarely accompanied by examinations of relevant descriptive information (indeed, such information is often not discussed at all); (2) are without exception assigned inappropriately conclusive interpretations; and, (3) are much more substantively interpreted by using straightforward descriptive analysis based on confidence intervals.

Our analysis is related to recent work by Dyckman and Zeff (2014), Kim and Ji (2015), Ohlson (2015), Dyckman (2016), Kim, Ji, and Ahmed (2017), Harvey (2017), and Stone (2018). The focus of these studies, however, is on the reliability of null hypothesis rejections, including the relevance of accompanying supplemental procedures for judging the integrity of such rejections. They are particularly concerned with the deficiencies of hypothesis based testing analysis relative to more descriptive based methods and the importance of replication as a mechanism for instilling confidence that null hypothesis rejections are not spurious. Our analysis differs from this literature in that it focuses on the integrity of interpretations provided for null outcomes. Unlike rejections, null outcomes are non-events from a hypothesis testing perspective. Interpretation of null outcomes is a purely descriptive, not a test of hypothesis, challenge.

On one dimension, however, this study does have much in common with this integrity of rejection literature, particularly with arguments found in Dyckman and Zeff (2014, 2015) and Dyckman (2016). These studies observe that the general absence of statistical interval reporting is a significant shortfall in the existing accounting literature. As Dyckman and Zeff (2015) note, such intervals enable understanding from the perspective of "where we are led by the data to believe the finding of interest is to be found." (p. 520) A core contribution of our analysis is that it demonstrates that the literature does not seek understanding in this fashion, even when it is the only available approach to attaining much of any understanding at all.

The second core contribution of our analysis follows directly from the recent *ASA Statement on Statistical Significance and P-Values* (Wasserstein and Lazar, 2016). This statement was promulgated by the American Statistical Association in response to a belief that "(the *p*-value) is commonly misused and misinterpreted." Its stated purpose is to provide a "formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value." The *Statement*'s introduction concludes by asserting that it "articulates in nontechnical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community."

Principle number 2 in the *Statement* speaks to the general question of inferring the truth of the null hypothesis based on *p*-value outcomes. It reads, in full, as follows:

> ***P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.** Researchers often wish to turn a *p*-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The *p*-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

Principle Number 6 provides further null outcome specific clarification on this point when it states that "a relatively large *p*-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data."

Our analysis empirically documents that the accounting literature routinely employs high p-values and/or low value test statistics as a basis for advancing strong claims favoring the truth of underlying null hypotheses. That is, the literature's approach to null outcome interpretation (i.e., non-small p-values) is starkly at odds with the relevant guidance from the *ASA Statement*.

II. CLASSICAL HYPOTHESIS TESTING AND NULL OUTCOMES

Classical null hypothesis testing flows from the well-know if-then logical paradigm. A specific premise (null) is asserted to be true and that premise is evidenced by a necessary "then" outcome. In this paradigm the demonstrated absence of the necessary outcome negates the premise. Alternatively, and of direct relevance to null outcome interpretation, using the "then" outcome in and of itself as indicative of the truth of the premise is a well-known logical fallacy—affirming the consequent (e.g., see pp. 83-85, Damer, 2013). Simple observation of a necessary or expected outcome does not prove the antecedent that suggests it, since the same outcome may follow from many other premises. Hence, as a matter of simple logic, leaving aside any notion of how statistical inference draws upon such logic, using observed consequent consistent outcomes to infer anything about the plausibility of a null hypothesis, to say nothing of its truth, is problematic.

Statistical applications of this logical framework further complicate matters by introducing the notion of random error into the paradigm, meaning that the consequent is noisy, not truly observed. Hence, in the statistical setting noise is an explanation for any failure to observe a necessary outcome. Consequently, we are never in a position to falsify the premise with certainty. Instead, we rely on *p*-values to make rejection, or better, incompatibility with the evidence assertions. In the case of null outcomes, allowing for noisy measurement gives rise to irrefutable alternative hypotheses that are consistent with the evidence (some even more consistent with it), yet fundamentally contradict the postulated null. And, the *p*-value says nothing useful at all about the seriousness of this alternative hypothesis issue. This thinking is clearly seen in Principle 6 of the *ASA Statement* when it cautions against using a high *p*-value as a basis for inferring that a tested null hypothesis is true since "many other hypotheses may be equally or more consistent with the data." And, it is for these logical and practical reasons that in many disciplines null outcomes are routinely ignored entirely or dismissed as entirely inconsequential with language such as "unable to reject."

A closely related setting that also touches upon the affirmation of the consequent fallacy arises with respect to interpreting the alternative research hypothesis that typically accompanies a rejected null hypothesis. However, there are important distinctions between this sort of exercise and null affirmation interpretations. First, these interpretations begin from the perspective that the null hypothesis assertion is highly inconsistent with the examined evidence. Consequently, the analysis reasonably excludes a particularly relevant hypothesis (or, better, collection of hypotheses) from further consideration. In contrast, a null outcome in and of itself identifies no hypothesis that is inconsistent with the examined evidence. And, in fact, implicitly recognizes that the null and research hypothesis(es) are both consistent with the evidence.

Second, in the classical hypothesis testing structure the alternative research hypothesis is not, despite commonly encountered assertions to the contrary, directly tested. It is not presumed true and held up for falsification. Rather, it is posed as a reason, typically with some possibility of being valid, why the null premise may not be true. Hence, the alternative research hypothesis is accepted, not proven, when the null is rejected. In some cases, it is the only (often as a matter of definition) plausible alternative explanation.[2] In others, the accompanying analysis goes to great

[2] For example, if the null is that the effect is less than or equal to zero and the alternative is simply that this null is not true then falsification (setting aside for the moment that the presence of random error negates the absolute proof

lengths to rule out or control other alternative explanations for the null hypothesis rejection.[3] These analyses advance the case for the alternative in the form of an if and only if (i.e., "sufficient") argument. That is, they aim to persuade the reader that the research hypothesis is uniquely consistent with the examined evidence. At this point it is, of course, up to the reader to judge the persuasive merits of the presented case, recognizing that the answer involves considerations that go well beyond whatever *p*-values are in play.[4]

A Null Outcome Illustration

The preceding discussion is rather philosophical and hence abstract. It is sensible, therefore, to complement it with a more empirically grounded illustration as a means of demonstrating its more salient points. For this purpose we use a null outcome presented in a recent publication by Kim and Klein (2017). [5] This null outcome pertains to a test of whether there was an economy-wide change in the market value of firms in response to the passage of a rule change imposing mandates on audit committee composition and independence. The associated null hypothesis is motived by "market theory" which generally suggests that if the mandated changes are value-increasing then value-maximizing firms would have already changed voluntarily. Hence, the rule change should not be beneficial (increase market value) and may well be detrimental (decrease market value). In if/then terms then, the core assertion here is that if the rule change is

of anything) of the null proves the alternative. In contrast, if there is some sort of specific reasoning for why the effect is expected to be positive, falsification of the null does not prove the truth of this specific reasoning. Other reasons almost certainly exist for why the effect is not less than or equal to zero.

[3] Such refinements commonly are incorporated into the tested null hypothesis. (E.g., testing for no effect after controlling for a relevant correlated variable.)

[4] It is important to emphasize here that reaching a definitive inference based on judgement of the evidence is inherently inapplicable to interpreting null outcomes in statistical analysis settings. It will not work because it is impossible, by construction, to claim that contrarian alternatives to the advanced null hypothesis are even possibly unlikely in null outcome settings.

[5] Our choice of Kim and Klein here is, in part, motivated by the fact that it was selected for broader dissemination via an American Accounting Association press release captioned as" Longstanding mandate on corporate audit committees yields no benefit for investors, new research finds." (AAA, 11/1/2017) That is, it seemingly is viewed as a particularly noteworthy accounting research contribution.

not, on average, beneficial then the observed market response to the rule passage will be zero (or negative). The associated alternative research hypothesis stems from "entrenchment theory," which suggest that entrenched managers sacrifice market value to maintain and exploit their entrenched status. Forcing such entrenched manager firms to change their governance structure by means of the rule change may increase firm value, providing a rationale for questioning the market theory based if/then assertion.

The null outcome here is the absence of a statistically significant relation between firm market values in response to the rule change event. So, what does this outcome actually tell us? Well, if we follow the generally accepted test of hypothesis guidance for such interpretation, it tells us very little. We have no reliable basis for thinking that the examined evidence is inconsistent with the market theory no net positive benefit null. However, neither can we say that the examined evidence is inconsistent with the postulated alternative entrenchment theory hypotheses (i.e., positive values for the market valuation impact from the rule change.) Hence, based solely on the null outcome, we really can't say much of anything at all in the way of reliable inference here.

In light of this "nothing much to be said" takeaway and our study's focus on how the literature interprets such null outcomes, the actual interpretation provided for this null outcome by the article itself is also of considerable interest. The article's introduction starts this exercise with the observation: "We find, on average, no statistically significant cumulative abnormal returns…" (p. 188). This statement is correct, but says nothing substantive. Searching for nothing and finding it is hardly a difficult or inherently meaningful accomplishment. The sentence following this statement, however, expands upon this non-finding, claiming that "Thus, the market assigned no overall net benefit or cost to compliance." This assertion is an affirmation of the consequent. It is fallacy. Moreover, at even a basic descriptive level, it is a highly misleading interpretation. Table

1 of the article reports the examined evidence underlying these interpretations. The estimated average per event date effect (across 8 events) is a +4.8 basis point increase in market value to the implementation of the rule. In other words, the entrenchment theory alternative (i.e., that there was an increase in value due to the net benefit of the rule) is not merely consistent with the examined evidence, it is actually better supported by the evidence than the no effect null that is being put forth as truth here.[6]

III. THE DESCRIPTIVE PERSPECTIVE

Descriptive approaches to conducting statistical analysis are commonly advanced as an alternative or supplement to classical null hypothesis testing. The *ASA Statement*, for instance, discusses the relevance of other approaches, including "methods that emphasize estimation over testing," It, in particular, identifies "confidence, credibility, (and) prediction intervals" as examples. The *Publication Manual of the American Psychological Association* (2013) states "APA stresses that NHST (Null Hypothesis Statistical Significance Testing) is but a starting point and that additional reporting elements such as effect sizes, confidence intervals, and extensive description are needed to convey the most complete meaning of the results." (p. 33) In the specific case of null outcomes, Aberson (2002) indicates that "presenting results that 'support' a null hypothesis requires more detailed statistical reporting than do results that reject the null hypothesis."

[6] If, in particular, the null hypothesis is reformulated to be that the estimated effect of the rule is greater than zero (an unconventional but allowed phrasing for a null), then it would not be rejected either. Based on the article' affirmative approach to interpreting null outcomes the table 1 evidence should be taken as indicating that the market did indeed assign a net benefit to the rule change. An inference, that is likely consistent with the priors of the rule-makers who determined that the rule was needed. The more general lesson here, however, is that null affirmation interpretation paves the way for different researchers (or even a single researcher) to draw contradictory inferences from the same body of evidence simply as a matter of how the null is phrased.

In this analysis we advance confidence intervals as a particularly insightful approach for obtaining meaningful insights in null outcome settings. While confidence intervals are estimated from many of the same underlying constructs employed in null hypothesis testing (e.g., standard errors, alpha levels, estimated effect values), they do not center the analysis around specific hypotheses (i.e., the null hypothesis, in particular). Instead, a confidence interval identifies a range of effect values (or hypothesized effect values) that are plausibly consistent with the evidence, given some pre-set tolerance level for the acceptable level of uncertainty in this determination.

We have three reasons for focusing on confidence intervals. First, they are widely accepted and understood. Their determination is typically covered in introductory level statistics courses. The *Publication Manual of the American Psychological Association* (2009), in fact, states "complete reporting of all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are the minimum expectation for APA journals." (p. 33) Aberson (2002) argues that "reporting confidence intervals allows for stronger conclusions about the viability of null hypotheses than does reporting of null hypothesis test statistics, probabilities, and effect sizes." Second, confidence interval outcomes are readily mapped to null hypothesis testing outcomes. Specifically, for a given critical *p*-value a null hypothesis rejection corresponds to a setting where the hypothesized null value (or range of values) falls outside the confidence interval based on that same *p*-value. Alternatively, a null outcome corresponds to a setting where the hypothesized null value falls within this confidence interval. Third, relevant confidence intervals can generally be determined from values reported in the performance of null hypothesis testing. Hence, we can readily benchmark how studies interpret null outcomes relative to the broader descriptive understandings of such outcomes provided by the associated confidence intervals.

In appreciating the relevance of confidence interval analysis for null outcome settings it is particularly important to recognize that any associated null hypothesis testing analysis has, at this point, come up empty handed. The null hypothesis, which is the object of the test, is not inconsistent with the evidence. But, neither is the alternative research hypothesis, which is the rationale for questioning the null to begin with, inconsistent with the evidence. In fact, within the rigid framework that such testing operates, it follows that a null outcome has no inconsistent-with-the-evidence implications whatsoever for any considered hypothesis. Hence, given a null outcome, the immediate interpretative inference obtained from the confidence interval is an identification of a set of hypotheses or beliefs that are highly inconsistent with the evidence. And, if the confidence interval does no more than this then it is doing far more in terms of inference than is achieved by merely reporting a statistically insignificant outcome.

The confidence interval can also serve as a foundation for drawing more substantive descriptive inferences about effect magnitudes, as well as the general suitability of the analysis for drawing meaningful inferences. In particular, if a confidence interval is contextually narrow (i.e., standard errors are low and test power is high) then it may be feasible to advance an argument that the set of evidence consistent values in the confidence interval contains no consequential or substantively meaningful alternative (to the null) hypothesis values. For example, in the case of a no effect null, the confidence interval may indicate that the upper and lower bounds are both inconsequentially small. Importantly, a clearly necessary component of such a determination is a discussion and analysis pointing out why these bounds, in the context of the studied issue, are plausibly characterized as being “small” or inconsequentially different from the null value. Alternatively, the confidence interval may turn out to be very wide. This width indicates that the study is effectively devoid of power. Absent power, there is little reason to view the associated analysis as a serious attempt at testing any null hypothesis, or describing the relevant empirical landscape.

Finally, there are those outcomes where the associated confidence interval is not particularly wide, but neither is it narrow enough to reasonably argue that the set of evidence-consistent values are effectively indistinguishable from the null hypothesis value. Here, at a descriptive level a study can reasonably characterize a large number of alternative hypotheses, including many of the larger values that are consistent with the alternative research hypothesis, as being highly inconsistent with the examined evidence. Hence, the reporting study is providing some relevant insights about the empirical landscape of interest. Such identification may also point the way for more targeted and powerful approaches that narrow the set of evidence-consistent alternative hypotheses even further. It also indicates a need to look to aggregate evidence across studies using meta-analysis techniques. However, in engaging in such efforts, it is important to recognize that as the confidence interval narrows the ultimate outcome may well shift from null outcome to null hypothesis rejection (as power increases the likelihood of correctly rejecting false nulls).

Given this confidence interval perspective on interpreting null outcomes, we now revisit the Kim and Klein null outcome regarding the general market-wide response to the rule change they examine. The estimated average value they report is 0.048% per event date. As there are eight event dates the cumulative return over all eight dates is at least 0.384%. (8*.048), or 38.4 basis points.[7] The standard error for the per event date average is around 0.066, or 6.6 basis points. We

[7] Each event date consists of two days. And, per equation (1) of the article, they estimate return effects per day. It is unclear whether or not Table 1 doubles these coefficient estimates to produce true two day cumulative returns, or is simply reporting event-specific equation (1) coefficient estimates (i.e., averages of the two day event date returns). For purposes of this discussion we assume that the table is reporting two-day cumulative returns. We also, take the text assertion that the average return effect is +0.48% rather than +0.048% to be a typographical error.

determine this value by dividing the effect magnitude by the associated reported t-value (.048/.73). As the interest here is in the sum over the 8 events, the estimated standard error for the sum is .524% (8*.066), or 52.8 basis points. Hence, a two-standard deviation (which typically corresponds to a *p* value of slightly less than 5%), confidence interval for the overall effect has a lower bound of -67.2 basis points and an upper bound of 144 basis points.[8] Hence, alternative hypotheses that the net effect of the rule change was to increase firm values across the board by 100 or more basis points are not inconsistent with the examined evidence here. As, it is quite difficult to conceive of how 100+ basis points is somehow small or inconsequential, particularly when extended across the market as a whole, the viable descriptive inference here is that it is unlikely that the net impact of the rule change was extraordinarily positive.[9] Moreover, the examined evidence is most certainly not a descriptively sound basis for claiming that the market assigned no, or even little, benefit to the rule change.

IV. EMPIRICAL ISSUES

The preceding discussion raises two specific empirically addressable issues with respect to null outcome analysis reporting in the accounting literature. First, any sort of substantive interpretation of such an outcome requires empirical analysis and support that goes well beyond

[8] We employ two standard deviation confidence intervals to conform with the traditional emphasis on the .05 p-value dividing line between null and rejection outcomes. Apart from null hypothesis testing dogma, however, there is no compelling reason to employ such wide intervals. For instance, the well-known cone of uncertainty prediction intervals (a form of confidence interval) used in hurricane forecasting employ approximately one standard deviation confidence intervals for forecasts of “center path of the storm” tracks. Hence, this confidence interval exercise is rather tolerant of failing to include possible but comparatively unlikely paths in the prediction intervals they report. .

[9] In a follow-on analysis, Kim and Klein focus on the differential event period cumulative return effects between affected and unaffected firms. In their model (1), reported in table 4, the two standard deviation confidence interval for this difference has a lower bound of -14.6 basis points and an upper bound of 5.8 basis points. Hence, there is a strong descriptive case for asserting that the differential effect across these two groups is likely quite small. However, this evidence is relevant to the overall effect only under the further assumption that unaffected firms were in no way benefited by the rule. There may well be broader social welfare gains from the rule, obtained from leveling up the reporting and control playing field for instance or by making it impossible for in-compliance firms to revert to an out-of-compliance state. If so, then in-compliance firms would also benefit from the rule change. Consequently, this analysis is addressing a distinctly different question--the differential benefit from the rule change, not the overall unconditional benefit of the rule change.

the *p*-values and associated test statistics that led to the null outcome determination. Hence, we are interested in empirically documenting what sorts of analyses are, or are not, being provided in terms of providing relevant supporting evidence relevant to interpreting null outcomes. In particular, are null outcome interpretations based on nothing more than a high *p*-value or low test statistic value? Or, do they draw upon further empirical evidence such as that available from estimated effect magnitudes, estimate standard errors, or confidence intervals?

Second, we are interested in empirically documenting what sorts of interpretations the literature is providing for null outcomes. Nickerson (2000) identifies a number of false beliefs encountered in the context of classical hypothesis testing. One of these is the "Belief that failing to reject the null hypothesis is equivalent to demonstrating that it is true." (p. 260) Hence, particularly in light of the previously discussed problematic Kim and Klein null outcome interpretation, we are interested in whether null outcomes are interpreted in ways that are consistent with a the examined evidence. Or, are they commonly provided with overly conclusive interpretations? Interpretations that are unsupported by any of the reporting study's underlying examinations and analyses of the empirical evidence.

V. EMPIRICAL ANALYSES OF NULL OUTCOME REPORTING

Identification of Published Null Outcomes

Our empirical analysis is based on a set of null outcomes reported in the literature that we identified by means of a comprehensive examination of articles published in *The Accounting Review* in 2016 and 2017. We conduct our empirical analyses of how null outcomes are analyzed and interpreted at a fairly in-depth level. Indeed, the empirical analyses we present are arguably more qualitative than quantitative in nature. The fundamental observational unit here is, in fact, a complete research article. Large sample approaches are not particularly feasible and most certainly not cost-effective for analyzing full article (text) data points. Moreover, an overly large sample size would likely actually degrade the inferential validity of the analysis. Currently, very specific analysis of every single identified article is provided somewhere in our paper (inclusive of provided appendices). One can do this effectively for a limited number of articles, but not for large numbers. The analysis makes extensive use of specific examples discovered in the sample. The generalizability of this illustration by example approach actually decreases as the sample size increases since it is far easier to identify an intriguing collection of oddities in a sample of say 1,000 than in a sample of only 35. Finally, the small sample size together with the article level nature of the analyses allow any reader to readily evaluate or critique the underlying bases for our inferences on an article by article basis.

Based on the thinking that the analysis be based on relatively manageable number of null outcomes we next considered how to identify such a set in a fashion that, in particular, would mitigate criticism that we had "cherry-picked" articles for study. That is, given the nature of the analysis it seemed particularly desirable that we strive to make the article selection task both replicable (and analyzable) and free from (arbitrary) researcher choices as possible. Hence, we opted to examine articles from a single journal using a time period with natural start and end points. We also felt that the sample so-selected be representative of high quality accounting research, as broadly defined as possible. Given these objectives, *The Accounting Review* strikes us as the clearly obvious journal choice. It is the flagship journal of the world's largest association of accounting academics. It also has an extraordinarily diverse editorial board and employs a comparatively decentralized editorial decision-making structure. So, while we recognize that the direct generalizability of our analysis pertains to the population of articles appearing in recent issues of *The Accounting Review*, we also take the state of this population as the most reasonable single journal-based proxy for the general state of the accounting literature.

We identify articles reporting null outcomes based on three separate examinations of every article published in *The Accounting Review* over the 2016-17 time period (128 articles in total, 113 of which employ null hypothesis testing methods). Based on these examinations all reported null outcomes, regardless of importance, were identified in each paper. We then separated these outcomes into those that were deemed to play a central role in the paper and those that were not deemed to play such a role. Null outcomes that directly pertained to paper identified hypotheses and research questions as well as outcomes mentioned in a paper's abstract and introductory (pre-empirical analyses) sections are taken to have central roles while outcomes pertaining to robustness tests, validity checks, or control variables were not unless specifically identified in article abstracts or emphasized in article introductory sections. We located a total of 63 such central null outcomes from 35 separate articles based on this process. Appendix B provides a complete listing of the underlying research questions and hypotheses associated with these outcomes.

Table 1 lists the set of null outcome articles we identified along with some initial article level descriptive information. The set of articles span the major empirical research areas in accounting (auditing, financial, managerial, and tax) and encompass both archival and experimental studies. 14 of the studies address auditing, possibly indicating a predisposition for null outcomes in this domain. However, a contributing explanation is that auditing articles were particularly prevalent in *The Accounting Review* during the time period we study. Most of the articles (20) contain only a single null outcome. However, in one instance an article contains 7 null outcomes while another contains 5. 43 of the 63 outcomes are referenced in the article abstracts, and the abstracts of all but 6 of the articles contain some sort of explicit null outcome-based inference. As the abstract is highly prominent and word count restricted, these choices to abstract reference null outcomes indicate that the authors view them as speaking to important aspects of their articles. It is also inconsistent with the classical hypothesis testing perspective that null outcomes lack inferential merit. Finally, an explicit statement of the associated null hypothesis or prediction accompanies only 21 of the 63 null outcomes.

Evidence on Descriptive Analysis Provided for Null Outcomes

For our purpose, relevant descriptive analysis for a null outcome involves most any effort by an article that goes beyond a tabulated presentation of the estimated effect magnitude accompanied by a test of null hypothesis produced high *p*-value or low test statistic value. Such analysis may be nothing more than stating the estimated effect magnitude in the text discussion of the null outcome or a textual assertion that the estimated effect magnitude is small. Additionally, it may engage with the question of the precision associated with the estimated effect by either formally tabulating the associated standard error or, better, presenting and discussing the standard error magnitude in the body of the article. Finally, it may go so far as to present and analyze confidence intervals for null outcome linked estimated effects.

Table 2 presents the sorts of descriptive statistics that accompany each of the identified null outcomes we examine by article. Arguably, the most pertinent comprehensive descriptive statistic for a null outcome is a low p-value based confidence interval (CI). However, this key statistic is never reported.[10] Indeed, we were unable to locate a substantive discussion of the range of possible underlying null outcome consistent values in any of the 35 articles. Another item of descriptive relevance is the estimated standard error of the estimated null effect. If this standard error is "large" then the set of possible underlying effect values that is consistent with the observed

[10] Interestingly, Humphreys, Gary, and Trotman (2016) does report confidence intervals for several of its null rejection outcomes (p. 1457), but not for either of its two null outcomes.

outcome is also large, while if the standard error is "small" then this set of values is also small or, more to the point, precise. The standard error is, in particular, a readily obtainable measure of the underlying power of the statistical test. Hence, we might reasonably expect articles reporting null outcomes to be particularly keen about it. They are not. Only 5 of the 35 articles report standard errors of estimated null outcome effects. And, in these cases the standard error is simply tabulated. None of the articles mentions the standard error of the estimate in its text discussion.

The final three columns of table 2 focus on the text discussions of null outcomes with a focus on the degree to which articles pay attention to the most basic descriptive statistic, the estimated magnitude of the effect. Here again the level of omission is striking. Only 12 of the articles incorporate specific values of the estimated effects in their text discussion. And, the majority of these discussions are superficial (i.e., the effect is simply reported with no substantive accompanying interpretation as to why it should be taken as "small"). Another 5 articles describe the tabulated effect size without mentioning its specific magnitude. These claims uniformly take the form of unsupported assertions that the estimated effect is small. In contrast, 4 articles both present the estimated effect magnitude in the text and provide additional discussion of it. These discussions all involve somehow comparing the estimated effect magnitude to a relevant benchmark value (i.e, that the estimated effect is much smaller than the benchmark value).

Thirteen of the articles also specifically mention numeric p-value or test statistic values (e.g., t-values) for null outcomes in their text discussions of these outcomes. As stated in Principle 1 of the *ASA Statement* such values do reflect the degree of consistency between the null outcome and the underlying evidence. However, it is not clear how much they add on this dimension relative to the underlying estimated effect values.

The general absence of descriptive textual engagement with estimated effect magnitudes associated with null outcomes documented here is also notable in light of the sorts of descriptive analyses commonly provided for effect magnitudes for null rejections. When a null is rejected and the associated alternative is accepted then a common next step is a descriptive demonstration that the effect size of the alternative is "economically significant" or, more generally, large enough to care about.[11] For example, in the set of analyses we examine DeFond, Lim, and Zhang (2016) argue that the negative relation between client conservatism they document is "economically important" because, based on one measure, moving from the bottom to top conservatism decile reduces audit fees by 29%.[12] In an untabulated analysis, we found that such "it is large" substantive descriptive assessments are found in well over half of the articles reporting null hypothesis rejections published in *The Accounting Review* over the 2016-17 period. We found little evidence of any sorts of parallel "it is small" descriptive assessments for null outcomes in the articles examined here.[13]

Evidence on Null Outcome Interpretation

We examine how the question of how articles interpret null outcomes by reviewing and identifying associated interpretative statements provided by each null outcome reporting article's text discussion. We then classify each of these statements into one of the following five categories:

[11] Stone (2018), in fact, based on the premise that almost all (point) null hypotheses encountered in the accounting literature are truly false, argues that effect size is the actual relevant question in most null hypothesis rejection settings.

[12] Interestingly, DeFond et al. (2016) also report a parallel analysis of the relation between unconditional conservatism and audit fees that leads them to conclude that "auditors do *no*t strategically respond to *unconditional* conservatism by adjusting their fees" (emphasis theirs). They, however, provide no discussion as to why the magnitudes they document should be considered small. That is, in a setting where the relevant literature Aberson, 2002) argues that substantially more descriptive analysis is needed, we do the opposite, providing far less (arguably "zero") descriptive analysis for null outcomes than for null rejections.

[13] In the set of articles we examine, we encountered some form of coefficient "smallness" discussion of effect magnitudes in only four of them: Drake et al. (2016), Lennox (2016), Henry and Leone (2016), and Robinson et al. (2016).

Precisely Conclusive (PC); Generally Conclusive (GC); Selectively Conclusive (SC); Arguably Conclusive (AC); and Non-Conclusive (NC).

PC statements are those that are highly conclusive of the null being exactly true. Commonly, as is seen in the previous discussion of the Kim and Klein analysis, such statements present the no effect outcome as indicating that there is truly no effect at all.[14] For instance, Lennox (2016) states that he finds "*no change* in audit quality." Similarly, Choi et al. (2016) claim that "performance *does not differ* between … tournaments." (emphasis in both quoted statements ours). However, another form that such statements take is as denials of the validity of the alternative hypothesis. Such assertions range from assertions by Wieczynska (2016) that a null outcome constitutes a rejection of the alternative hypothesis (p. 1269) to somewhat milder assertions that an outcome is "inconsistent with" the alternative (e.g., Guenther et al., 2017; Kim and Klein, 2017; Lourenco, 2016).[15]

As is clear from even a very narrow reading, PC interpretations of null outcomes violate Principles 2 and 6 of the *ASA Statement*. High p-values are not an acceptable evidential basis for concluding or even inferring that a null hypothesis is true, or inferring that an alternative hypothesis is not true. Moreover, this basic structure to the inferential dimensions of null hypothesis testing is not at all new. It is, as discussed earlier, foundational to the logical structure underlying null hypothesis testing.

[14] In a few instances the precisely conclusive phrasing is accompanied by less conclusive qualifications such as "indicates", "implies." In general, we ignored such qualifications in our classifications since, as a matter of semantics, they do not negate the precisely conclusive language that follows. And, there were not a sufficient number of these qualifications to merit a separate category (e.g., precisely conclusive with qualifications).

[15] An attribution of "inconsistency" has the appearance of lacking a high degree of conclusiveness. However, the problem here is that a null outcome is not, absent additional descriptive insights, at a very fundamental level possibly inconsistent with anything. That is, one is not be able to reliably reject a hypothesis that the effect is positive, that it is negative, or that it is zero.

We identify a null outcome interpretation as Generally Conclusive when it advances the notion that the tested null hypothesis is approximately true. Claims that the effect is "small", "similar", or "comparable" fall into this category. We also include claims of insignificance in this category when the discussion provides no accompanying context indicating that it is specifically discussing statistical significance. A key distinction between statements that are classified as GC relative to those that are classified as PC is that effective descriptive analyses (something the table 2 evidence indicates is almost entirely absent in the set of articles we examine) could provide a basis for justifying GC interpretations. That is, a descriptive analysis, particularly one in the form of confidence interval presentation and discussion, could very well plausibly establish the interpretation that an effect is not large or dissimilar.

Another common approach to interpreting null outcomes is to state that they are consistent with or supportive of the null hypothesis or that they are not supportive of the associated alternative hypothesis. We label these sorts of statements as Selectively Conclusive because from a descriptive perspective they are, essentially, cherry picking the set of available individually arguably acceptable (but incomplete) descriptions for the null outcome. This acceptability perspective is arguably in line with the language found in Principle 1 of the *ASA Statement* where it observes that a *p*-value summarizes "the incompatibility between a particular set of data and a proposed model of the data…..The smaller the *p*-value, the greater the statistical incompatibility of the data with the null hypothesis."[16] However, as noted previously, Principle 6 of the *ASA Statement* also clearly indicates that a null outcome does not in any way rule out the consistency of other hypotheses, particularly the alternative research hypothesis, with the evidence. That is, in

[16] In fact, textbook presentations of hypothesis testing sometimes use the "consistent with the null" approach in describing null outcomes. Contextually, however, such presentations are also quite clear to avoid making anything conclusive out of this description.

general, a null outcome is consistent with any relevant hypothesis and, in particular, is most certainly not inconsistent with either the tested null hypothesis or any associated alternative research hypothesis. Collectively, from a descriptive perspective this ASA guidance argues against SC interpretations for the central findings of an analysis apart from a recognition that reasonable contrarian hypotheses are also consistent with the observed evidence.[17]

We divide those null outcome descriptions that do not fall into the first three conclusiveness categories between those we deem to be arguably conclusive and those that we deem to be non-conclusive. Most of those identified as arguably conclusive are so identified because of how they present the null outcome's absence of statistical significance. The notion of statistical significance and, in particular, the lack thereof, pose a particularly difficult challenge when presenting a null outcome. A representationally faithful discussion of a null outcome certainly needs to report the fact that it lacks statistical significance. Yet, at the same time, it should avoid conveying any sense of conclusiveness to this outcome because such an outcome does not say an effect is absent nor, absent additional descriptive analyses, does it even say much of anything about what magnitudes of effects are likely or unlikely. Given this perspective, stating that an effect is "statistically insignificant," while accurate, is also arguably advancing the case the effect is either non-existent or insubstantial, which are inferences that do not necessarily follow from a lack of statistical significance.[18] In contrast, descriptions such as "unable to reject, "not reliably different," and "no

[17] As a matter of convention, null hypotheses are commonly phrased as inclusive of zero or no difference. However, the underlying hypothesis testing framework does not require this. It is perfectly valid to propose and test a hypothesis that an effect is strictly less than zero, strictly greater than zero, or that it exceeds or falls below some non-zero threshold value, in much the same manner that one proposes and tests a null that an effect equals or is inclusive of zero. One wonders, however, how palatable describing a failure to reject a null hypothesis that there has been a 20% reduction in the likelihood of reporting a restatement (which would be true in both the Lennox (2016) and Kim and Klein (2017) analyses) in an SC fashion (e.g., "there is no evidence that the effect is less than 20%," or, "the evidence is consistent with a 20% reduction") would be to the vast majority of readers, reviewers, and editors?

[18] Lindsay (1994) addresses the presence of this "tendency to equate scientific significance with statistical significance" bias in the accounting literature. And, in fact, a very obvious example of this sort of fundamental

reliable evidence of" are more neutral. We classify these sorts of interpretations as being non-conclusive.

Tables 3 and 4 present summary analyses based on this five-level categorization system for the descriptive language employed in presenting null outcomes. Table 3 focuses on how article abstracts interpret null outcomes while table 4 focuses on how article text, apart from the abstract, interpret null outcomes. Detailed information on textual null outcome descriptions is available upon request from the corresponding author.

29 of the 35 articles in our study discuss some aspect of a null outcome in their abstracts. There are a total of 43 underlying null outcomes that are addressed in these abstracts. Table 3 lists the exact abstract phrasing employed in discussing each of these outcomes. Based on our five-way categorization system we deemed 27 of the 43 as described in a precisely conclusive fashion. Another 8 are deemed generally conclusive while the remaining 8 are selectively conclusive. In summary, the evidence here is overwhelming, in article abstracts the literature as a matter of course presents inherently inconclusive empirical evidence in misleadingly conclusive terms. Moreover, the substantial majority of the time the language employed is highly conclusive.

Table 4 presents summary data for a similar analysis of statements pertaining to null outcomes found in the texts of the 35 articles. Summary counts of PC, GC, SC, AC, and NC statements are provided by article. As was true for abstracts, precisely conclusive terms predominate. All but two of the articles employ such language at some point to describe reported null outcomes. However, the two exceptions, Henry and Leone (2016) and Lennox (2016), each employ precisely conclusive language in their abstracts (see table 3). Hence, none of the articles

misinterpretation of the concept of statistical insignificance is seen in the set of null outcomes in one of the articles we examine in this study. Specifically, Lin and Wang (2016) justify their assertion that "innovation premium is not related to takeover probability" explicitly because "the coefficient on *TakoverProbalility* X *Innovaton Efficiency* is statistically insignificant." (p. 965) steers entirely clear of falsely asserting at some point that an observed null outcome indicates that the tested null hypothesis is precisely true. Equally alarming, only two articles, Cannon and Bedard (2017) and Henry and Leone (2016), manage to provide a clearly inconclusive description of a null outcome. And, each of these articles achieves this rarified level of candor precisely once: Cannon and Bedard, when it observes for its hypotheses H5b and H6b that "model results do not reject the null for both constructs" (p. 99); and Henry and Leone when it observes that "tests of differences cannot reject the null hypothesis that the explanatory power of models incorporating the alternatives is equivalent." (p. 155)

VI. CONFIDENCE INTERVAL ANALYSES

The most glaring descriptive deficiency in the null outcome interpretations we document is the absence of confidence interval reporting. In this section we illustrate the relevance and efficacy of confidence interval based descriptive analysis by providing such analyses for five of the null outcomes we identified in our prior analyses. Our goal in doing this is twofold. First, as a follow-on to Dyckman and Zeff (2014) we illustrate that conducting such analyses is both feasible and substantive. Second, we use these analyses to underscore the point of just how much an accompanying confidence interval alters a null outcome's interpretation, at least relative to how the original article interprets it. Given these objectives, we further admit to the fact that the null outcomes we address here were chosen to some degree for simplicity, variety and effect. However, Appendix C provides similar analyses for many more of the null outcomes identified in our analysis. We could likely have used any five of them to reach the same substantive insights that emerge from those chosen for presentation and discussion here.

Lennox (2016)

Lennox (2016) tests the null hypothesis that "There is no change in audit quality after companies reduce their APTS purchases following the new rules" using three measures of audit quality. Null outcomes are obtained for all three measures, and in this evaluation we treat each of these measures as a separate outcome. In his paper's text discussion Lennox asserts that the estimated coefficients for the change effects are "very small" and that the evidence "suggests no significant change in misstatements, tax-related misstatements, or going-concern opinions." The main results are presented for the independent variable TREATxPOST in his table 6 for (1) a full sample; and, (2) a matched sample. In this re-examination we focus on the more powerful (per Lennox) full sample outcomes.

We generate two standard deviation confidence intervals for each of the full sample TREATxPOST estimates by first dividing the reported effect by its associated t-value as standard error estimates are not directly reported. This yields standard error estimates of .118, 1, and .175 for misstatements tax-related misstatements, and going concern opinions. As the respective estimated effects are -0.04, -0.01, and .29 the following two standard deviation confidence intervals are obtained:

(1) Misstatements: lower bound (LB) -0.276; upper bound (UB): +0.196

(2) Tax Misstatements: LB = -2.01; UB = +1.99

(3) Going Concern Opinion: LB = -0.06; UB = + 0.64 .

As the tax misstatements standard error derivation is problematic due to the non-reporting of coefficient estimate significant digits we limit our descriptive analyses to the confidence intervals for misstatements and going concerns opinions.

As logistic regressions are employed the coefficient and bound magnitudes are not directly interpretable. However, they can be converted into odds ratios which are substantively interpretable. In the case of misstatements converting the above UB and LB values into odds ratios indicates that in the post restriction period hypothesized effect values as high as an increase of 21.65% or as small as a decline of 24.12% are not inconsistent with the examined evidence. Hence, based on this additional analysis we can certainly infer that the rule did not lead to quite large declines, it is quite difficult to see how they would justify an inference that the evidence indicates that the decline in misstatements was, at most, small. And, these bounds are certainly inconsistent with a claim such as that found in the article's abstract that the evidence shows that no decline at all occurred in response to the implementation of APTS restrictions.

In the case of going concern opinions converting the above LB and UB values into likelihoods yields a rather small value of a 6% decline for the LB. The UB value, however, is +89.65%. That is, based on the underlying evidence examined in this study it is not possibly to reliably rule out hypotheses that the likelihood that an affected firm reported a going concern opinion (where higher going concern opinion rates signal an improvement in audit quality) nearly doubled after the implementation of the restrictions.[19] In this case it is difficult to see why an analysis would be comfortable saying anything at all about the plausibility of the null hypotheses being true or the range of likely effect values consistent with the data being small.

[19] As going concern opinions are rare events the range of extreme likelihoods here is arguably misleading. That is, if there is only a .5% chance of a going concern opinion to begin with then a doubling of this value only increases it to 1%, which does not seem all that big in the overall scheme of things. However, this line of thinking raises the more substantive question of why a study would choose to employ such an inherently low impact measure? That is, would we really expect the sort of change in rules being evaluated here to cause some sort of meteoritic rise in going concern opinions? Offhand, we think most would be surprised if this sort of rule changed moved the needle by as much as 25%. Consequently, this low economic magnitude effect perspective on the outcome is better viewed as a criticism of the study's choice to use such a low impact measure to begin with, not as grounds for debating what sort of outcome should be viewed as "not small." Hence, given that a measure is been deemed suitable for inferential analysis, it is difficult to conceive why a sizable percentage change in it can possibly be viewed as inconsequential.

Finally, and uniquely among the null outcome papers we examine, Lennox provides analyses targeting the concern that his tests may lack the power to support his no effect conclusion. His approach is particularly relevant to our analysis in that it relies on hypothesis tests of non-zero null effects. Specifically, he evaluates whether a hypothetical effect of a given magnitude would have been rejected. Essentially, this approach is a very limited form of power curve analysis. That is, an attempt to measure the type 2 error rate. The key component to such an analysis, of course, is the determination of the hypothetical (assuming the null is false) effect's magnitude. Larger effects are easier to detect than smaller ones so the type 2 error rate is necessarily dependent on the size of the effect that is being searched for. A power curve reports rejection likelihoods for a given p-level as the size of the unobserved effect varies.

Lennox, however, does not report power curves. Rather, the study focuses on a specific effect magnitude--the estimated pre-existing difference in effect between the affected and unaffected sample firms. In the case of misstatements this effect is -0.27, and we are informed that an underlying offsetting effect of this magnitude would be rejected at the .05 level. This is roughly in line with what our confidence interval analysis lower bound of -0.276 (i.e., a 24% reduction in likelihood) implies when one takes into consideration the fact that it is based on rounded numbers (the reported estimate and t-value) and a slightly higher significance level threshold.[20] That is, if the average decline in restatement likelihood had exceeded 24%, which is around what the underlying difference in affected and unaffected firms is prior to the rule change, then we would almost certainly have rejected the no change null. What is far less clear, however, is why we should

[20] Moreover, the -0.276 confidence interval determination is much more straightforward, and is not interpretatively tied to a specific, likely selectively chosen, counterfactual. That is, one clear takeaway from the Lennox analysis is that relative to confidence interval analysis, hypothesis testing is a comparatively awkward approach to conducting meaningful descriptive analyses of hull outcomes.

take reliably ruling out the presence of a 24% or more reduction in restatement likelihood as somehow advancing a conclusion that the underlying change is small, to say nothing about a claim that there is "no change in audit quality," in any sort of compelling fashion.

Kim and Klein (2017)

In addition to reporting on the net equity valuation impact of listing standard changes, which we examine in the initial portion of this paper, Kim and Klein (2017) also report an analysis addressing the impact of these changes on restatement and fraud likelihoods as well as earnings management levels. Based on this analysis they conclude that there is no evidence of a change in restatements, fraud related restatements, or earnings management in response to the rule changes. These inferences are largely based on tests of coefficient estimates for the independent variable PostxOOC in table 7 of their article.

Kim and Klein report standard errors for coefficient estimates allowing the direct determination of two standard deviation confidence intervals for the PostxOOC coefficients as follows:

Restatement:	LB = -1.05; UB = 0.366
Fraud Restatement:	LB = -1.236; UB = 0.860
Earnings Management:	LB = -0.014; UB = 0.014 .

The first two sets of estimates are from logistic regressions. Hence, we convert the relevant values to likelihood ratios, yielding LBs of -65.00% and -70.95%. That is, the evidence examined here is not a reliable basis for ruling out an alternative hypothesis that the rule change reduced restatement likelihoods by over 60% and fraudulent restatement likelihoods by over 70%. Hence, while it certainly true that there is no reliable evidence that restatement likelihoods decreased, neither is there any reliable evidence to dispute contentions that they declined dramatically.[21]

As there is no obvious absolute scale for assessing what constitutes a high versus a low level of earnings management (EM) activity, evaluating the EM bounds is somewhat more of a challenge. Relative analysis, however, is still feasible. In this regard, table 3 of Kim and Klein indicates that the EM variable's standard error is .071. Hence, the above lower bound value of -0.012 amounts to around .197 of a one standard error in EM variation change. While this bound does not strike us as readily thought of as essentially equivalent to 0, it does seem to be rather small. Regrettably, Kim and Klein do not engage the question of why it should be so thought of in their analyses.

Robinson, Stomberg, and Towery (2016)

Robinson et al. (2016) present four null outcomes. Here we focus on the first of these, which pertains to whether the relation between settlements and tax expense differs after the implementation of FIN 48 (Financial Accounting Standards Board 2006, ASC 740-10, *Accounting for Uncertainty in Income Taxes*). The key variable in this analysis is SETTLEIND*FIN48IND, reported in column 4 of their table 3. The estimated coefficient for this variable of 0.009 lacks significance at conventional levels, which the authors interpret as indicating that there is "no evidence that FIN 48 significantly changed the ability of income tax expense to predict future tax cash flows."

As Robinson et al. do not report standard errors we again derive an estimate by dividing the estimated coefficient (0.009) by the reported t-value (1.55), yielding an estimated standard

[21] Kim and Klein, as well as a number of other logistic based analyses, present ROC values. Such values evaluate the overall discriminatory validity of the model. They do not, however, speak to the discriminatory saliency of individual elements of the model, which is the core issue here, apart from the setting where the variable of interest is the sole explanatory variable (which is certainly not the case here).

error of .0058. Consequently, the pertinent two standard deviation confidence interval is -0.0026 to +0.0206. These magnitudes are not in themselves inherently meaningful. However, as Robinson et al. in fact make use of, the estimated value of the stand-alone SETTLEIND effect of -0.024 is a particularly relevant basis for judging magnitude here. Specifically, the estimated interaction effect that is of central interest here is argued (under the alternative hypothesis) to be an offset to this stand-alone effect. Hence, we can divide the upper bound by the absolute value of this coefficient estimate to determine a plausible upper bound on the percentage of the effect that is being offset. Doing this yields an upper bound of 85.8%. Hence, a hypothesis that FIN 48 reduced the SETTLEIND effect that existed prior to its implementation by as much as 85% is not inconsistent with the examined evidence here. So, while the evidence is inconsistent with a conjecture that FIN48 fully eliminated the pre-existing SETTLEIND effect (i.e., a 100% reduction), that seems to be about the limit of what can be said about it. There is certainly no basis here to rule out alternative hypotheses claiming that FIN 48 had a rather consequential impact on the relation between settlements and tax rates.

Lourenco (2016)

Lourenco examines how feedback interacts with other incentives in a field experiment setting, concluding that "feedback is independent of other incentives," which is a form of a null outcome. Table 3 of her article presents her main results wherein all of the interactions involving the feedback indicator variable (FEED) lack significance. For purposes of this evaluation we focus on the specific null outcome for the three-way interaction MONEY*FEED*EXP, where MONEY indicates whether a monetary incentive is provided and EXP indicates whether the given observation is in a treatment or non-treatment state (determined weekly over time). The estimated effect for this variable is -6.91 with an associated standard error of 7.13. Consequently, its two standard deviation confidence interval is -21.17 to 7.35. The dependent variable is sales performance measured by sales scaled by a baseline goal. Table 2 of the article indicates that this variable has a standard deviation of between 23 and 25. Dividing the upper and lower bounds of the confidence interval by these the midpoint of these two values, 24, yields a confidence interval measured as a percentage of a standard deviation of the dependent variable of -88.21% to +30.63%. While these magnitudes indicate that the evidence is not supportive of the presence of extraordinarily large conditional feedback effects, they hardly seem sufficiently small to argue that such effects are not materially present. .

Alternatively, one might evaluate the confidence interval here by using the statistically significant effect on the MONEY*EXP variable as a benchmark. That is, if FEED is fixed at 1 rather than 0 then the MONEY*EXP effect equals the sum of the MONEY*EXP and MONEY*FEED*EXP coefficients. Hence, an operative question is whether a lower bound value for MONEY*FEED*EXP can flip the sign of the MONEY*EXP variable? And, in fact, it does just this. The MONEY*FEED estimate equals 13.30, which is substantially smaller in terms of absolute magnitude than the MONEY*FEED*EXP lower bound of -21.17. Hence, the evidence examined here is not inconsistent with the possibility that FEED flips the sign of the MONEY*EXP effect.

Fredrickson and Zolotoy (2016)

Table 6 of Fredrickson and Zolotoy presents examinations of whether individual and institutional investors exhibit visibility driven queuing behavior in processing earnings announcements. They find statistically significant evidence of queuing in high individually held firms but obtain a null outcome for high institutionally held firms.[22] Based on this analysis they conclude that "competing earnings announcements do not distract institutional investors." One of the key reported effects in their analysis is a value of +0.61 for the UExQUEUE_ABOVE variable. This effect should be negative if queuing is taking place, so this outcome is directionally consistent with their "conclusion."

When we turn to confidence intervals, however, things get a good bit murkier. As Frederickson and Zolotoy do not provide standard errors we again resort to backing out an estimate based on the reported coefficient value and t-statistic. In this case we obtain an estimated standard error of 1.605. Hence, the two standard deviation confidence interval here has an LB of -2.60 and an UB of 3.82. An obvious benchmark for evaluating, in particular, the estimated LB is the reported UExQUEUE_ABOVE estimate for individual investor held firms. This value is -2.63. Hence, we cannot reliably rule out an alternative hypothesis that queueing effects among institutional held firms equal or exceed the best estimate of queueing effects among individual held firms.[23]

Summary

Collectively, there are three distinct takeaways from the preceding confidence interval analyses. First, the effect-associated confidence interval supports the interpretation that the underlying effect is, at most, "small" in only one instance, the Kim and Klein earnings quality assessment. Hence, there is little support here for taking null outcomes as per se reliable indicators

[22] This is a particularly well known form of the inferring the consequent fallacy in that a difference between two groups (a null rejection) is advanced based on based on separate analyses where one of the analyses results in a rejection of the null while the other results in a null outcome. See Gelman and Stern (2006) for a discussion.

[23] The reseach design framework employed by Fredrickson and Zolotoy is also of some relevance to the general theme of our study in that it is predicated on obtaining null outcomes in selected sub-groups. Interestingly, however, across tests the memberships of these sub-groups change such that firms that argued to have the effect in one test are included in the group that is argued to exhibit no effect at all in another test. That is, the only way the no effect null is possibly true is if it is true for all sub-groups examined, including those where it is expected to be (and, in fact is) rejected. Or, in other words, the design itself is inherently a self-contradiction. A self-contradiction that would not have occurred had the study simply avoided using a design built upon the dubious foundation of obtaining null outcome "results."

that the examined evidence shows that an underlying effect is substantively indistinguishable from the hypothesized null value or not materially consistent with the relevant alternative research hypothesis. Moreover, this inference is not unique to these five studies. Appendix C provides similar analyses for many of the other null outcome analyses identified in this paper. The picture there is no less severe. Null outcomes that, in our subjective assessment at least, pass under the "it is small" bar are rare.

Second, confidence interval analysis provides useful non-trivial insights about ranges of effect magnitudes that are consistent with the examined evidence. For instance, the Lennox (2106) analysis clearly indicates that the APTS rule implementation did not lead to a dramatic reduction in restatement likelihoods, since the relevant bound here is a roughly 20% reduction. In contrast, it is less clear whether such a less-than-dramatic decrease assertion fits the audit committee composition rule setting examined by Kim and Klein, since the possibility of a 60% reduction in restatement likelihood is not reliably ruled out by the evidence. Moreover, in both of these cases the confidence interval analyses make clear that there is more work to be done. Additional evidence needs to be collected, examined, and integrated that, in particular, will narrow these confidence interval ranges to a point where we can more tightly identify the range of likely effect magnitudes.

Third, on a stand-alone basis descriptive analyses tend to lack tension. A report that the evidence reliably indicates that a mandated change in audit committee composition decreased fraudulent restatement likelihoods by no more than 70.95% (our reinterpretation of Kim and Klein) clearly lacks the punch of a claim that the mandate "yields no benefits to investors." (Caption of the AAA press release http://aaahq.org/Outreach/Newsroom/Press-Releases/11-1-17- for the Kim and Klein study.) Hence, judging the merits of such descriptive evidence of non-definitive-outcomes may require some adjustment to the current literature's voracious appetite for tension in its publications.

VII. MULTIPLE NULL OUTCOMES

In a number of cases the null outcomes we examine are reported in multiples. That is, there is a guiding overarching hypothesis that is evaluated using alternative measures or sub-samples, yielding multiple null outcomes and, in a few instances, a mixture of null and statistically significant opposite direction outcomes.[24] At a very general level such collective null outcomes provide greater confidence in the possibility that any underlying effect(s) are either inconsequential or non-existent. However, formalizing the extent to which such collective outcomes heighten such confidence levels is highly situational. Just as is true for a stand-alone outcome, reliable interpretation of such multiple interrelated null outcomes demands a highly descriptive engagement with the evidence. For instance, suppose that three separate tests all yield null outcomes where there the associated upper bound is a 70% reduction in some undesirable behavior or outcome. These homogenous outcomes certainly instill confidence in the notion that this lower bound is likely a good bit smaller than 70%.[25] However, it would take a much larger number of such tests (or tests with tighter ranges) to get anywhere close to the point of being able to think that any such effect is, at most, of negligible magnitude.[26]

[24] In general, in multiple outcome settings where one or more of the tests result in opposite direction significance we only focus on the subset of null (statistically insignificant) outcomes. If the vast majority or all of the tests of a hypothesis result in opposite direction significance (see Lawrence, Siriviriyakul, and Sloan (2016) for an example), we viewed it as equivalent to a rejection of the null (i.e., we did not use it in our analysis).

[25] One particularly germane technique to evaluate such multi-outcome settings is meta-analysis (see Dyckman and Zeff for a discussion), which would also take into account any presumed underling non-independence across tests. Non-independence, in particular, makes it difficult to assess whether one can take each test as providing new information, which should affect beliefs, relative to simply re-reporting information already contained in the other tests, which should not affect beliefs.

[26] Moreover, there is no guarantee here that the null outcome will survive the multiple testing gauntlet as the heightened level of power that accompanies such testing means that smaller and smaller non-zero effects are reliably

Moreover, three confounding factors tend to limit the uncertainty reduction properties of such multiple null outcome occurrences. First, they are commonly inherently redundant. Hence, from an informational perspective the second, third, or fourth test is often bringing in very little new descriptively relevant understanding of the underlying effects relative to what underlies the first test. For instance, Towery (2017) reports null outcomes using change in federal cash tax payments and change in total tax payments as alternative dependent variables. While the total tax payments measure certainly potentially adds something informative relative to the federal cash tax payments measure, it is unclear how much new information one of these measures adds relative to the other.

Second, they also often focus on subsets of the overall data where the effect, if it truly exists, is likely to be strongest. While rejections within such targeted subsets can provide persuasive evidence in favor of rejecting the null, the converse is not true. As they necessarily use less data than what is employed in the overall examinations, from a null outcome perspective they are again redundant, and, moreover, they are also less powerful, ceteris paribus, than examinations based on the entire set of available observations.[27]

Finally, in many cases it is unclear whether one should view these sorts of examinations as multiple tests of the same null hypothesis, in which case some sort of aggregation or accumulation of test outcomes could be appropriate, or as separate tests of specific distinct questions that reflect possible ways the overarching hypothesized effect might or might not manifest itself. That is, should the tests of earnings management and restatement likelihood effects reported in Klein and

detected. In fact, the general absence of any substantive efforts to collapse multiple null outcomes into single overall test outcomes speaks quite directly to the inappropriateness of using multiple null outcomes to advance a no effect conclusion.

[27] Absent ceteris paribus, they are possibly more powerful when the differential effect between the subsample-included observations is analysis is large enough to offset the loss in power from the necessary decline in sample size.

Kim (2017) be taken as two separate tests of separate independent possible outcomes from the rule change? In which case they should be evaluated separately since neither is necessary for the other. Or, should they be taken as two tests evaluating a hypothesized common outcome of the rule change? In which case they can be aggregated, after taking into account the likely degree of jointness they exhibit.[28]

VIII. CONCLUSION

In the widely accepted structure for conducting conventional statistics-based hypothesis testing null outcomes are appropriately interpreted as a basis for being unable to reject the null. The structure also quite clearly prohibits their use as a basis for making conclusive assertions regarding the truth of the associated null hypothesis. These precepts are repeated (twice) in the *ASA Statement on Statistical Significance and P-Values.* A central takeaway from our analyses of how articles interpret the null outcomes they report is that the accounting literature has quite forgotten how to say "unable to reject," but excels at violating the "is true" proscription. There is, to our knowledge, no methodological justification for this seemingly opportunistic departure from accepted statistical practice.

Descriptive inferential perspectives, on the other hand, do not impose the rigid interpretive structure on data analysis such as that mandated by conventional null hypothesis testing. However, they also do not typically provide a basis for somehow linking a null outcome with highly definitive assertions about an associated null or research hypothesis of interest. Substantive descriptive analysis can, through clear identification of the set of evidence-consistent alternative hypotheses, facilitate judgements as to as to whether the set of evidence-consistent alternatives are not substantively distinguishable from the stated null hypothesis. We, however, find little evidence

[28] Note that if the underlying metrics are perfectly jointly determined (i.e., they always come together or none come at all) then they are also inherently perfectly redundant.

that articles in the accounting literature are pursuing such a descriptive path, particularly with respect to exercising judgement, when interpreting null outcomes. The typical analysis simply reports a large *p*-value or a small test statistic value along with a tabulated estimated effect, an effect that is rarely mentioned, to say nothing of discussed in the text in any meaningful way. It then directly proceeds to make some form of misleading claim of how such an outcome demonstrates the truth of the postulated null hypothesis.

Our analysis is also of some relevance to the ongoing debate about the degree to which the accounting academic literature is biased against publishing papers reporting null outcomes (Lindsay, 1994; Bamber, Christensen, and Gaver, 2000; Dyckman and Zeff, 2014). This "bias," however, is inherent to the conventional hypothesis testing structure.[29] And, without it the entire structure falls apart. The answer to this "bias," in our opinion, is not to try and bend the rules of conventional hypothesis testing to form some sort of "equitable" counterbalance. Rather, to echo a point more broadly advanced by Dyckman and Zeff (2104, 2015) and Dyckman (2016), the answer is to shift to a descriptive perspective of data inference. That is, to adopt the position that there is considerable merit in establishing descriptive understandings of fundamentally interesting problems, questions, and settings. Indeed, we would argue that what is currently nominally identified as "bias against the null" is, in reality, a manifestation of a very pervasive "bias against descriptive analysis." Descriptive is never going to match up with null hypothesis testing (when it returns null rejections) in terms of providing seemingly definitive yes/no answers to questions.

[29] Interestingly, the underlying structure of hypothesis testing is firmly rooted in a bias in favor of the null. That is, the null is only rejected if the evidence is compellingly inconsistent with it. And, it is for precisely this reason that inferring the truth of the null from a null outcome is unacceptable. The tradeoff one makes when adopting a conventional hypothesis testing approach to inference is giving up on drawing reliable conclusive affirmative inferences about the truth of the null in exchange for possibly being able to make very reliable inferences about it being false or inconsistent with the evidence. From this perspective, complaining about bias against the null is a bit like complaining about paying for a lottery ticket that, after the fact, didn't win the lottery.

And, in a publication environment where tension is critical, that is a rather severe handicap to operate under.

Finally, we advance the notion of reporting and examining confidence intervals as an initial step toward providing rigor to null outcome analysis. By focusing on the range of effects that are consistent with evidence, rather than a specific null hypothesis value (or range of null hypothesis values), confidence intervals discipline articles to describe any reported null outcomes in more representationally faithful fashions. It is, in particular, difficult to conceive of an article claiming a null hypothesis as "truth" at some point when the subsequent discussion of the evidence addresses a range of evidence-consistent values that include values directly contradicting such a claim. In fact, on a more general level, the literature would be far better served if authors, readers, and listeners when writing, or reading, or hearing statements such as "no relation", "no evidence of", "no change", "no difference", "insignificant", "inconsistent", "similar," etc., would make it a a practice to ask the question—"Is there a confidence interval for that?"

.

References

Aberson, C. (2002). Interpreting null results: improving presentation and conclusions with confidence intervals. *Journal of Articles in Support of the Null Hypothesis,* 1 (3): 36-42.

American Accounting Association (2017). Longstanding mandate on corporate audit committees yields no benefit for investors, new research finds. AAA press release. (http://aaahq.org/Outreach/Newsroom/Press-Releases/11-1-17-)

Bamber, L. S., Christensen, T. E., and Gaver, K. M. (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 (2): 103–129.

Bills, K.L., Lisic, L.L., and Seidel, T.A. (2017). Do CEO succession and succession planning affect stakeholders' perceptions of financial reporting risk? Evidence from audit fees. *The Accounting Review,* 92(4), 27-52.

Brasel, K., Doxey, M., Grenier, J., and Redffett, A. (2016). Risk disclosure preceding negative outcomes: the effects of reporting critical audit matters on judgments of auditor liability. *The Accounting Review,* 91(5): 1345-1362.

Brazel, J. F., Jackson, S. B., Schaefer, T. J., and Stewart, B. W. (2016). The outcome effect and professional skepticism. *The Accounting Review*, 91(6), 1577-1599.

Cannon, N. H., and Bedard, J. C. (2017). Auditing challenging fair value measurements: Evidence from the field. *The Accounting Review*, 92(4), 81-114.

Casas-Arce, Martinez-Jerez, F.A., Narayanan, V. (2017). The impact of forward-looking metrics on employee decision-making: the case of consumer lifetime value. *The Accounting Review,* 92(3): 31-56.

Chen, K. C., Cheng, Q., Lin, Y. C., Lin, Y. C., & Xiao, X. (2016). Financial reporting quality of Chinese reverse merger firms: The reverse merger effect or the weak country effect?. *The Accounting Review*, 91(5), 1363-1390.

Choi, J., Newman, A. H., and Tafkov, I. D. (2016). A marathon, a series of sprints, or both? Tournament horizon and dynamic task complexity in multi-period settings. *The Accounting Review*, 91(5), 1391-1410.

Damer, T. Edward. (2013) *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments 7th edition*. Wadsworth, Cenage Learning.

DeFond, M., Lim, C., and Zang, Y., (2016). Client conservatism and auditor-client contracting. *The Accounting Review*, 91(1): 69-98.

Drake, K., Goldman, N., and Lusch, S. (2016) Do income tax-related deficiencies in publicly disclosed PCAOB Part II reports influence audit client reporting of income tax accounts?. *The Accounting Review,* 91(5): 1411-1439.

Dutta, S., and Patatoukas, P.N. (2017) Identifying conditional conservatism in financial accounting data: theory and evidence. *The Accounting Review,* 92(4): 191-216.

Dyckman, T. R. (2016). Significance Testing: We Can Do Better. *Abacus*, 52(2), 319-342.

Dyckman, T. R., and Zeff, S. A. (2014). Some methodological deficiencies in empirical research articles in accounting. *Accounting Horizons*, 28(3), 695-712.

Dyckman, T. R., and Zeff, S. A. (2015). Accounting research: past, present, and future. *Abacus*, 51(4), 511-524.

Erickson, D., Hewitt, M., and Maines, L. (2017). Do investors perceive low risk when earnings are smooth relative to the volatility of operating cash flows? Discerning opportunity and incentive to report smooth earnings. *The Accounting Review* 92 (3): 137-154.

Farrell, A. M., Grenier, J. H., and Leiby, J. (2017). Scoundrels or stars? Theory and evidence on the quality of workers in online labor markets. *The Accounting Review*, 92(1), 93-114.

Francis, B. B., Hunter, D. M., Robinson, D. M., Robinson, M. N., and Yuan, X. (2017). Auditor Changes and the Cost of Bank Debt. *The Accounting Review*, 92(3), 155-184.

Frederickson, J. R., and Zolotoy, L. (2016). Competing Earnings Announcements: Which Announcement Do Investors Process First?. *The Accounting Review*, 91(2), 441-462.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60(4), 328-331.

Gong, Q., Li, O. Z., Lin, Y., and Wu, L. (2016). On the benefits of audit market consolidation: Evidence from merged audit firms. *The Accounting Review*, 91(2), 463-488.

Guenther, D. A., Matsunaga, S. R., and Williams, B. M. (2017). Is tax avoidance related to firm risk?. *The Accounting Review*, 92(1), 115-136.

Hall, C. M. (2016). Does ownership structure affect labor decisions? *The Accounting Review*, 91(6), 1671-1696.

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

Henry, H., and Leone, A., (2016). Measuring qualitative information in capital markets research: comparison of alternative methodologies to measure disclosure tone. *The Accounting Review,* 91(1): 153-178.

Humphreys, K., Gary, M., and Trotman, K. (2016). Dynamic decision making using the balanced scorecard framework. *The Accounting Review,* 91(5): 1441-1465.

Kelly, K., Presslee, A., and Webb, R. A. (2017). The Effects of Tangible Rewards versus Cash Rewards in Consecutive Sales Tournaments: A Field Experiment. *The Accounting Review*, 92(6), 165-185.

Khan, M., Serafeim, G., and Yoon, A. (2016). Corporate sustainability: first evidence on materiality. *The Accounting Review*, 91(6): 1697-1724.

Kim, J., and P. Ji. (2015). Significance testing in empirical finance: A critical review and assessment. *Journal of Empirical Finance,* 34: 1-14.

Kim, J. H., Ji, P., and Ahmed, K. (2017). Significance Testing in Accounting Research: A Critical Evaluation Based on Evidence.

Kim, S., and Klein, A. (2017). Did the 1999 NYSE and NASDAQ Listing Standard Changes on Audit Committee Composition Benefit Investors?. *The Accounting Review*, 92(6), 187-212.

Krishnan, J., Krishnan, J., and Song, H. (2017). PCAOB international inspections and audit quality. *The Accounting Review*, 92(5), 143-166.

Laurion, H., Lawrence, A., and Ryans, J. P. (2017). US audit partner rotations. *The Accounting Review*, 92(3), 209-237.

Lawrence, Siriviriyakul and Sloan, (2016) Who's the fairest of them all? Evidence from closed-end funds, *The Accounting Review,* 91(1), 207-227.

Lennox, C. S. (2016). Did the PCAOB's restrictions on auditors' tax services improve audit quality? *The Accounting Review*, 91(5), 1493-1512.

Li, L., Qi, B., Tian, G., and Zhang, G. (2017). The contagion effect of low-quality audits at the level of individual auditors. *The Accounting Review*, 92(1), 137-163.

Lin, J., and Wang, Y., (2016) The R&D premium and takeover risk. *The Accounting Review,* 91(3): 955-971.

Lindsay, R. M. (1994). Publication system biases associated with the statistical testing paradigm. *Contemporary Accounting Research*, 11(1), 33-57.

Lourenço, S. M. (2016). Monetary incentives, feedback, and recognition—complements or substitutes? Evidence from a field experiment in a retail services company. *The Accounting Review*, 91(1), 279-297.

Nelson, M. W., Proell, C. A., and Randel, A. E. (2016). Team-oriented leadership and auditors' willingness to raise audit issues. *The Accounting Review*, 91(6), 1781-1805.

Nessa, M. (2017) Repatriation tax costs and U.S. multinational companies' shareholder payouts. *The Accounting Review,* 92(4), 191-216.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, 5(2), 241-301.

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

Patatoukas, P. N., and Thomas, J. K. (2016). Placebo tests of conditional conservatism. *The Accounting Review*, 91(2), 625-648.

Publication Manual of the American Psychological Association 6th Edition (2013), American Psychological Association. 2013.

Robinson, L. A., Stomberg, B., and Towery, E. M. (2016). One size does not fit all: How the uniform rules of FIN 48 affect the relevance of income tax accounting. *The Accounting Review*, 91(4), 1195-1217.

Schroeder, J. H., and Shepardson, M. L. (2016). Do SOX 404 control audits and management assessments improve overall internal control system quality?. *The Accounting Review*, 91(5), 1513-1541.

Stone, D. (2018). The new "statistics" and nullifying the null: Twelve actions for improving quantitative accounting research quality and integrity. *Accounting Horizons,* 32(1), 105-120.

Towery, E. M. (2017). Unintended consequences of linking tax return disclosures to financial reporting for income taxes: Evidence from Schedule UTP. *The Accounting Review*, 92(5), 201-226.

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

Wieczynska, M. (2016). The "Big" Consequences of IFRS: How and When Does the Adoption of IFRS Benefit Global Accounting Firms?. *The Accounting Review*, 91(4), 1257-1283.

Table 1

Null Outcome Articles in *The Accounting Review*: 2016-2017

Article	Area	Design	Number of Null Outcomes		Number Stated as Null Hyp.
			In Paper	In Abstract	
Bills et al. (2017)	Audit	Archival	2	2	0
Brasel et al. (2016)	Audit	Exper.	2	2	1
Brazel et al. (2016)	Audit	Exper.	1	1	0
Cannon & Bedard (2017)	Audit	Exper.	7	0	2
Casas-Arce et al. (2017)	Man.	Archival	1	1	1
Chen et al. (2016)	Financial	Archival	1	1	0
Choi et al. (2016)	Managerial	Exper.	2	1	0
DeFond et al. (2016)	Audit	Archival	1	0	0
Drake et al. (2016)	Tax	Archival	1	1	0
Dutta & Patatoukas (2017)	Financial	Archival	1	1	1
Erickson et al. (2017)	Financial	Exper.	1	0	1
Farrell et al. (2017)	Method	Exper.	3	2	1
Francis et al. (2017)	Audit	Archival	1	1	0
Frederickson & Zolotoy (2016)	Financial	Archival	3	3	1
Gong et al. (2016)	Audit	Archival	1	1	1
Guenther et al. (2017)	Tax	Archival	2	2	0
Hall (2016)	Financial	Archival	1	0	1
Henry & Leone (2016)	Financial	Archival	1	1	1
Humphreys et al. (2016)	Managerial	Exper.	2	2	0
Kelly et al. (2017)	Managerial	Exper.	1	1	0
Khan et al. (2016)	Financial	Archival	1	1	0
Kim and Klein (2017)	Audit	Archival	5	2	0
Krishnan et al. (2017)	Audit	Archival	1	1	1
Laurion et al. (2017)	Audit	Archival	1	1	0
Lennox (2016)[30]	Audit	Archival	3	3	3
Li et al. (2017)	Audit	Archival	1	1	1
Lin and Wang (2016)	Financial	Archival	2	2	0
Lourenco (2016)	Managerial	Exper.	2	2	2
Nelson et al. (2016)	Audit	Exper.	2	1	0
Nessa (2017)	Tax	Archival	1	1	0
Patatoukas and Thomas (2016)	Financial	Archival	1	0	1
Robinson et al. (2016)	Financial	Archival	3	2	1
Schroeder & Shepardson (2016)	Audit	Archival	1	0	0
Towery (2017)	Tax	Archival	1	1	1
Wieczynska (2016)	Financial	Archival	3	2	0
Totals			63	43	21
% Articles Non-Zero				82.9%	48.6%

[30] Lennox tests a single broadly stated hypothesis using three measures. As his discussion and analysis varies by measure we treat these as three separate hypotheses for purposes of our analyses.

Table 2
Descriptive Statistic Presentations for Null Outcomes

Article	# of Null Out-comes	Reports CI or Range Analysis	Reports Std. Error	Text Presentation		
				Value of Estimated Effect	Other Descr.	Test stat. Or p-value
Bills et al.	2	0	0	0	0	1
Brasel et al.	2	0	0	2	0	2
Brazel et al.	1	0	0	1	0	1
Cannon & Bedard	7	0	0	0	0	0
Casas-Arce et al.	1	0	1	0	1	0
Chen et al.	1	0	0	0	0	0
Choi et al.	2	0	2	2	0	2
DeFond et al.	1	0	0	0	0	0
Drake et al.	1	0	0	0	1	0
Dutta & Patatoukas	1	0	0	0	0	0
Erickson et al.	1	0	0	0	0	1
Farrell et al.	3	0	0	3	0	3
Francis et al.	1	0	0	0	0	0
Frederickson&Zolotoy	3	0	0	0	0	2
Gong et al.	1	0	0	0	0	0
Guenther et al.	2	0	0	0	0	0
Hall	1	0	0	1	1	0
Henry & Leone	1	0	0	1	1	0
Humphreys et al.	2	0	0[31]	1	0	2
Kelly et al.	1	0	0	0	0	1
Khan et al.	1	0	0	1	0	0
Kim and Klein	5	0	5	1	0	1
Krishnan et al.	1	0	0	0	0	0
Laurion et al.	1	0	0	0	1	0
Lennox	3	0[32]	0	1	3	3
Li et al.	1	0	0	0	0	0
Lin & Wang	2	0	0	0	1	0
Lourenco	2	0	2	0	0	0
Nelson et al.	2	0	0	1	0	2
Nessa	1	0	0	0	0	0
Patatoukas & Thomas	1	0	0	0	1	0
Robinson et al.	3	0	0	1	1	3
Schroeder&Shepardson	1	0	0	0	0	0
Towery	1	0	0	0	0	0
Wieczynska	3	0	3	0	0	0
Totals	63	0	11	16	12	24

[31] Group means and standard deviations are tabulated for one of the null outcomes. Standard errors of differences in means are not reported.

[32] Tests of alternative non-zero benchmarks are conducted, establishing that any underlying effect that may be present is smaller than these (rather large) benchmarks.

	# of Null Outcomes	Reports CI or Range Analysis	Reports Std. Error	Text Presentation		
				Value of Estimated Effect	Other Descr.	Test stat. Or p-value
% Articles Non-Zero		0%	14.3%	34.3%	25.7%	37.1%
% Outcomes Non-Zero		0%	17.5%	25.4%	19.0%	38.1%

This table reports what specific information items are and **are not** provided by articles for the null outcomes they report. In cases where a paper reports multiple null outcomes counts are provided where the maximum value is the paper's number of null outcomes (as listed in the second column of the table). Text presentation columns refer to the item being reported in the text of the paper, not to tabulated presentations.

Table 3

Text Discussion of Null Outcomes in Article Abstracts

Article	Null Outcome Statement	Type
Bills et al.	1. “as evidenced by *a lack of* an audit pricing adjustment” 2. “we *do not find evidence* of a deterioration in audit quality”	PC SC
Brasel et al.	1. “we find that CAM disclosures *only reduce* auditor liability for undetected misstatements that, absent CAM disclosure, are relatively difficult to foresee” 2. “CAM disclosures that are unrelated to subsequent misstatements *neither increase nor reduce* auditor liability judgments relative to the current regime.”	PC PC
Brazel et al.	“consultation *did not effectively mitigate* the outcome effect”	PC
Casas-Arce et al.	“the use of CLV *did not negatively impact* pricing” (note, this is linked with a similar assertion regarding default risk that is not subjected to statistical testing.)	PC
Chen et al.	“the financial reporting quality of U.S. RM firms *is similar*”	GC
Choi et al.	“with *similar performance* in the latter two tournaments”	GC
Drake et al.	“Deloitte’s clients report valuation allowances and UTB balances that *are not significantly different* than other annually inspected auditors”	GC
Dutta & Patatoukas	A series of placebo tests provides additional support for the construct validity	GC
Farrell et al.	1. “online workers are *at least as willing* as students” 2. “performance-based wages, which are *just as effective* in inducing high effort as high fixed wages”	PC PC
Francis et al.	“we *find no effect* resulting from the forced auditor changes”	PC
Frederickson & Zolotoy	1. “We *find no support* for queuing based on the latter” 2. “Earnings announcements made by firms that are more visible…—*but not by* firms that are less visible—mitigate” 3. “individual investors—*not* institutional investors—drive the queuing effect.”	SC PC PC
Gong et al.	“*unaccompanied by* a deterioration in audit quality”	PC
Guenther et al.	“measures of tax avoidance …are generally *not associated with”* 1. “future tax rate volatility” or 2. “future overall firm risk”	 PC PC
Henry & Leone	1. “word-frequency tone measures *are as powerful as* the Naıve Bayesian machine-learning tone measure from Li (2010)”	PC
Humphreys et al.	1. “For managers presented with causal linkages with delays, long-term profit generation is higher than the control group, *but is not significantly different* from the causal linkages without delays treatment” 2. “Learning *is found to plateau* for the causal linkages without delays treatment and is not present for the control group.”	GC PC
Lin & Wang	1. “*but not to* innovation efficiency” 2. “*but not the* innovation efficiency premium”	PC PC
Kelly et al.	“We *do not find significant* effects of reward type”	GC

Article	Null Outcome Statement	Type
Khan et al.	“firms with good ratings on immaterial sustainability issues *do not significantly outperform* firms with poor ratings on the same issues”	GC
Kim and Klein	“we find *no evidence* of	
	1. “higher market value or”	SC
	2. “better financial reporting quality”	SC
Krishnan et al.	“we *find no systematic differences* for accruals or for value relevance”	PC
Laurion et al.	“we *find no evidence of a change* in the frequency”	SC
Lennox	“I find *no change* in audit quality”	
	(for 1. accounting misstatements;	PC
	2. tax-related misstatements;	PC
	3. going concern opinion likelihoods.	PC
Li et al.	“we *find little evidence* that an audit failure also casts doubt”	GC
Lourenco	1. “*feedback is independent* of the other incentives”	PC
	2. “feedback in the form of knowledge of results *has no impact*”	PC
Nelson et al.	1.“but *not by* concerns about the … repercussions”	PC
Nessa	“I *do not find evidence* that repatriation tax costs decrease U.S. MNCs’ share repurchases”	SC
Robinson et al.	1. “*we find no evidence* that FIN 48 increased …”	SC
	2. “*we find no evidence* that investors identify ….”	SC
Towery	“firms ……*do not claim fewer* income tax benefits…”	PC
Wieczynska	1. “adoption is *not associated with an increase* …” (before adoption)	PC
	2. “adoption is *not associated with an increase* …” (after adoption)	PC
Articles with PC descriptions of null outcomes in abstract		18
Percentage of abstract identified null outcomes presented as PC		62.8%

This table reports descriptions of null outcomes identified in 29 article abstracts. Each description is classified into one of the following five types based on its conclusive nature: Precisely Conclusive (PC), Generally Conclusive (GC), Selectively Conclusive (SC), Arguably Conclusive (AC), and Non-Conclusive (NC). See appendix A for further details on each of these categories.

Table 4

Textual Analysis of Article Discussions of Null Outcomes

Article	Null Hypothesis Description Counts by Conclusive Nature				
	PC	GC	SC	AC	NC
Bills et al.	5	1	4	1	0
Brasel et al.	3	0	0	2	0
Brazel et al.	4	0	0	3	0
Cannon & Bedard	8	3	1	0	1
Casas-Arce et al.	2	1	1	0	0
Chen et al.	4	3	2	0	0
Choi et al.	4	4	0	0	0
DeFond et al.	3	1	0	0	0
Drake et al.	2	3	0	0	0
Dutta & Patatoukas	1	0	3	0	0
Erickson et al.	3	0	0	0	0
Farrell et al.	3	4	5	4	0
Francis et al.	3	2	0	0	0
Frederickson & Zolotoy	5	0	6	3	0
Gong et al.	3	1	0	0	0
Guenther et al.	3	9	4	0	0
Hall	1	0	2	0	0
Henry & Leone	0	4	0	0	1
Humphreys et al.	4	1	2	0	0
Kelly et al.	5	1	1	0	0
Khan et al.	3	1	0	0	0
Kim and Klein	6	4	10	2	0
Krishnan et al.	3	1	3	1	0
Laurion et al.	1	3	0	0	0
Lennox	0	5	3	0	0
Li et al.	1	4	1	0	0
Lin & Wang	7	1	0	0	0
Lourenco	5	0	4	3	0
Nelson et al.	5	1	1	0	0
Nessa	4	1	3	0	0
Patatoukas and Thomas	3	1	0	0	0
Robinson et al.	3	5	5	0	0
Schroeder & Shepardson	3	1	2	1	0
Towery	1	1	2	0	0
Wieczynska	4	0	1	0	0
Totals	115	67	66	20	2
Articles	33	27	23	9	2

This table summarizes how article texts (excluding the abstract) describe null outcomes. Counts are provided by article for five distinct descriptive types: Precisely Conclusive (PC), Generally Conclusive (GC), Selectively Conclusive (SC), Arguably Conclusive (AC), and Non-Conclusive (NC). See appendix A for further details on each of these categories.

Appendix A

Classification of Null Outcome Text Discussions

Classification	Definitions and Examples
PC: Precisely Conclusive	Definitive statements that the null is precisely true or the alternative is unconditionally false. Examples: did not, is no difference, find no effect, equals…, unaccompanied by, (alternative) is rejected; not different from, independent, no association, inconsistent with (alternative), etc.
GC: Generally Conclusive	Statements indicating that any effect is negligible or inconsequential. Examples: insignificant (w/o any statistical reference); small, little, similar, etc.
SC: Selectively Conclusive	Statements that selectively point out that outcome is: (1) consistent with null or, (2) unsupportive of alternative. Examples: Consistent with null; find no evidence for alternative, find no support for, etc.
AC: Arguably Conclusive	Statements that can be taken as conclusive, although it is not clear that they are or are intended to be. Example: Statistically Insignificant.
NC: Non-conclusive	Clearly inconclusive statements. Examples: Unable to reject; lacks statistical significance, not reliably different, unclear, etc.

Appendix B
Listing of Hypotheses, Predictions. and Questions Associated with Null Outcomes

Article	#	Hypothesis/Question
Bills et al.	2	1. "H3: Audit fees will increase to a lesser extent for companies with a new CEO who is considered an heir apparent before taking office than for companies with a new CEO who is an insider, but not considered an heir apparent before taking office." (p. 30) (this exact hypothesis is rejected, but the test gives rise to the null outcome of an insignificant effect relative to fees when there is no change in CEO that underlies the abstract assertion of a "lack of an audit pricing adjustment". 2. "We next examine whether uncertainty due to CEO succession is associated with audit quality." (p. 40)
Brasel et al.	2	1. "but (do) not (observe a significant decrease) within restoration liability." (p.1351/2) (partial null outcome for H1.) 2. "How do jurors' auditor liability judgments compare when the audit report discloses a CAM that is unrelated to the undetected misstatement versus when the audit report is silent regarding CAMs?" (p. 1349)
Brazel et al.	1	H2:" When subordinate auditors consult with their superiors during the course of exercising skepticism, the outcome effect in auditor evaluations is reduced." (p. 1582)
Cannon & Bedard	7	1. H2: "Auditors will be more likely to use a valuation specialist to assist the engagement team as estimation uncertainty for the FVM increases." (p.86) 2. H5b: "The likelihood of booking an audit adjustment that decreases income will not differ based on the estimation uncertainty for the FVM." (p. 87) 3. H6b: The likelihood of booking an audit adjustment that decreases income will not differ based on the level of inherent and control risk assessments for the FVM. (p. 87) 4. H7a: The likelihood of auditors discussing a possible audit adjustment with client management will increase when a valuation specialist is used by the engagement team. 5. H7b: The likelihood of booking an audit adjustment that decreases income will increase when a valuation specialist is used by the engagement team (p. 88) 6. H8a: The likelihood of auditors discussing a possible audit adjustment with client management will increase when an independent estimate of the FVM is developed. (p. 88) 7. H8b: The likelihood of booking an audit adjustment that decreases income will increase when an independent estimate of the FVM is developed. (p. 88)
Casas Arce et al.	1	"Our model predictions with respect to price (or, equivalently, to credit risk) are ambiguous." (p. 37).
Chen et al.	1	H1: Ceteris paribus, the financial reporting quality of U.S. RM firms is lower than that of U.S. IPO firms/We find that the financial reporting quality of U.S. RM firms is similar to that of matched U.S. IPO firms (p. 1368)
Choi et al.	2	1. H3b: When dynamic task complexity is high, strategy experimentation is greater in a hybrid tournament than in a grand tournament. (p. 1396) 2. RQ1b: When dynamic task complexity is high, does performance in a hybrid tournament differ from performance in a grand tournament? (p. 1397) (Grand Similar to Hybrid) in abstract.

DeFond et al.	2	“we find that auditors do not strategically respond to unconditional conservatism by adjusting their fees, GCO frequency, or propensity to resign.” (p. 71) “We also find that unconditional conservatism is not associated with lawsuits against auditors or client restatements” (p. 71)
Drake et al.	1	“we conduct our tests of valuation allowances and UTBs in subsequent years and note that Deloitte clients continue to report similar levels of valuation allowance as clients of other annually inspected auditors” (p. 1412)
Dutta & Patatoukas	1	“Next, we introduce construct validity tests using placebo test variables that should be free of the effect of conditional conservatism.” (p. 208)
Erickson et al.	1	H1: (partial) “will perceive relatively high risk only when both operating cash flows and earnings are volatile” (p. 141)
Farrell et al.	3	1. H1a: Workers in online labor markets report their private information less honestly than do students. (p. 97) 2. H1b: Workers in online labor markets exert less effort than do students. (p. 97) 3. H2b: When tasks are more intrinsically interesting, the efforts of workers in online labor markets will not differ between performance-based and flat wages. (p. 98) (partial support, partial no support)
Francis et al.	1	Finally, we find no effect resulting from the forced auditor changes due to Arthur Andersen. (abstract) This is an untabulated analysis highlighted throughout the paper.
Frederickson & Zolotoy	3	1. H1: Find no support for queuing based on the latter (Abstract) 2. H1: The number of announcing firms queued above firm i will be associated positively with the degree of market distraction, whereas the number of announcing firms queued below firm i will not distract the market (p. 443) 3. (The queuing effect will be less pronounced for institutional investors than for individual investors, p. 444) Analyzed as additional analysis, states that individual investors—not institutional investors—drive the queuing effect. (Abstract)
Gong et al.	1	we need to show that a reduction in audit hours due to audit firm mergers is not accompanied by any deterioration in audit quality; unaccompanied by a deterioration in audit quality, (p. 474)
Guenther et al.[33]	2	1. H1b: Low effective tax rates are positively associated with future tax rate volatility. (p. 119) (2 signif. in opposite direction outcomes, 5 null outcomes) 2. H2: Lower effective tax rates are positively associated with future stock price volatility. (p. 120) (3 sign. In opposite direction, 6 null outcomes).
Hall	1	I find no evidence that reducing labor costs in response to financial reporting and regulatory pressure affects future performance. (p. 1672)
Henry & Leone	2	1. “Our tests of alternative weighting methods for word-frequency tone measures compare the equal weighting method based solely on word frequencies (*wf*) and the inverse document frequency (*idf*) weighting method advocated in Loughran and McDonald (2011)” (p.155) 2. “we next compare word-count tone measures with the machine-

[33] Guenther et al. examine three main hypotheses using multiple measures of tax avoidance. For the first hypothesis significant opposite (of directional null) effects are widespread. This outcome is excluded from the analysis. For the remaining two hypotheses significant opposite effects are found for a few of the measures, while the remaining measures yield null outcomes. We treat these latter two hypotheses as encompassing null outcomes.

		learning measure used in Li (2010)" (p. 155)
Humphreys et al.	2.	1. "H2b: Managers presented with a set of strategic objectives with causal linkages and delays will generate higher performance on a dynamic task than those presented with the same objectives without delays." (p. 1446) 2. "A general linear model (GLM) repeated measures within-subjects analysis of learning rates is also conducted." (p 1454)
Kelly et al.	1	H1: Total sales performance for both tournaments will be higher for retailers eligible for tangible rewards than retailers eligible for cash rewards (p. 170)
Khan et al.	1	"firms with high residual changes on immaterial sustainability topics do not outperform firms with low residual changes on the same topic" (p. 1698)
Kim and Klein	5	1. Overall: "We first test for significant differences in stock returns between firms in and out of compliance in 1998. (p. 194) 2. Benefits: "If non-compliant firms with relatively poor financial reporting quality benefit most from the 1999 rules, then the coefficients b3 in Equation (3a) and b4 and b5 in Equation (3b) will be significantly positive for firms with restatements or with higher earnings management. (p. 195) 3. Costs. "We include these three variables as our cost variables in Equations (3a) and (3b), and predict negative coefficients on b5 in Equation (3a) and b7 and b8 in Equation (3b)." (p.196) 4. "We find no evidence that out-of-compliance firms with higher earnings management (financial reporting quality) or restatements (audit quality) prior to the proposed changes earned higher returns than out-of-compliance firms with better financial reporting quality." (p. 188) 5. "We measure whether desired changes (less earnings management, fewer restatements, less fraud) are seen after the implementation of the 1999 rules" (p.204)
Krishnan et al.	1	RQ3: "For clients cross-listed in the U.S., does the inspection effect on audit quality differ for inspection reports with and without audit deficiencies?" (p. 149)
Laurion et al.	1	H1: "Audit partner rotation is associated with ……a decrease in misstatements." (p. 214)
Lennox	3	"There is no change in audit quality after companies reduce their APTS purchases following the new rules." (p. 1497) In our own language: 1. No change in accounting misstatements 2. No change in tax-related misstatements 3. No change in going concern opinion likelihoods
Li et al.	1	We compare the audit quality of non-failed auditors who are in the same office as a failed auditor and that of auditors in offices that do not experience audit failures. (p. 138)
Lin & Wang	2.	1. "We find that a firm's innovation efficiency… is not related to its likelihood of becoming a takeover target" (p. 957) 2. "We expect and find that takeover risk is not responsible for the abnormal return associated with innovation efficiency" (p. 957)
Lourenco	2	1. H1: There is no interaction between monetary incentives and performance feedback in terms of their impact on performance; that is, monetary incentives and feedback are independent. (p. 283)

		2. H2: There is no interaction between recognition and performance feedback in terms of their impact on performance; that is, recognition and feedback are independent. (p. 284)
Nelson et al.	2	1. H3: Alignment between issue and supervisor concerns has less of an effect on an auditor's willingness to speak up about an issue when the auditor's supervisor is more team-oriented. (p. 1786) 2. Experiment 4: Analyses examine the extent to which the effect of team-oriented leadership on assessed willingness to speak up is mediated by three distinct constructs suggested by prior management research: team members' (1) team identification, (2) leader commitment, and (3) concern over consequences associated with speaking up. (p. 1782)
Nessa	1	H2: Repatriation tax costs are negatively associated with the level of share repurchases by U.S. MNCs. (p. 221)
Patatoukas & Thomas	1	Figure 1 "Predict upward bias, which should explain PT's lagged earnings bias." (p. 627)
Robinson et al.	3	Overall for 1. And 2.: The ability of income tax expense to predict future tax cash flows does not change as a result of FIN 48. (p. 1199) 1. observing a change in how settlements affect tax expense from pre- to post-FIN 48 provides evidence that FIN 48 changed the way income tax expense maps into future cash tax outflows (p. 1206) 2. We estimate this series of equations using our full sample of firms (and subsamples of firms most likely affected by FIN 48). Observing significant changes in the predictive ability of tax expense for future tax cash flows in the FIN 48 regime for these subsamples provides evidence consistent with differences across time being attributable to FIN 48 rather than other factors (p. 1208) (some sub-samples are opposite direction significant) 3.If investors correctly determine when excess reserves are incorporated into firms' tax expense accruals, then the level of tax expense should be less negatively related to levels of expected future cash outflows. Therefore, we would expect a positive coefficient on TaxExpense SubSample. On the other hand, if investors do not distinguish among these two types of firms, then the coefficient on TaxExpense SubSample should be no different from zero. (p. 1212)
Schroeder &Shepardson	1	H2: Management assessments of internal controls over financial reporting are associated with internal control system quality improvements. (p. 1518)
Towery	1	H1: Claims for uncertain tax positions do not change in response to Schedule UTP (p. 205)
Wieczynska	3	1. (H3b): The likelihood of switching from small audit firms to global ones increases in the year of IFRS adoption in countries with … (weak) regulatory regimes. (p. 1262) 2. H4a …: The likelihood of switching from small audit firms to global ones increases one year …. before IFRS adoption in countries with strong ….. regulatory regimes. (p. 1262) 3. (H4b): The likelihood of switching from small audit firms to global ones increases …. (two years) before IFRS adoption in countries with ….. (weak) regulatory regimes. (p. 1262)

Appendix C

Supplemental Confidence Interval Analyses of Null Outcomes

Article	Confidence Interval Analysis
Bills et al.	Table 4 of Bills et al. reports the analysis of the relation between heir new CEOs and audit fee changes. The estimated coefficient on New CEO Heir is +0.019. The implied standard error for this estimate is 0.015. Hence a two standard error confidence interval for the effect on fee change is -0.023 to +0.053. The same analysis reports that the estimated mean effect of the CEO change being to an outsider is a 9.86% increase in fee. Hence, this analysis is unable to reject a null that the new CEO heir fee effects are as much as 50% of the fee increases experienced when an outsider CEO is hired. While this effect is certainly smaller (consistent with the rejection of the null hypothesis of no difference in audit fee change), in context it does not seem at all consistent with assertions that there is a reliable basis in the reported evidence for believing that the New CEO Heir fee effect is immaterial or non-existent.
Brasel et al.	Brasel et al. evaluate the impact of the auditor reporting an unrelated CAM on verdict outcomes. The baseline (control) estimated level is a negligent judgement 42.1 % of the time. This estimated level drops to 36.4% when an unrelated CAM is reported by the auditor, a decline of 5.7 percentage points. . The standard error for this estimate is around 6% points. Hence, the estimated confidence interval for the effect ranges from -17.7 percentage points to + 6.3 percentage points. It is quite difficult to fathom how one credibly advances a claim that unrelated CAMs "neither increase nor reduce auditor liability judgements" is a plausible interpretation of such evidence.
Brazel et al.	OutcomeXConsult lacks significance in Table 2 leading to the conclusion that "outcome bias is not mitigated by either form of consultation." Neither effects or t-values are reported. Hence, it is effectively impossible to construct confidence intervals based on the provided information.
Cannon & Bedard	Canon and Bedard obtain a number of null outcomes. But their analysis is based on a small sample size and low explanatory power models. Consequently, they obtain low precision estimates as a matter of course. Hence, effects must be sizable to have much chance of being reliably detected in this analysis. As a representative example, we evaluate their examination of the likelihood that a valuation specialist is used when Level 3 assets are present. The estimated effect reported in table 4 for this examination is -0.73, which certainly does not favor the notion that a valuation specialist are called in due to the presence of Level 3 assets. However, the implied standard error for this estimate is a rather substantial 0.97. Hence, the two standard error upper bound on this estimate is +1.21. Or, in terms of likelihoods, roughly 235%. Hence, based on the evidence considered in this analysis we cannot reliably rule out the possibility that the presence of level 3 assets increases the likelihood that a valuation specialist is consulted by 235%. This level of in-precision hardly seems the basis for asserting that there is no association

	between LEVEL3 and the use of a specialist as representing a "key result" or a "new finding". (p.106)
Casas-Arce et al.	Table 9 of Casas-Arce et al reports its analysis of the determinants of mortgage pricing. Its no decrease in price inference rests on the fact that the difference between the Branch and internet Mortgages Base *Post-CLV* coefficient estimates of +0.346 lacks significance (fn. 23). The reported standard errors for the two coefficient estimates are 0.434 to 0.770. Hence, an estimator for the standard error for the difference in mean between the two groups is the square root of the sum of these two values, 1.051. The associated two standard error confidence interval for the difference in the change in pricing is from -1.754 Basis points to +2.447 basis points. While such values certainly strike us as small, they are also arising in what seems to be a highly competitive market setting. In highly competitive settings pricing differences are generally expected to be rather tiny.
Chen et al.	Chen et al. (2016) test the null hypothesis that the financial reporting quality of U.S. reverse merger (RM) firms does not differ from that of U.S. IPO firms using four accounting quality measures: restatements; accounting errors, accounting irregularities, and a battery of five accrual quality measures. Based on the null outcomes from these tests they conclude that in terms of accounting quality U.S. RM firms "do not differ from" U.S. IPO firms. This conclusion is important for their analysis as it allows them to avoid specifying how to meaningfully equate differences in reporting quality for U. S. firms with differences in reporting quality for Chinese firms. The results are presented in their table 3 and table 4, as measured by the coefficients for the RM variable. The estimated effects of the RM process for these variables are presented in tables 3 and 4 of their paper. In this analysis we exclude the last two accrual-based measures in table 4 because we could not devise a reasonable approach to assess their magnitudes given the limited amount of descriptive information available to us. In terms of the first three measures (all restatements, accounting errors, accounting irregularities) the RM effect estimates are 0.593, 0.696, and 0.244. All three are positive, which is directionally consistent with RM firms exhibiting lower reporting quality than IPO firms. As these are all from logit regressions, we can convert their estimates into odds ratios of 80.94% more likely to restate, 100.57% more likely to experience an accounting error, and are 27.6% more likely to report irregularities. While these effects lack statistical significance, they most certainly are not near 0. Hence, it is hard to see how they justify an inference that there is no difference or even only a small difference in quality between U.S. IPO and RM firms. The three accrual quality measures that are evaluatable are: absolute value of discretionary accruals (DA), the absolute value of working capital accruals (DD), and the absolute value of discretionary revenue (DR) into our analysis. The estimated effects for these three measures are -0.008, 0.006, and 0.001 respectively. Taking the means of matched U.S. IPO firms from their table 1 as scaling variable (0.17, 0.08, and 0.07) the RM relative to IPO differences are around -4.7%, 7.5%, and 1.4% of their mean values.These do not seem

	particularly large, so, at the magnitude of effect level, the differences are arguably small. Further, we generate two standard deviation confidence intervals for these six RM estimates by first dividing the reported effect by its associated z-value or t-value, as standard error estimates are not directly reported. This yields standard error estimates of 0.371, 0.470, 0.841, 0.015, 0.010, and 0.009, respectively. Then, the following two standard deviation confidence intervals are obtained: (1) All restatements: LB = -0.149; UB = 1.335 (2) Accounting errors: LB = -0.244; UB = 1.636 (3) Accounting irregularities: LB = -1.438; UB = 1.926 (4) The absolute value of DA: LB = -0.038; UB = 0.022 (5) The absolute value of DD: LB = -0.014; UB = 0.026 (6) The absolute value of DR: LB = -0.017; UB = 0.019 Converting the first three of these into odds ratios gives us: (1) All restatements: LB= -13.84%; UB = +280.00%; (2) Accounting errors: LB = -21.65%; UB = +413.46%; (3) Accounting irregularities: LB = -76.25%; UB = +586.20%. The upper bound values, which are particularly relevant in the context of this analysis, here are astronomical. It is not clear how one can say much of anything at all substantive here at all against the possibility that U.S. RM firms have far higher restatement, error, and irregularity rates than do U.S. IPO firms. For the three accrual quality measures the mean value scaled UBs are 12.94%, 32.5%, and 27.14%. That is, the presented evidence here does not rule out the possibility that the accrual quality of IPO firms is as much as 32.5% higher than the accrual quality of RM firms. Hence, there again does not seem to be a very plausible basis for thinking that we can reliably infer that the difference between IPO and RM firms here is small.
Choi et al.	Table 2 of Choi et al. presents a null outcome with respect to whether a difference in the level of strategy experimentation differs between participants in the grand tournament setting and participants in the hybrid tournament setting. The estimated mean difference in strategy experimentation is 0.30 with an associated standard error of 0.34. Hence, the two standard error confidence interval for this difference is from -0.38 to +0.98. There are two plausible scales available here for evaluating these magnitudes. The first is the estimated mean value of experimentation across the two groups, which seems to be around 4.0. Using 4.0 as a scale results in a scaled confidence interval of -9.5% to +24.5%. Alternatively, the standard deviation for the experimentation variable seems to be roughly 1.1. Using this value to scale the bounds gives a confidence interval in units of the underlying variable's standard deviation of between -34.55% and +89.10%. While these bounds seem to reliably rule out the possibility of moderately lower and substantially larger experimentation means in the grand experiment, they do not seem nearly precise enough to

	infer that the level of strategy experimentation is similar in the two tournaments.
DeFond et al.	An absence of descriptive information for the independent unconditional conservatism measures precludes substantive descriptive analysis of the reported coefficients and associated (not reported but estimable) confidence intervals. The paper's companion conditional conservatism tests employ decile ranks. If the unconditional conservatism measures are also decile-ranked then meaningful descriptive analyses of effect sizes and likely ranges is feasible from the reported numbers. However, the text of the paper never states that this is done, and some language used actually implies that "as is" rather than transformed variables are used. Moreover, simple inspection of the reported magnitudes and associated test statistics strongly suggests that transformations are not used in these analyses. In particular, if rank transformations are assumed, estimated effect magnitudes are seemingly astronomical for one of the two unconditional conservatism measures and remarkably miniscule for the other.
Drake et al.	Table 6 of Drake et al. examines UTB (uncertain tax benefit) and change in UTB values for Deloitte clients relative to these values for other clients by year. As pertinent descriptive information is provided for UTB we focus on this set of results here. The 2012 estimated UTB difference for Deloitte is -0.0011, which is consistent with Deloitte clients actually reporting lower UTB values (which favors the authors' position that the UTB values of Deloitte clients are no longer higher than the clients of other auditors.) The two standard error upper bound on this estimate is +0.0007. Given that average UTB level is 0.013 with a standard deviation of 0.023 it seems reasonable to view this upper bound value of being quite small, consistent with the generally conclusive interpretations provided by for it by the authors. However, as has been noted for a few other articles, the underlying evidence for such an interpretation is never formalized in the text presentation.
Dutta & Patatoukas	Dutta and Patatoukas cite several null outcomes from placebo tests to validate their construct of conditional conservatism. Here we focus on two of these tests: (1) the test of the null hypothesis that the spread between the lagged accrual variances conditional on the sign of future unexpected returns is zero; (2) the test of the null hypothesis that the spread between the lagged cash flow variances conditional on the sign of future unexpected returns is zero. Empirical results are presented in their Table 6. The estimated spreads between the conditional variances of bad news and good news lagged accruals and lagged cash flows are -0.45% and -0.05%, respectively. They suggest that the conditional variance of bad news lagged accruals is 7.27% lower than that of good news one, and the conditional variance of bad news lagged cash flows is 1.26% lower than that of good news one. Since the estimates lack significance at conventional levels, the authors conclude that they "find no evidence of asymmetry in the conditional variances of lagged earnings components". We generate two standard deviation confidence intervals for these spreads by first dividing the reported effect by its associated t-value, as standard error estimates

	are not directly reported. This yields standard error estimates of 0.009375 and 0.002381, respectively. Then, we obtain two standard deviation confidence intervals of the spreads: LB = -2.33% and UB = 1.43% for lagged accruals and LB = -0.53% and UB = 0.43% for lagged cash flows. Converting them into percentages of overall conditional variances of good news variables gives us "LB" = -37.64% and "UB" = 23.10% for lagged accruals and "LB" = -13.35% and "UB" = 10.83% for lagged cash flows. That is, the presented evidence does not rule out the possibility that rather sizable differences exist in accrual and cash flow variances conditional on whether future return is positive or negative.
Erickson et al.	Erickson et al. does not provide any descriptive statistics for the null outcome besides observing that the p value >.50.
Farrell et al.	Farrell et al. examines honesty rates for students relative to online subjects (wokers) under comparable high pay conditions under two trust contracts and obtains null outcomes under both conditions. They conclude that "online workers' honesty in reporting does not differ from that of student paricipants". Group differences (student honesty rate less online honesty rate) are +8.6 percentage points for the "trust contract" and -6.5 percentage points for the modified trust contract. Based on the reported t-values of 1.46 and 0.96 for these two differences the estimated standard errors here are 5.89 and 6.77 percentage points. Hence, the relevant two standard deviation confidence intervals are LB = -3.18, UB =+20.38 for the "trust contract", and LB = -20.04, UB = +7.04 in the "modified trust contract." The LB estimates are of central interest here as the underlying hypothesis concerns the possibility that online participants are more dishonest. Hence, these values suggest that online participants are, at most, only slightly more dishonest than student participants in the "trust contract" setting. In the "modified trust contract" setting, however, it is hard to fathom how the possibility that the online worker honesty rate is as much as 20 percentage points lower than that of the student workers is compatible with an inference that no substantive difference in honesty rates is present under this sort of contract. More generally, as this paper contains multiple null outcomes, it is pertinent to note that the 20 to 30 percentage point confidence intervals documented here are representative of the sort of confidence intervals found throughout the article. Hence, in general, the analysis is not producing the sort of high precision estimates needed to substantiate propositions that underlying effects are reliably near zero, small or even, for that matter, something other than possibly very large.
Francis et al.	No descriptive information provided.
Frederickson & Zolotoy	Discussed in main body of article.
Gong et al.	In table 4 a null outcome occurs in the test of whether client accrual quality changed in the post-audit firm merger period. The estimated coefficient magnitude is -0.004, which the paper takes as indicating that "client firms' accrual quality is not affected by firm mergers." Based on the reported t-value of -0.947 the implied value of the standard deviation here is .0042. Hence, the two standard deviation CI is: LB = -0.0124, UB = +0.0044. The accrual

<table>
<tr><td></td><td>measure, AbsDA, has a standard deviation of 0.084. Using this to rescale the CI in terms of standard deviations of AbsDA gives: LB = -.14.8% and UB = +5%. Absent further qualitative insights these values appear to be consistent with the inference that any sort of accruals quality effect that is present, particularly downside effect, is reliably small.</td></tr>
<tr><td><b>Guenther et al.</b></td><td>Guenther et al. (2017) examine the prediction that tax avoidance policies that reduce ETRs (Effective Tax Rates) are associated with a greater degree of tax rate volatility. Table 4 presents their main results and we focus on the null outcomes for the 5-year GAAP ETR and 3-Year Adjusted GAAP ETR measures, which are two of the four measures Guenther et al. identify as central to their analysis. The estimated coefficient effects for these two measures are 0.014 and 0.053 respectively. As the associated standard deviations for 5-Year GAAP ETR and 3-Year Adjusted GAAP ETR are 0.105 and 0.101, these coefficients imply that a one standard deviation shift in 5-Year GAAP ETR is expected to produce a corresponding future volatility shift of .0015 while a one standard deviation shift in 3-Year Adjusted GAAP ETR is expected to produce a corresponding future volatility shift of .0054. As the mean and standard deviation for future volatility are 0.134 and 0.203, these shifts correspond to 1.1% and 4.0% of mean volatility or, .7% and 2.7% of a standard deviation in volatility. In general, these sorts of magnitudes strike us as negligible.

This negligibility assessment, however, is specific to the estimated effect magnitudes and does not take into account the level of precision associated with these magnitudes. Doing so requires confidence intervals. Based on the tabulated t-values, the standard errors associated with 5-Year GAAP ETR and 3-Year Adjusted GAAP ETR are 0.034 and 0.038 respectively. Hence, the associated two standard deviation confidence intervals for the coefficient estimates are:

(1) 5-Year GAAP ETR: LB=-0.054; UB=0.082
(2) 3-Year Adjusted GAAP ETR: LB=-0.023; UB=0.129

The table below translates these bounds into implications of a one standard deviation increase in a given ETR measure for future volatility measured as a percentage of: (1) the mean of future volatility; and, (2) the standard deviation of future volatility.

<table>
<tr><th rowspan="2"></th><th colspan="2">% of Mean Future Volatility</th><th colspan="2">% of S.D. of Future Volatility</th></tr>
<tr><th>LB</th><th>UB</th><th>LB</th><th>UB</th></tr>
<tr><td>5-year ETR</td><td>-4.3%</td><td>+6.5%</td><td>-2.8%</td><td>+9.5%</td></tr>
<tr><td>3-year ETR</td><td>-1.7%</td><td>+4.3%</td><td>-1.1%</td><td>+6.2%</td></tr>
</table>
</td></tr>
</table>

	The tabulated LB values here strike us as being quite small. The UB values, which are more salient to the issues framed by the paper, are larger, but still strike us as at least being arguably small.
Hall	Hall examines the effect of labor cost cuts on future performance and concludes that there is “no evidence that using labor cost reductions to meet financial reporting and regulatory benchmarks improves future financial performance.” (P1691) The author uses ROA for each of subsequent years (ROAit+1, ROAit+2 and ROAit+3) as the dependent variables to measure the future financial performance in Table 7. This inference is based on tests of coefficient estimates for the independent variables SMINCR*LOWLC and LOWCAP*LOWLC. SMINCR is equal to 1 if the bank reports a small increase. LOWCAP is equal to 1 if the bank’s Tier 1 Capital Ratio is in the lowest quartile of the distribution of all banks in the sample. LOWLC is equal to 1 if the bank has abnormally low labor costs in year t. (1) The estimated coefficients on SMINCR*LOWLC are, in terms of basis points (bps), -0.8, 5.5, 11.4 for Public Banks and -0.8, -2.4, -3.7 for Private Banks. (ROAit+1, ROAit+2 and ROAit+3 are multiplied by 100 in the regression per Table7 ROPA defnintions), While these estimates seem rather small, it is important to note that baseline ROAs for banks is generally around 100 basis points (it averages 122 basis points for the paper’s sample per table 3). Hence, an 11.4 basis (the estimated effect for t+3 ROA for public banks) point value is actually rather substantial. Based on the associated t-statistic the estimated standard errors associated with these estimates are -3.6, 5.4, 10.8, 2.7, 4.0 and 5.2. Hence, the two standard deviation confidence intervals are: LB= -8.1bp, -5.4bp, -10.1bp, -6.1bp, -10.4bp and -14.1bb; UB= 6.5bp, 16.4bp, 32.9bp, 4.5bp, 5.6bp and 6.7bp. From the perspective of real activities manipulation, which seems to be the primary focus of this analysis, these bounds indicate that we cannot reliably rule out a future period ROA decline as large as 10.1 basis points for public banks (ROAt+3) or as large as 14.1 basis points for private banks (ROAt+3). imply that ROA will decrease 14.1% point. While for most types of firms a one period increase in ROA of 10 to 14 basis points would be reasonably viewed as small, lending support for the conclusion that any ROA impact here is negligible, bank ROAs are inherently quite low. Hence, we cannot reliably rule out the possibility that material adverse real activities manipulation consequences are in play here. (Note, if we look instead at the future improvement aspect then the bounds are quite a bit larger, meaning we almost certainly should not view this evidence as reliably indicating that there is not a material upside benefit present.).
Humphreys et al.	The text discussion of the H2b results seems to employ values at odds with the reported statistics in Table 1. Hence, we are not sure what set of numbers we should use to conduct a confidence interval analysis. Detailed statistics are not provided for the second null outcome.

Kelly et al.	Kelly et al assess whether sales performance differs conditional on whether cash or tangible awards are provided. In the opening baseline tournament they conclude that "there is no overall difference in sales" between the two reward types. The evidence for this conclusion is reported in table 3. Here we employ the robust regression estimates reported in panel B as their interpretation is not confounded by the presence of an interaction term. The estimated effect of Tangible Reward in panel B is +$48.52. The implied standard error for this estimate, however, is $110. Hence, a two standard error confidence interval here yields an upper bound value of $268.52. As mean prior year sales levels are $866, this upper bound translates into a 31.17% differential change in sales. Hence, the analysis cannot rule out the possibility that the form of reward improved sales levels by as much as 31%. This seems to be far too large a value to justify a claim of there clearly being no significant difference in sales effect between the two reward types.
Kim and Klein	Discussed in main body of article
Krishnan et al.	Krishnan et al. compare the inspection effects for auditors with and without deficiency reports and find no systematic differences for accruals or for value relevance. (1) For accruals, in Table 6, the estimated coefficients on POSTINSPEC*DEF are 0.012 for two-year window and 0.011 for four-year window. POSTINSPEC is indicator variable which equal to 1 for fiscal periods following the inspection. DEF is indicator variable which equal to 1 for observations of cross-listed clients of inspected auditors with deficiencies. We can use the p-values to back into the two-tailed t-values. Based on the associated t-statistic the estimated standard errors associated are 0.21 for two-year window and 0.014 for four-year window. Hence, the relevant two standard deviation confidence intervals are LB = -0.031, UB = +0.055 for two-year window and LB = -0.018, UB = +0.040 for four-year window. We can use the estimated value of the POSTINSPEC effect of -0.044 and -0.025 as relevant basis for judging magnitude here. After dividing the upper bound by the absolute value of these coefficients, we get upper bounds on the percentage of the effect that is being offset of 124.33% and 158.78 %. Furthermore, because the plausible upper bounds on the percentage of the effect that is being offset are greater than 100%, we cannot rule out the possibility that the post-inspection effect is eliminated, reversed even, for firms with deficiencies. Hence, there is simply no reliable evidence here that the underlying effect is not large. 2) For value relevance, only effects are reported. Hence, it is impossible to construct confidence intervals.
Laurion et al.	Laurion et al. report the null outcome for the presence of a relation between misstatement likelihoods after partner rotations. The estimated logit coefficient is 0.367 and based on the associated t-statistic the estimated standard deviation

	associated with it is 0.314. Hence, the two standard deviation confidence interval is -0.261 to 0.943. Conversion of these bounds into likelihoods yields a range of -23% to +157%. Hence, based on the set of data examined in this analysis it is quite impossible to rule out the possibility that partner rotations resulted in very sizable increases in misstatement levels.
Lennox	Discussed in main body of article.
Li et al.	Li et al. (Jan. 2017) test the null hypothesis that there is no difference between audit quality of non-failed auditors in the same office as a failed auditor and that of auditors in offices as a non-failed auditor. Specifically, they regress indicator variable ABS(AB_ACC) and AB_ACC>0 on FAIL_X_ COLLEAGUE. Table 6 reports estimated effects of 0.001 for FAIL_0_ COLLEAGUE and FAIL_10_COLLEAGUE for four models' estimations. The non-reporting of further digits in these three estimations means we cannot reliably estimate the unreported associated standard errors as the underlying value possibly ranges from 0.00149 (50% higher than .001) to 0.0005 (50% lower than 0.001). In the fourth cases the estimated effect is 0.002, which narrows the range of possible values (as a percentage of 0.002) considerably. The associated t-value of 1.36 then implies an underlying standard error of around 0.0015, which in turn yields a two standard error confidence interval of -0.001 to +0.005. However, the audit quality dependent variable in this instance is the subsample of firm-years with positive abnormal accruals, but the article does not report descriptive statistics for this subsample. Hence, we were unable to devise a strategy for evaluating the magnitudes of these bounds.
Lin & Wang	Lin and Wang obtain two null outcomes, both of which concern a measure of innovation efficiency. Unfortunately, while descriptive statistics are provided for most of the variables they utilize, no such statistics are provided of the innovation measure. Consequently, we can provide a pertinent descriptive analysis for only one of them—the absence of a significant relation between the interaction of efficiency with takeover probability and equity returns. The analysis is based on the reported standard deviation of the *Takeover Probalility* variable of 0.16 (per p. 965 of article) and the table 4 reported coefficient estimates of 0.0126 and 0.146 for *innovation efficiency* and its interaction with *Takeover Probability.* Multiplying the interaction coefficient by the standard deviation of *Takeover Probability* indicates how much a one standard deviation difference in takeover probability shifts the overall relation between *innovation efficiency* and returns. This product is 0.023. Consequently, a one standard deviation shift in *Takeover Probability* shifts the overall relation between *innovation efficiency* and returns by 0.023, which is nearly double the magnitude of the general relation between *innovation efficiency* and returns. That is, the presented evidence here clearly cannot reliably rule out the possibility that the relation between *innovation efficiency* and equity returns is highly sensitive to *Takeover Probability* level.
Lourenco	Discussed in main body of article.
Nelson et al.	Nelson et al. report a null outcome for the test of whether alignment between issue and supervisor concerns has less of an effect on an auditor's willingness

	to speak up about an issue when the auditor's supervisor is more team-oriented. They conduct ANOVA test for the intersect TOL * Concern * Issue. In Table 3 Panel B, the corresponding p-value is 0.29. They conclude that they do not find support for the three-way interaction between audit issue, supervisor concern, and team-oriented leadership. While the authors do not report the estimated conditional mean for the TOL*Concern*Issue, we can infer it from other effect estimates that are reported in the table 3 (panel A). Specifically, the effect of alignment between issue and supervisor concerns in Team-Oriented Leadership group is 82.70+ 79.26-64.30-69.00 =28.66, and the effect of alignment between issue and supervisor concerns in Non-Team-Oriented Leadership group is 60.96+65.88-49.18-45.05=32.61. Hence, non-team oriented leadership is estimated to have a 3.95 point (speaking up is measured on a 1 to 100 point scale) increases in speaking up comfort level under these conditions. As the average level of speaking up comfort across the two groups here is 64.5 ((73.8+55.2)/2) this amounts to a 6% increase which is certainly not that large, but is certainly not essentially 0. Moving to the confidence interval determination, we infer a standard error estimate based on the reported p-value for the F-test of 0.29. This p-value corresponds to a t-value of around 1.05, suggesting that the standard error is around 3.76 (3.95/1.05). Hence, the two standard error confidence interval here ranges from -3.57 to +11.47. Or, in terms of percent of mean confidence level, from -5.5% to +17.8%. Hence, the analysis is unable to rule out the possibility that non-team oriented leadership increases speaking up comfort levels by as much as 17.8% relative to average.
Nessa	Column (5) of table 5 in Nessa reports a null outcome for the unconditional relation between repatriation costs and the level of repatriation exhibited by firms. The estimated effect is -0.0317 and the implied standard error is 0.0793. Hence, a two standard error confidence interval is from -0.1903 to +0.1269. The mean value of repatriation costs is 0.0023. Multiplying these bounds by this mean provides insights about the implications of this confidence interval for an "average repatriation cost firm." Specifically, the estimated repatriation level effect for such an average firm ranges between -0.00044 to + 0.00029. The average (unconditional) repatriation level here is 0.0175. Consequently, when expressed as a percent of this level these bounds become -2.5% to +1.7%. These values seem broadly consistent with assertions that the impact of repatriation costs on repatriation levels is small.
Patatoukas & Thomas	Table 4 of Patatoukas and Thomas reports null outcomes for the relations between expected return and three different expected earnings constructs. All three estimates are negative, and the focal concern in the paper is that they are possibly positive. Hence, the negative estimates are consistent with the general inference that the relation is not positive. The estimated two standard error upper bounds for the three measures, however, are +.008, +.018, and +.012 respectively. Dividing these by the estimated value (.251) of the asymmetric timeliness coefficient (which is what is being decomposed here) provides percentage of total effect upper bounds of 3.2%, 7.17%, and 4.78%. These magnitudes seem reasonably consistent with the idea, at least from a positive

	side perspective, that the effect is fairly negligible or, in the words of the article, "close to zero." This assessment does, however, depend on the sorts of magnitudes we might expect to see here. It is also important to recognize here that this effect is being advanced as a mechanism for explaining an anomalous "Placebo" effect, which in this analysis is estimated at 0.159. When the above three upper bounds are scaled by this magnitude we get values 5.0%, 11.3%, and 7.5%. Hence, this effect can only reliably explain a small fraction of the anomalous effect that is in question here. (We thank Jake Thomas and Panos Patatoukas for suggesting this Placebo effect scaling perspective.)
Robinson et al.	Discussed in main body of article.
Schroeder & Shepardson	Table 6 of Schroeder and Shepardson report null outcomes for tests of whether the management assessment requirement affected accrual quality. Here we evaluate the metric based on the unexplained residual variation in accruals (UAQ_NOISE) as this measure can be reasonably scaled by the sample mean, which provides a reasonable basis for understanding underlying effect magnitudes. The estimated effect of imposing the assessment requirement equals -0.0014 in column (2). As negative values are consistent with reduced levels of unexplained variation, this estimate is directionally consistent with the conjecture that the assessment requirement improved accrual quality. Dividing by the average value of UAQ_NOISE for non-accelerated filers in the 2007 to 2011 time period of 0.042 (reported in table 2 of their analysis) converts this value into a percentage: 3.33%. Hence, the estimated effect suggests that a best guess estimate of the assessment requirement reduced unexplained variation in accruals by 3.33%. While this is certainly not a huge change, it is hard to say absent further descriptive perspective that it is negligible. The imputed standard error (in this case t-values are imputed from reported p-values for purposes of imputing the unreported standard error) here is .002, meaning that the two standard error lower bound of the estimated value is actually -0.0054, or 12.86%. Hence, based on the reported evidence in this study, one cannot rule out the possibility that the assessment requirement resulted in a reduction in unexplained accruals of well over 10%. This possibility does not quite square with the article's conclusion that "our results suggest that SOX 404(a) management assessments do not yield significant improvement in internal control system quality."
Towery	Towery reports a null outcome for tests of whether firms subject to Schedule UTP do not experience a decrease in FedCashETR and CashETR. The coefficients on SchUTPInd reported in table are -0.0092 for FedCashETR and -0.0168 for cashETR. The associated standard errors, estimated from the reported t-statistics, are 0.012 and 0.017. Hence, the respective confidence intervals are: -0.0332 to +0.0148 and -0.0505 to +0.0168. The underlying standard deviations for these two changes in tax rate variables are 0.0973 and 0.2251 implying standard deviation scaled bounds of -0.341 to +0.152 and -0.224 to +0.075 which are not overly large but not small, particularly on the downside, either. However, the fact that these are change variables undercuts the usefulness of this scale. The mean or the standard deviation of the levels of

	these variables would be far more meaningful scales. Neither of these is reported in detail, but Figures 1 and 2 do provide information about their levels. In particular, FedCashETR seems to average around 0.06 while CashETR seems to average around 0.12. Unfortunately, the decimal level scaling used in these figures does not seem to be the same as that used for the changes, since the adjusted bound values based on them are insensibly large.
Wieczynska	Panel C of Wieczynska reports a null outcome for whether there is a change in the likelihood that firms in weak enforcement regimes switch auditors in the IFRS adoption year (for their country). The estimated coefficient in the binary change model is 0.03, with a standard error of 0.15. Hence, the two standard error confidence interval here is. -0.27 to +0.33. Converting these values into likelihoods we get a confidence interval of -23.7% to 39.1%. That is, based on its evidence, this study is unable to rule out the possibility that the likelihood that firms changed their auditor upon their country's adoption of IFRS increased by as much as 39%. While this evidence could likely support a conclusion that auditor changes did not dramatically increase, they do not seem to justify a "rejection" of the proposition that they did not increase at all.

In six cases (**bolded authors)** the estimated bounds are consistent with the effect being, in our judgement, plausibly thought of as "small." In five cases (*italicized authors*) the information reported in the article was insufficient for the determination of meaningful confidence intervals.