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

July, 2019

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
Hong Kong Polytechnic University and Xiamen University
zhangyang_xmu@foxmail.com

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

This paper has benefited from the comments of workshop participants at Lingnan University, Elizabeth Gutierrez, Tom Omer, Jake Thomas, and several anonymous reviewers.

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

ABSTRACT

This study evaluates how accounting researchers analyze and report null outcomes based on an examination of recent accounting research publications. As null outcomes reflect an inability to reject a null they, unlike rejections, do not lend themselves to specifically conclusive interpretations. Rather, drawing useful inference from them requires fundamental descriptive analysis. 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. The analysis also proposes and illustrates the use of descriptive techniques in the form of confidence intervals as a more appropriate approach for interpreting null outcomes.

I. Introduction

This study addresses of how the academic accounting literature reports and interprets null outcomes. We define a null outcome as an instance where a non-directional test of a (null) hypothesis results in *p*-values that are larger than what is required to sustain a rejection inference.[1] Or, in other words, the associated effect lacks statistical significance at conventional levels. Null outcomes pose a particular challenge for conventional *p*-value-based inference. According to Principle No. 1 of the recently promulgated *American Statistical Association (ASA) Statement on Statistical Significance and P-Values* (Wasserstein and Lazar, 2016) a *p*-value summarizes the incompatibility of the examined evidence with a proposed (null hypothesis) explanation for it. The language accompanying Principle No. 2 of the *Statement* directly connects this incompatibility-with-the-evidence perspective to null outcome interpretation when it states that a *p*-value is not "a statement about the truth of a null hypothesis." Greenland et al. (2016) further assert that "it is false to claim that statistically non-significant results support a test (i.e., null) hypothesis" (parenthetical clarification ours).[2] A thought echoed in the Wasserstein, Schrim, and Lazar (2019) introduction to *The American Statistician*'s recent special issue on *p*-values "what not to do" list: "Don't believe that an effect is absent just because it was not statistically significant." In summary, null outcomes do not reflect an incompatibility of the evidence with any identified hypothesis, nor should they be taken as providing meaningful support for the related null hypothesis.[3]

While the *ASA Statement* clearly formalizes the limited inferential value of null outcomes, the methodological underpinnings for its arguments and assertions are hardly new. The idea that

[1] Importantly, a null outcome is not the same as a null hypothesis. A null outcome is produced by a postulated (null) hypothesis in conjunction with an examination of relevant evidence. Hence, there is a substantive difference between advocating the reporting of null outcomes as a descriptively useful exercise and advocating the use of null outcomes as a definitive basis for drawing supportive inferences about associated null hypotheses.
[2] All 7 of the Greenland et al. authors served on the ASA committee that drafted the *ASA Statement*.
[3] "Absence of evidence is not evidence of absence." (article title, Altman and Bland, 1995)

high *p*-values should not be taken as supporting associated null hypotheses follows directly from the fact that the underlying structure of hypothesis testing is firmly rooted in a bias in favor of the tested (i.e., null) hypothesis. A null is only rejected if the evidence is compellingly incompatible with it. It is for precisely this reason that inferring anything about the truth of the null hypothesis solely 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 highly reliable inferences about it being false or incompatible with the examined evidence.[4]

The circumstances leading up to the promulgation of the *ASA Statement* is also a primary motivation for our analysis. The ASA produced this statement in response to their belief that "(the *p*-value) is commonly misused and misinterpreted." The statement identifies its purpose as providing a "formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the *p*-value." That is, the ASA clearly believes that *p*-value misinterpretation is a widespread problem across many academic disciplines. Our analysis examines whether their belief holds for the accounting discipline by examining how it interprets reported null outcomes. Does it avoid providing high *p*-value based conclusive interpretations of such outcomes, which would reflect compliance with the *ASA Statement* principles? Or, does it as a matter of course rely on high *p*-values, and little else, to advance **falsely** evidenced (Greenland et al., 2016) supportive or conclusive claims based on such outcomes?

Our analysis also addresses the use of descriptive analysis provided for null outcomes in the accounting literature. The motivation here also stems from the *ASA Statement*, as it advocates

[4] From this perspective, the commonly encountered complaint of a bias against the null in conducting *p*-value-based inference is a bit like complaining about paying for a lottery ticket that, after the fact, didn't win the lottery.

the use of descriptive analysis as a valuable relevant alternative or supplement to *p*-value based inferenced. Statistics is foremost a descriptive science. Much like financial reporting condenses events and transactions into summary reports, statistics seeks to condense the information existent in a collection of empirical observations (data) into measures such as means and variances that, under specified assumptions, efficiently aggregate the information in the examined data. Hence, while *p*-values and the associated classical hypothesis testing frameworks that give them sustenance are inherently ill-suited to providing useful insights in null outcome settings, the underlying descriptive content of fundamental statistical measures is not so impaired. In other words, null outcomes are open to meaningful rigorous examinations and inferences that flow from statistic-based descriptive engagement. We examine both the extent to which the accounting literature does not provide relevant descriptive analyses for null outcomes and how it can improve on this dimension by engaging in substantive confidence interval assessment of null outcomes.

We address these two issues based on a comprehensive examination of all articles published in *The Accounting Review* over the 2016-2017 time period. 35 of these articles report null outcomes of central importance to the presented analysis, where importance is evidenced by the outcome being connected to a formal hypothesis statement or being discussed in the article's abstract or its introductory sections. The analysis evaluates the null outcomes in each of these articles on two distinct dimensions: (1) the interpretative language provided for the null outcome; and, (2) the presence or absence of descriptive interpretation for its reported null outcomes. Collectively, these analyses reveal that the accounting literature routinely provides highly conclusive interpretations for null outcomes and rarely provides any substantive descriptive analysis or magnitude assessment for such outcomes. That is, when presenting null outcomes the

literature routinely uses the wrong language to interpret them and does not employ appropriate methods to evaluate them.

Finally, our analysis is related to recent work by Basu (2013), Basu and Park (2014), Dyckman and Zeff (2014), Kim and Ji (2015), Ohlson (2015), Dyckman (2016), Kim, Ji, and Ahmed (2018), Harvey (2017), Myer, Witteloostuijn, and Beugelsdijk (2017) and Stone (2018) among others. The focus of these articles, however, is largely on the integrity of low p-value (i.e., rejection) outcomes, best practices for interpreting them, and, in a few cases, why null hypothesis testing is a generally flawed approach to empirical inference. In those instances where these studies do address null outcomes, the discussion commonly takes the form of an observation about the literature not reporting them often enough. These studies provide few insights regarding best practices for reporting them, which is the overarching focus of our analysis.

2. Classical Hypothesis Testing and Null Outcomes

Classical null hypothesis testing flows from the well-known if-then logical paradigm. A specific antecedent 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 affirmative observation of 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 an

observed antecedent (null) consistent consequent 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 noisily, not truly observed. Hence, in statistical inference settings noise is an explanation for any failure to observe a necessary outcome. A researcher is simply not able to falsify the premise with certainty. Instead, researchers rely on *p*-values to make rejection, or better, evidence incompatibility assertions. In the case of null outcomes, allowing for noisy measurement gives rise to irrefutable alternative explanations that are consistent with the null outcome evidence (some even more consistent with it), but also 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 the elaboration provided for Principle 6 of the *ASA Statement* which 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." Hence, the classic textbook perspective of null outcome reporting is to report them with neutral language such as "unable to reject."

A closely related setting that also involves the affirmation of the consequent fallacy arises with respect to interpreting the alternative research hypothesis after rejecting the associated null. However, there are important distinctions between this sort of exercise and null affirmation assertions. First, these interpretations begin from the perspective that the null hypothesis claim is 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, such an outcome implicitly indicates that the examined evidence is compatible with both the null and alternative research hypothesis(es).

Second, in classical hypothesis testing 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, this alternative 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.[5] In others, the accompanying analysis goes to great lengths to rule out or control other explanations for the null hypothesis rejection.[6] 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 alternative 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.

The preceding discussion is philosophical and hence somewhat abstract. Hence, we now complement it with a more empirically grounded illustration. For this purpose, we use a null outcome presented in a recent publication by Kim and Klein (2017).[7] This null outcome pertains

[5] 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 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.

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

[7] Two factors motivate our choice of Kim and Klein here. First, it is one of two null outcome reporting studies in our sample that exclusively report null outcome "findings." The other study, Lennox (2016), is discussed in depth in a later section. Second, it was selected for broader dissemination via an American Accounting Association press release under the caption: "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 notable research contribution.

to a test of whether there was an economy-wide change in the market values of firms in response to the passage of a rule change imposing mandates on audit committee composition and independence. The associated implied 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 is arguably detrimental (decrease market value). In if/then terms the relevant assertion here is that if the rule change is not, on average, beneficial then the mean market response to the rule passage will be zero (or negative). The associated alternative research hypothesis stems from "entrenchment theory," which suggests that entrenched managers sacrifice market value to maintain and exploit their entrenched status. If the rule change lessens the entrenchment power of these managers firms then firm value should increase rather than decrease or be unaffected, as implied by the market theory-based null.

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 is this outcome truly saying? If one follows generally accepted test of hypothesis interpretive guidance, it tells us very little. The examined evidence is not incompatible with the no net positive benefit null. However, neither is there a basis for claiming that the evidence is incompatible with alternative entrenchment theory hypotheses (i.e., positive values for the market valuation impact from the rule change.) Hence, based solely on the null outcome, all one can infer here is that the examined evidence does not convincingly disfavor either hypothesis.

Given this "nothing much definitive 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 of considerable interest. The study's introduction states: "We find, on

average, no statistically significant cumulative abnormal returns…" (p. 188). This statement is an arguably reasonable description for a null outcome (although Wasserstein et al. (2019) disagree). The sentence following it, however, expands upon this non-significance assertion. It advances the claim that that "Thus, the market assigned no overall net benefit or cost to compliance." This assertion is an affirmation of the consequent. It is fallacy.[8] Moreover, at even a basic descriptive level, it is misleading. 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 points increase in market value. 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 compatible 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.[9]

3. The Descriptive Perspective

Descriptive approaches to conducting statistical analysis are commonly advanced as an alternative or supplement to classical null hypothesis testing. In this regard the *ASA Statement on Statistical Significance and P-Values* emphasizes "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 Significance Testing) is but a starting point and that

[8] Alternatively, a description emphasizing the inconclusive nature of the evidence (e.g., the benefit of the rule change is unclear) would be logically sound.

[9] If 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's 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.

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)

Descriptive approaches are particularly relevant for null outcomes. By definition, a null outcome means that the test of hypothesis analysis has returned empty handed. The examined evidence is compatible with the null, which is the focus of the test. But, the evidence is also compatible with the alternative research hypothesis, which is the rationale for questioning the null to begin with. In fact, within the rigid framework that such testing operates, it follows that a null outcome has no incompatible-with-the-evidence implications whatsoever for any considered hypothesis. In contrast, a null outcome does not negate the relevance of substantive descriptive analysis. In this regard Aberson (2002) argues that descriptive analysis is more important for null outcomes, stating that "presenting results that 'support' a null hypothesis requires more detailed statistical reporting than do results that reject the null hypothesis."

3.1 CONFIDENCE INTERVAL RELEVANCE

In this analysis we advance confidence intervals (CIs) as a readily implemented relevant approach for obtaining insights in null outcome settings. CIs are widely invoked as a counterbalance to common mis-interpretation and over-interpretation of *p*-value based test of hypothesis inferences. The *Publication Manual of the American Psychological Association* (2009), states "complete reporting of all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are the minimum expectation for APA journals." Stone (2018) argues for the "mandatory reporting" of CIs by accounting journals. Greenland et al. (2016), while recognizing that CI misinterpretation is a concern, nevertheless advances interval estimates as an important aid "in evaluating whether the data are capable of discriminating among various hypotheses about effect sizes, or whether statistical results have been misrepresented as supporting one hypothesis

when those results are better explained by other hypotheses." Five of Cumming's (2014) 25 guidelines for improving psychological research reference CIs. (p. 33) Specific to null outcome settings, 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."[10]

Strictly interpreted, a CI is a statement of a range of values that, in expectation, contains the true value of the underlying population parameter a specified percentage (i.e., the confidence level) of the time. Consistent with this definition, our analysis follows Greenland et al. (2016), Greenland (2018, 2019), and Amirhand, Trafimow, and Greenland (2019) by interpreting the CI as presenting a range of effect size hypotheses or conjectures that are compatible with the evidence. An effect size conjecture is evidence-compatible if the relevant test of its equality with the true underlying parameter value returns a null outcome. The idea here is that in null outcome settings descriptive interpretation should focus on the range of hypothetical effect values that satisfy the qualifying condition of being null outcomes, not myopically focus on specific effect values that happen to align with a chosen null hypothesis or reference point as exclusive representations of the evidence.

3.2 CI-BASED EFFECT SIZE ASSESSMENT

The descriptive implications of a CI for effect size flow from its identification of the set of effect size beliefs that are compatible with the presented evidence. A CI upper (lower) bound identifies the largest (smallest) effect size belief that is compatible with the examined evidence. Compatibility, however, simply supports the idea that an effect is an evidence-compatible possibility. It does not suggest that it is a certainly or even particularly likely. Neither is the CI a

[10] CIs do need to be interpreted with some care. Greenland et al. (2016) identify 5 common misinterpretations of confidence intervals. (As contrasted with 18 commonly encountered misinterpretations of p-values.)

definitive statement regarding the posterior distribution of effect sizes. Such a distribution, in particular, requires the formal incorporation of prior beliefs into the assessment of the examined evidence.[11]

In null outcome settings a central descriptive concern is what the examined evidence implies regarding the smallness of the underlying effect size, where effect size is measured based on deviation from the postulated null hypothesis value(s). A particularly useful construct in such settings is what Goodman, Spruill, and Komaroff (2019) identify as the "minimum practically significant distance" (MPSD) value. An MPSD value is the "smallest observed distance from equaling exactly the null that could be considered meaningfully large." Goodman et al. draws upon MPSDs as a basis for conducting "thick null" minimum effect plus p-value (MESP) hypothesis testing, wherein the null is rejected only if the estimated effect is statistically significant and its estimated magnitude lies outside of the null defined MPSD interval.[12] In null outcome settings a thick null specification gives rise to the possibility that the CI bounds place an effect size exclusively inside the MPSD defined region around the null value. Such an outcome indicates that the examined evidence is highly compatible with the underlying effect being materially indistinguishable from the postulated null hypothesis value(s).

Betensky (2019), in fact, uses the MPSD concept as a basis for sorting CI evidence into three distinct categories. She argues that if a high confidence level CI falls within a plausible MPSD defined interval then the researcher takes the evidence as signaling that the underlying effect of interest is effectively non-existent. If, instead, the CI extends beyond plausible MPSD based interval values, then the evidence is simply inconclusive with respect to whether or not a

[11] Under the assumption of uninformative priors then, in many cases, the CI can be taken as presenting a posterior distribution of expected effect size beliefs.

[12] Assessment of the "economic significance" of estimated effect sizes that are commonly encountered in the accounting literature are an ex post variation on this concept.

substantive effect is present.[13] Finally, if the CI falls entirely outside of the MPSD interval then the evidence is consistent with the existence of a substantive underlying effect size.[14]

4. Empirical Issues

The preceding discussions raise two specific empirically addressable issues with respect to null outcome reporting in the accounting literature. First, any sort of substantive interpretation of a null outcome requires empirical analysis and support that goes well beyond the *p*-values and associated test statistics that led to its 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. Are null outcome interpretations typically 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 CIs?

Second, we are interested in empirically documenting what sorts of language the literature is providing for null outcomes. Nickerson (2000), among others, identifies several 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

[13] Blume, McGowan, Dupont, and Greevy (2018), within a more general interval based testing structure identified as "second-generation *p*-values," propose using the overlap between the interval null and the examination of the evidence-based CI as a measure of the degree to which the evidence supports the interval null.

[14] This interpretative structure is also addressed in depth by Blume, Greevy, Welty, Smith, and Dupont (2019) as a form of "interval null hypothesis" testing. They proposed that interval widths (e.g., MPSDs) be based on "limits on physical or experimental precision in outcome measurements, measurement error, clinical significance, and/or scientific relevance." The structure also loosely parallels Bayesian region of practical equivalence (ROPE) based inference (Kruskal, 2011).

consistent with 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.

5. Empirical Analyses of Null Outcome Reporting

5.1 IDENTIFICATION OF PUBLISHED NULL OUTCIOMES

Our empirical analysis draws on a set of null outcomes reported in the literature obtained from a comprehensive examination of articles published in *The Accounting Review* in 2016 and 2017. We conduct our empirical analyses of the analysis and interpretation of null outcomes at a rather in-depth level. The fundamental observational unit here is 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 of them. The analysis makes extensive use of specific examples discovered in the sample. The generalizability of this illustration by example approach actually declines as the sample size increases since it is far easier (bordering on trivial) 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 mitigates criticism that

we have "cherry-picked" articles for study. That is, 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 examine articles from a single journal using a time period with natural start and end points. We also desired the sample to be representative of high-quality accounting research, as broadly defined as possible. Given these objectives, *The Accounting Review* seems a clearly obvious first 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 this population as the most reasonable single journal-based proxy for the general state of the accounting literature.

We identify articles reporting null outcomes from 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, are identified in each paper. We then separate 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. We take 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 as having central roles in a study. We exclude outcomes pertaining to robustness tests, validity checks, or control variables, unless article abstracts specifically mention them or article introductory sections emphasize them. We located 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. 15 of the studies address auditing, possibly indicating a predisposition for null outcomes in this domain. Most of the articles (20) contain only a single null outcome. However, one 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. Finally, an explicit statement of the associated null hypothesis or prediction accompanies 21 of the 63 null outcomes.

5.2 NULL OUTCOME DESCRIPTION

For our purpose, relevant descriptive analysis for a null outcome involves most anything 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.[15]

[15] Several of the articles also provide multiple null outcomes in support of a given null outcome-based conclusion. It is tempting to view such null outcome consistency as supporting the truth of the tested null hypothesis. But, as Greenland et al. (2016) observe such a belief is invalid. Multiple null outcomes "should not be taken as implying

Table 2 presents the sorts of descriptive statistics that accompany each of the identified null outcomes we examine by article. Arguably, the most readily obtainable comprehensive descriptive statistic for a null outcome is a high confidence level CI. However, this key statistic is never reported.[16] Indeed, we were unable to locate a substantive discussion of the range of effect values that are compatible with the examined evidence for any of the examined null outcomes.[17] 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 compatible with the observed 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 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. No article 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 null outcome associated estimated effects in their text discussion. Moreover, most 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

that the totality of evidence supports no effect." (p. 343) Indeed, formal aggregations of null outcomes (e.g., by meta analyses) may well supplant a multitude of individual null outcomes with an overarching null rejection.

[16] 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.

[17] We also found no references to discipline- or setting-specific generally accepted effect size guidelines such as that provided in Cohen (1992) for assessing the importance of behavioral intervention effects.

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 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 implied by Principle No. 1 of the *ASA Statement* such values do reflect a form of compatibility between the postulated null hypothesis and the examined evidence. However, it is not clear how much they add on this dimension relative to what is readily inferable from the reported estimated magnitudes of the underlying effect values. They also tell the reader nothing as to whether the source of this compatibility is due to effect size or to the poor quality (i.e., noisiness) of the examined evidence.

The general lack of descriptive textual engagement with estimated effect magnitudes associated with null outcomes documented here stands in stark contrast to 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.[18] In an untabulated analysis, we found that over half of the articles reporting null hypothesis rejections in *The Accounting Review* over the 2016-17 period provide some form of substantive "it is large" descriptive assessment of the estimated effect size.[19] That is, in settings where the relevant literature (Aberson, 2002) argues for providing substantially more descriptive

[18] 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. We would further argue that effect smallness is also the true relevant question in null outcome settings.
[19] Only four articles (Drake et al. (2016), Lennox (2016), Henry and Leone (2016), and Robinson et al. (2016)) provide some form of non-trivial coefficient "smallness" discussion of effect magnitudes.

analysis, the accounting literature does the opposite. It provides far less descriptive analysis for null outcomes than for null rejections.

5.3 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: 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 present at all.[20] 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. For instance, studies commonly state that an outcome is "inconsistent with" the alternative (e.g., Guenther et al., 2017; Kim and Klein, 2017; Lourenco, 2016). Most glaringly, Wieczynska (2016) asserts that a null outcome constitutes a rejection of the alternative hypothesis (p. 1269).[21]

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

[20] In a few instances the precisely conclusive phrasing is accompanied by less conclusive qualifications such as "indicates", "implies." In general, we ignored such qualifications unless it was very clear that they are negating 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).

[21] An attribution of "inconsistency" has the appearance of lacking a high degree of conclusiveness. However, a null outcome is not, absent additional descriptive insights, possibly inconsistent with anything. That is, it is not a basis for rejecting a hypothesis that an effect is: (1) positive; (2) negative; or, (3) zero.

concluding or even inferring that a null hypothesis is true, or thinking 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.

We identify a null outcome interpretation as Generally Conclusive when it advances the notion that the tested null hypothesis is, for practical purposes, approximately true. Claims that an 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. The GC category also aligns with MPSD (minimum practical significant distance) and ROPE (region of practical equivalence) notions regarding immaterial effect sizes. For instance, a CI falling entirely within a relevant MPSD defined interval would merit GC interpretation. However, GC interpretations based on nothing more than a statistically insignificant outcome violate Principle 5 of the *ASA Statement:* "A *p*-value, or statistical significance, does not measure the size of an effect."[22]

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 cherry picking from the set of available individually arguably acceptable (but incomplete) descriptions for the null outcome. Principle 6 of the *ASA Statement*,

[22] "Small" effect assessment is also connected to the notion of performing pre-test power assessment based on hypothetical large enough to matter or effect values (see, for example, author instruction 2-10 for submissions to the registered reports *Journal of Accounting Research* conference (2017) as reported in the appendix of Bloomfield, Rennekamp, and Steenhoven (2018)). Properly implemented, such pre-test assessments identify whether a proposed analysis has sufficient power to reliably detect the smallest meaningful effect that might be present. An analysis possessing such power sufficiency is, ex ante, **expected** to identify a meaningful effect, if it is present. It is not, however, equivalent to an ex post assessment of the degree to which the examined evidence is compatible with the absence of a meaningful effect.

however, clearly indicates that a null outcome does not in any way rule out the consistency of other hypotheses, particularly relevant alternative hypotheses, with the evidence. Greenland et al. (2016), in fact, assert that "It is simply false to claim that statistically non-significant results support a test hypothesis, because the same results may be even more compatible with alternative hypotheses--even if the power of the test is high for those alternatives."

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. Arguably conclusive identifications mostly involve the attribution of statistical insignificance interpretations to null outcomes. The notions of statistical significance and the lack thereof pose a particularly difficult challenge when presenting null outcomes. 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 observing a large *p*-value.[23] In contrast, descriptions such as "unable to reject, "not reliably different," and "no reliable evidence of" are more neutral. We classify these sorts of interpretations as being non-conclusive.

[23] Lindsay (1994) addresses the presence of this "tendency to equate scientific significance with statistical significance" bias in the accounting literature. An example (there were several to choose from) of this false equivalency in the set of null outcome studies examined here is found in Lin and Wang (2016) when they 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)

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 these categorizations of textual null outcome descriptions is available upon request from the corresponding author. 29 of the 35 articles in our study discuss null outcome in their abstracts. Collectively, table 3 indicates that these abstracts present 43 individual null outcomes. 27 of these are described in a precisely conclusive fashion. Another 8 are described using generally conclusive language, while the final 8 are described with selectively conclusive language. 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.

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 steers entirely clear of, at some point, asserting in highly specific terms that an observed null outcome evidences the unequivocal truth of the associated null hypothesis. Equally alarming, only two articles, Cannon and Bedard (2017) and Henry and Leone (2016), manage to provide a clearly inconclusive descriptions of a null outcome. Cannon and Bedard, does so when it observes for its hypotheses H5b and H6b that “model results do not reject the null for both constructs” (p. 99). Henry and Leone do so when they observe that “tests of differences cannot reject the null

hypothesis that the explanatory power of models incorporating the alternatives is equivalent." (p. 155)

6. Confidence Interval Analysis

The most glaring descriptive deficiency in the null outcome interpretations we document is the absence of CI (or any other sort of interval or uncertainty assessment) analysis. In this section we explore both the mechanics of implementing relevant CI analysis and how such analysis is likely to impact the presentation and discussion of null outcomes. The analysis draws from an extensive compilation of CI determinations for the null outcomes examined in the prior sections of our analysis. This compilation is provided in Appendix C. Overall, it reveals that while most articles report sufficient information for an interested critical reader to self-generate relevant CIs, doing so often requires substantive effort and thought.[24] The resulting CIs, based on our subjective assessments, are also only occasionally narrow enough to be taken as falling within plausible (in our judgement) MPSD thick null hypothesis intervals. That is, the evidence provided by the typical null outcome is typically incompatible with broad assertions regarding the inconsequentiality of the underlying effect size.

6.1 A COMPREHENSIVE ILLUSTRATION

Almost all the null outcome reporting articles identified by our analysis report a mixture of null outcomes and null rejections. Two of them, however, only report null outcomes. One of these is Kim and Klein (2107), which we discuss in some depth in prior sections. The other is Lennox (2016), which tests the null hypothesis that "There is no change in audit quality after

[24] Six of the articles do not provide, in our reading of them, sufficient information for generating relevant confidence intervals.

companies reduce their APTS purchases following the new rules" using three measures of audit quality—(1) accounting misstatements; (2) tax-related misstatements; and, (3) going concern audit opinions. Its table 6 presents the main results for the independent variable TREATxPOST for: (1) a full sample; and, (2) a propensity matched sample. While the table presents estimate effect magnitudes, it does not provide associated standard errors. Consequently, we infer them from the reported t-statistic and estimated effect values (i.e., s.e. = Effect/*t*). However, in two of the six analyses, the tax misstatements full sample and the going concern matched sample, the reported effect and t magnitudes are near 0 (-.01 to -.02). In such cases the non-reporting of significant digits means inferring the magnitude of the associated standard errors from them is highly problematic. Hence, we drop these two cases from further consideration here.

Lennox employs logistic regressions, meaning that effect magnitudes are appropriately evaluated in the form of likelihood ratios. We follow a two-step procedure to obtain CIs in likelihood ratio form. First, we obtain two standard error CIs around the actual estimated effects by first multiplying each estimate's implied standard error by 2 and then adding and subtracting this value from it.[25] This procedure yields the following initial CIs, expressed in terms of lower (LB) and upper (UB) bounds around the four feasibly evaluated effect magnitudes:

(1) Full Sample Misstatements	LB = -0.276, UB = 0.196
(2) Matched Sample Misstatements	LB = -0.121, UB = 0.541
(3) Matched Full Sample Tax Misstatements	LB = -0.371, UB = 0.951
(4) Full Sample Going Concern Opinion	LB = -0.059 UB = 0.639

[25] We employ two standard error CIs 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 CI) used in hurricane forecasting employ approximately one standard error CIs for forecasts of "center path of the storm" tracks.

Second, we convert these into likelihood ratios by means of exponential transformations of the form e^B-1, where B is the value of the bound being transformed. This yields ratio bounds of:

(1) Full Sample Misstatements LB = -24.1%, UB = 21.65%

(2) Matched Sample Misstatements LB = -11.4%, UB = 71.77%

(3) Matched Full Sample Tax Misstatements LB = -31.0%, UB = 158.8%

(4) Full Sample Going Concern Opinion LB = -5.73% UB = 89.46% .

All four of the CIs defined by these bounds span 0, compatible with the underlying impact of the rule change truly being 0. However, none of the CIs is particularly tight. The narrowest has a range of nearly 48 percentage points. The largest has a range of nearly 190 percentage points. In our opinion these sorts of CIs fall short of attaining the level of precision required for advancing MPSD grounded assertions regarding the inconsequential impact of the rule change on quality.

While Lennox aggressively interprets his "findings" as favoring the truth of the study's (implied) no or little effect null hypotheses, a CI based interpretation of this evidence might read something like:

> Overall, this evidence suggests that the imposition of tax services restrictions on audit quality remains unclear. While the evidence somewhat favors the possibility that it resulted in (possibly sizable) declines in restatement-based measures of audit quality, we cannot convincingly rule out the conjecture that these restrictions led to modest quality improvements. In contrast, the evidence somewhat favors the conjecture that the rule change increased going concern opinion likelihoods, reflecting audit quality improvement. However, we cannot rule out conjectures that these likelihoods declined slightly, which would reflect a deterioration in audit quality.

This alternative discussion of the Lennox evidence illustrates how engaging rather than suppressing CIs and standard error analysis constrains article tendencies to slip into misleading reporting and interpretation of null outcomes. This reinterpretation makes clear that while the study presents a solid first empirical assessment of the empirical landscape for a matter of empirical accounting interest, it is not and should not be the final such assessment. The answer to the question

posed in the article's title, "Did the PCAOB's Restrictions on Auditors' Tax Services Improve Audit Quality?" is, at best, only partially answered (e.g., the restrictions did not result in a substantial improvement) based on the examined evidence. A more definitive understanding needs better evidence.

6.2 AN MPSD GROUNDED "SMALL" EFFECT ILLUSTRATION

Appendix C provides CI determinations and analyses for most of the null outcomes identified in the earlier section of our analysis. It also identifies eight instances where an article reports a null outcome that, after determination of appropriate relevant CIs, where the underlying evidence is compatible with a well-grounded MPSD (minimum practical significant distance) grounded claim that the effect of interest is inconsequential or "small." One of the articles so identified is Patatoukas and Thomas (PT, 2016). Here we use their article to illustrate the relevance of CIs for such "is small" settings.

Table 4 of the PT analysis reports null outcomes for the relations between expected return and three different expected earnings constructs. The focal issue is whether an underlying expected earnings risk factor effect ("upward bias") is present and is possibly large enough to explain the puzzling relation between lagged earnings and returns documented in Patatoukas and Thomas (2011). All three estimates are negative, however, consistent with a maximum likelihood-based inference that the relation is not positive, and the proposed conceptualization here requires that the relation be positive. The estimated two standard error CIs for the effects associated with these three metrics are:

Random Walk:	LB = -0.066, UB = 0.008;
Firm Fixed Effect:	LB = -0.026, UB = 0.018;
Industry-Adjusted:	LB = -0.038, UB = 0.012.

The magnitudes of these initial CI values are not readily meaningful, a far from uncommon occurrence in our examinations. Interpreting them requires a conversion or transformation into values that are substantively interpretable in terms of magnitude. In this instance, the effect in question constitutes a proposed explanation for the puzzling positive predictive relation between lagged earnings and returns. Hence, the estimated magnitude of this relation (0.159) is a highly relevant scale here. Specifically, division of the above UB values by .159 yields values reflecting bounds on the highest percentage of the effect that is compatible with the examined evidence. In this case these bounds are 5.0%, 11.3%, and 7.5%. The relevant interpretation is that the examined evidence is highly compatible with the position that an expected earnings risk factor is, at most, a very small player here. So, while the evidence does not establish that the risk factor effect is completely absent (the "no bias" assertion in Figure 1 of PT), it certainly strongly supports the inference that the explanatory saliency of any such effect is minimal.

6.3 "IS POSSIBLY ZERO" AND "IS DIFFERENT"

Null outcome settings commonly involve establishing two connected empirical propositions: (1) that the effect in question is possibly zero; and, (2) that the effect in question differs (typically is smaller in an absolute sense) from some other relevant benchmark effect. In PT, for instance, the implicit objectives are establishing that the proposed risk factor effect: (1) is possibly zero; and, (2) is not large enough to explain the puzzling lagged earnings effect. PT's approach to getting these points across is to advance the notion that the risk factor effect "is zero" and the lagged earnings effect is not zero. And, if one could truly infer that the risk factor effect "is zero" and nothing else then this approach is sensible and effective as the single "is zero" outcome fully addresses both propositions. However, as is clear from the earlier sections of this paper, inferring that an effect "is zero" based on a null outcome and nothing more is fallacy.

Moreover, if the “is zero” assertion is fallacious then, absent the provision of additional analysis, it necessarily follows that further assertions that are conditioned on its truth (i.e., that it is too small to explain the lagged earnings effect) are unfounded absent the provision of further evidence.

In contrast, as is seen in our PT re-examination, a CI presentation also effectively addresses the two propositions challenge, but does so without engaging in fallacious inference. That is, since the provided CIs span zero, the evidence is compatible with the effect possibly being zero. Moreover, the associated percentage-of-effect-explained CIs, at least on the high side, seem to fall well within any plausible MPSD limit. Hence, these CIs efficiently convey the twofold message that the examined evidence is compatible with a possibly zero effect and incompatible with the effect equaling or exceeding a highly relevant reference effect value. Moreover, the CIs in this case convey the additional message that the evidence is broadly compatible with the conjecture thqat the underlying effect size is “small.”

An important qualitative corollary to the use of CIs in addressing such connected proposition settings is that they also reveal instances when the “is possibly zero” proposition holds, but MPSD robustness is dubious. For instance, page 1506 of Lennox (2016) provides a rather challenging, in our judgement at least, discussion of several test of hypothesis analyses nominally addressing the power of his various tests. For the restatement metrics these analyses demonstrate that any audit quality improvements arising from the rule are likely smaller than arguably relevant benchmarks such as estimated pre-rule change restatement likelihood differences between affected and unaffected firms. The case for the effects being thought of as MPSD relevant follows only if these benchmarks are themselves within the MPSD domain. As the relevant benchmark likelihoods in play here involve restatement likelihood reductions of between 24% to 63%, the case for such an assertion seems dubious. In contrast, consider how what Lennox presents using an ill-suited test

of hypothesis structure compares to the CI presentations we propose as an alternative. For each of the three feasible restatement likelihood examinations the associated CI's show: (1) the examined evidence is compatible with the effect possibly being 0; (2) the evidence is compatible with the possible presence of a modest negative effect size; and (3) the effect is likely smaller than the pre-rule change difference as that difference falls outside of the CI. And, these CIs convey this picture of the evidence by simply reporting pairs of numbers.[26]

A second qualitative corollary to the use of CIs in these sorts of connected proposition settings is that reporting them deters articles from making erroneous assertions about differences in effects based on a null outcome for one effect paired with a rejection outcome for the other effect. "The difference between significant and not significant is not itself statistically significant." (title, Gellman and Stern, 2006). The Fredrickson and Zolotoy null outcome analysis is a relevant example of this form of misinterpretation and its consequences. Their table 6 presents separate examinations of whether individual and institutional investors exhibit visibility driven queuing behavior in processing earnings announcements. They find statistically significant evidence of queuing (a reaction to distracting influences) in high individually held firms but obtain a null outcome for high institutionally held firms. They take these paired outcomes as supporting their H2 hypothesis that "The queuing effect will be less pronounced for institutional investors than for individual investors." However, they never directly test this proposition. They simply rely on the fact that one test yields a rejection while the other does not. That is, the inference is conditioned

[26] In the case of going concern opinions the Lennox test of hypothesis approach yields null outcomes. He chooses to interpret these as reflecting a lack of sufficient power to establish a difference. However, the alternative perspective here is that is that the underlying effect is present and, in fact, is sufficiently large enough to offset the pre-existing difference in going concern opinion likelihoods between rule-affected and not-rule-affected firms. Here again the CI presentation accurately captures the relevant empirical evidence: (1) the evidence is compatible with the effect being 0; (2) the effect is compatible with the effect being substantially positive; (3) the evidence is compatible with the effect being as large as the pre-rule change difference.

on a fallacious "is zero" interpretation of the institutional investor null outcome.[27] A CI presentation here would reveal the rather extensive overlap in the confidence intervals for these paired outcomes, descriptively undercutting the implied claim of a significant difference existing between institutional and individual queuing in this setting.[28]

6.4 WIDE CONFIDENCE INTERVAL IDENTIFICATION

In general, powerful research designs yield narrow CIs that, in turn, facilitate MPSD-based null outcome interpretation. When research designs lack power then wide CIs are generally a consequence. Reporting CIs in such settings both reveals the lack of power inherent in the analysis and maps its impact in terms of identifying just how wide a range of underlying effect sizes are compatible with the evidence. Table 5 illustrates this point. It reports five seemingly overly wide CIs from our Appendix C examinations. The nominal appropriateness of these overly wide designations are, in our opinion, self-evident. Hence, we do not elaborate on them here. Presumably, if these CIs had been article-reported rather than constructed (by us) post-publication then article interpretations would be far less likely to assign PC descriptions to them. Such reporting would also incentivize articles to provide additional interpretive context that speaks to the practical relevance of such highly imprecise evidence. That is, taking Laurion et al. as an

[27] The two standard error CI for the institutional investor estimate of 0.61 is: LB = -2.60 and UB = +3.82. The estimated individual investor effect is -2.63 (a negative sign is consistent with distraction driven queuing) and associated two standard error CI is: LB=4.06, UB = -1.40. Hence, nearly 50% of the individual investor CI overlaps the institutional investor CI, which is incompatible with the notion that the evidence here supports the presence of a statistically significant difference in effect sizes between the two investor types.

[28] The research design framework employed by Fredrickson and Zolotoy is also of some relevance to the general theme of our study in that it requires obtaining zero effect inferences 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 using null outcomes to advance fallacious "is zero" inferences.

example, a study might lay out the case for the policy relevance of thinking that that restatement likelihoods probably increased by no more than 157% after a partner rotation.

7. Conclusion

A well-known educational illustration addressing null outcome interpretation (Jawlik, 2016) shows a man offering an engagement ring to a statistician who answers him by stating "I fail to reject the null hypothesis." The man, of course, expresses confusion about the meaning of this response as in his mind there are two possible answers to his implied question, yes or no. The subsequent panels of the illustration, however, point out that the statistician has indeed truly addressed the underlying issue. Prior to offering the ring the man was, presumably, uncertain about the status of the relationship. Otherwise, why ask? The statistician's answer is saying that this state of affairs has not changed. The uncertainty that prompted the question remains unresolved. He has no basis for taking the statistician's answer as a rejection of the null hypothesis that the definitive answer that will eventually emerge will be yes, or as a rejection of the null hypothesis that this answer will be no.

Extending this illustration to our evidence on null outcome interpretation as currently practiced by the academic accounting discipline suggests a setting where the man refuses to take uncertainty as an acceptable answer. Instead of viewing the STATA, SPSS, SAS, etc. "unable to reject" answer as indicating that the evaluated evidence has not resolved the uncertainty that prompted the examination, the literature resorts to interpretive alchemy.[29] It uses the assumed null to transmute evidence indicating material uncertainty about what is truly going on into certain

[29] Gelman (2016) notes that "Statistics is often sold as a sort of alchemy that transmutes randomness into certainty.."

claims that the null hypothesis is "true" or "supported." By doing so, it replaces evidence-based knowledge acquisition with knowledge acquisition by assumption.[30] I

Though null outcomes are not of themselves a suitable basis for making definitive assertions about null hypotheses, they are amenable to descriptive understanding. Substantive descriptive analysis can, through clear identification of the set of evidence-compatible alternative hypotheses or effect values, facilitate judgements as to as to whether these values are substantively indistinguishable from the relevant null hypothesis. It can also identify settings where the empirical evidence is providing little useful insight about what is truly going on. We, however, find little evidence that articles in the accounting literature are pursuing such a descriptive path 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 the text rarely mentions, to say nothing of discusses in a meaningful way. It then aggressively advances misleading claims about how such an outcome demonstrates the truth of the associated null hypothesis.

Our analysis is also of some relevance to the ongoing debate about the degree to which the accounting academic literature exhibits a "bias" against publishing papers drawing null inferences based on null outcomes (Lindsay, 1994; Bamber, Christensen, and Gaver, 2000; Dyckman and Zeff, 2014). This "bias," however, is inherent to the conventional hypothesis testing structure. And, without it the entire structure falls apart. The answer to this "bias," in our opinion, is not to 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. We should approach

[30] Which is not to say that knowledge cannot be reasonably acquired by assumption (e.g., analytical modeling). But empirical based knowledge acquisition, by definition, means that there is a substantive empirical component to this acquisition process.

empirical analysis from the perspective of "where we are led by the data to believe the finding of interest is to be found." (Dyckman and Zeff, 2015) Indeed, in our opinion 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. And, in a publication environment where "tension" is the most critical ingredient, that is a rather severe handicap to operate under.

We also advance the notion of reporting and examining confidence intervals as an initial step toward providing rigor to null outcome analysis. A confidence interval provides a comprehensive picture of the range of effect values that are compatible with a study's evidence. When thoughtfully implemented, it does so using a measurement scheme that possesses magnitude of effect interpretability. Confidence interval reporting also disciplines articles to interpret null outcomes in representationally faithful fashions. It is difficult to conceive of an article advancing a null hypothesis as "truth" when the subsequent discussion substantively engages a range of evidence-compatible values that materially contradict such a claim.

Finally, on a more prescriptive level, we think the literature would be far better served if authors, readers, and listeners when writing, reading, or hearing statements such as "no relation", "no evidence of", "no change", "no difference", "insignificant", "similar," etc., would make it a practice to ask the question—"Is there a confidence interval for that?" Because absent the descriptive context provided by such intervals (or some other similar form of supporting descriptive examination), these sorts of assertions likely lack substantive support in the presented empirical evidence.

.

References

ABERSON, C. “Interpreting Null results: Improving Presentation and Conclusions with Confidence Intervals.” *Journal of Articles in Support of the Null Hypothesis*, 1 (2002): 36-42.

ALTMAN, D.G., AND J.M. Bland. “Absence of Evidence is not Evidence of Absence.” *British Medical Journal*, 311 (1995): 485.

AMERICAN ACCOUN?TING ASSOCIATION. “Longstanding Mandate on Corporate Audit Committees Yields No Benefit for Investors, New Research Finds.” AAA press release (2017): (http://aaahq.org/Outreach/Newsroom/Press-Releases/11-1-17-)

AMERICAN PSYCHOLOGIOCAL ASSOCIATION. *Publication Manual of the American Psychological Association 6th Edition* (2013).

AMRHEIN, A., D. TRAFIMOW, AND S. GREENLAND. “Inferential vs. Descriptive Statistics: There is No Replication Crisis if We Don’t Expect Replication.” *The American Statistician*, 73.sup1 (2019): 262-270.

BAMBER, L.S., T.E. CHRISTENSEN, AND K.M. GAVER. “Do We Really “Know” What We Think We Know? A Case Study of Seminal Research and its Subsequent Overgeneralization.”*Accounting, Organizations and Society*, 25 (2000): 103–129.

BASU, S. “Is There a Scientific Basis for Accounting? Implications for Practice, Research, and Education,” *Journal of International Accounting Research*, 14 (2015): 235-265.

BASU, S., AND H.-U. PARK. “Publication Bias in Recent Empirical Accounting Research.” Unpublished working paper (2014). https://ssrn.com/abstract=2379889.

BETENSKY, R. “The *p*-Value Requires Context, Not a Threshold.” *The American Statistician*, 73.sup1 (2019): 115-117.

BILLS, K.L., L.L.LISIC, AND T.A. SEIDEL. “Do CEO Succession and Succession Planning Affect Stakeholders’ Perceptions of Financial Reporting Risk? Evidence From Audit Fees.” *The Accounting Review*, 92 (2017): 27-52.

BLOOMFIELD, R., K. RENNEKAMP, AND B. STEENHOVEN. “No System is Perfect: Understanding How Registration-Based Editorial Processes Affect Reproducibility and Investment in Research Quality.” *Journal of Accounting Research*, 56 (2018): 313-362.

BLUME, J., R. GREEVY JR., V. WELTY, J. SMITH, AND W. DUPONT. “An Introduction to Second-Generation *p*-Values.” *The American Statistician*, 73.sup 1 (2019): 157-167.

BLUME, J., L. MCGOWAN, W. DUPONT, AND R. GREEVY JR. "Second-Generation *p*-Values: Improved Rigor, Reproducability, & Transparency in Statistical Analyses." *PLoS ONE*, 13 (2018): e0188299. https://doi.org/10.1371/journal.pone.018299.

BRASEL, K. M. DOXEY, J. GRENIER, and A. REDFFETT. "Risk Disclosure Preceding Negative Outcomes: The Effects of Reporting Critical Audit Matters on Judgments of Auditor Liability. *The Accounting Review*, 91 (2016): 1345-1362.

BRAZEL, J. F., S.B. JACKSON, T.J. SCHAEFER, AND B.W. STEWART. "The Outcome Effect and Professional Skepticism. *The Accounting Review*, 91(2016), 1577-1599.

CANNON, N.H., AND J.C. BEDARD. Auditing Challenging Fair Value Measurements: Evidence from the Field. *The Accounting Review*, 92(2017), 81-114.

CASAS-ARCE, F.A. MARTINEZ-JEREZ, AND V. NARAYANAN. (2017)." The Impact of Forward-Looking Metrics on Employee Decision-Making: The Case of Consumer Lifetime Value. *The Accounting Review,* 92.3 (2017): 31-56.

CHEN, K. C., Q. CHENG, Y.C. LIN, AND X. XIAO. "Financial Reporting Quality of Chinese Reverse Merger Firms: The Reverse Merger Effect or the Weak Country Effect?" *The Accounting Review*, 91(2016): 1363-1390.

CHOI, J., A.H. NEWMAN, AND I.D. TAFKOV. "A Marathon, a Series of Sprints, or Both? Tournament Horizon and Dynamic Task Complexity in Multi-Period Settings." *The Accounting Review*, 91(2016), 1391-1410.

COHEN, A. "A Power Primer." *Psychological Bulletin,* 112 (1992): 155-159.

CUMMING, G. "The New Statistics: Why and How." *Psychological Science*, 25(2014): 7-29.

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

DEFOND, M., C. LIM, and Y. ZANG. "Client Conservatism and Auditor-Client Contracting." *The Accounting Review*, 91(2016): 69-98.

DRAKE, K., N. GOLDMAN, AND S. LUSCH. "Do Income Tax-Related Deficiencies in Publicly Disclosed PCAOB Part II Reports Influence Audit Client Reporting of Income Tax Accounts?" *The Accounting Review*, 91(2016): 1411-1439.

DUTTA, S., AND P.N. PATATOUKAS. "Identifying Conditional Conservatism in Financial Accounting Data: Theory and Evidence." *The Accounting Review,* 92.4 (2017): 191-216.

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

DYCKMAN, T.R., AND S.A. ZEFF. "Some Methodological Deficiencies in Empirical Research Articles in Accounting." *Accounting Horizons*, 28(2014): 695-712.

DYCKMAN, T.R. AND S.A. ZEFF. "Accounting Research: Past, Present, and Future." *Abacus*, 51(2015): 511-524.

ERICKSON, D., M. HEWITT, AND L.MAINES. “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 (2017): 137-154.

FARRELL, A.M., J.H. GRENIER, and J. LEIBY. “Scoundrels or Stars? Theory and Evidence on the Quality of Workers in Online Labor Markets.” *The Accounting Review*, 92.1 (2017): 93-114.

FRANCIS, B.B., D.M. HUNTER, M.N. ROBINSON, AND X. YUAN. “Auditor Changes and the Cost of Bank Debt.” *The Accounting Review*, 92.3 (2017), 155-184.

FREDRICKSON, J.R., AND L. ZOLOTOY. “Competing Earnings Announcements: Which Announcement Do Investors Process First?” *The Accounting Review*, 91(2016), 441-462.

GELMAN, A. “The Problems with P-Values Are Not Just with P-Values.” *The American Statistician*, supplemental materials to *ASA Statement on P- Values and Statistical Significance*, 70 (2016).

GELMAN, A., AND H. STERN. “The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant.” *The American Statistician*, 60(2006), 328-331.

GONG, Q., O.Z. Li, Y. LIN, AND L.Wu. “On the Benefits of Audit Market Consolidation: Evidence from Merged Audit Firms.” *The Accounting Review*, 91(2016), 463-488.

GOODMAN, W., S. SPRUILL, AND E. KOMAROFF. “A Proposed Hybrid Effect Size Plus P-Value Criterion: Empirical Evidence Supporting Its Use.” *The American Statistician* 73.sup 1 (2019): 168-185.

GREENLAND, S. “The Unconditional Information in *P*-Values, and its Refutational Interpretation Via *S*-Values.” (2018) Working paper.

GREENLAND, S. “Valid *P*-Values Behave Exactly as They Should: Some Misleading Criticisms of *P*-Values and Their Resolution with *S*-Values.” *The American Statistician* 73.sup 1 (2019): 106-114.

GREENLAND, S. S.J. SENN, K.J. ROTHMAN, J.B. CARLIN, C. POOLE, S.N. GOODMAN, AND Z. ALTMAN. “Statistical Tests, *P* values, Confidence Intervals, and Power: A Guide to Misinterpretations.” *European Journal of Epidemiology* 31 (2016): 337-350.

GUENTHER, D.A., S.R. MATSUNAGA, AND B.M. WILLIAMS. “Is Tax Avoidance Related to Firm Risk? *The Accounting Review*, 92.1 (2017), 115-136.

HALL, C. M. “Does Ownership Structure Affect Labor Decisions?” *The Accounting Review*, 91(2016), 1671-1696.

HARVEY, C. R. “Presidential Address: The Scientific Outlook in Financial Economics. *The Journal of Finance*, 72(2017), 1399-1440.

HENRY, H., AND A. LEONE. Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone. *The Accounting Review,* 91(2016): 153-178.

HUMPHREYS, K. M. GARY, and K. TROTMAN. Dynamic Decision Making Using the Balanced Scorecard Framework. *The Accounting Review,* 91(2016): 1441-1465.

JAWLIK, A.J. *Statistics From A to Z: Confusing Concepts Clairifed.* John Wiley and Sons. (2016) ISBN 9781119272038

KELLY, K., A. PRESSLEE, AND R.A. WEBB. "The Effects of Tangible Rewards Versus Cash Rewards in Consecutive Sales Tournaments: A Field Experiment." *The Accounting Review,* 92.6 (2017): 165-185.

KHAN, M., G. SERAFEIM, AND A.YOON. "Corporate Sustainability: First Evidence on Materiality." *The Accounting Review*, 91(2016): 1697-1724.

KIM, J., AND P. JI. Significance Testing in Empirical Finance: A Critical Review and Assessment. *Journal of Empirical Finance,* 34 (2015): 1-14.

KIM, J., P. JI, AND K. Ahmed, K. Significance Testing in Accounting Research: A Critical Evaluation Based on Evidence. *Abacus* 54 (2018): 524-546.

KIM, S., AND A. KLEIN. "Did the 1999 NYSE and NASDAQ Listing Standard Changes on Audit Committee Composition Benefit Investors?" *The Accounting Review,* 92.6 (2017): 187-212.

KRISHNAN, J., J. KRISHNAN, AND H. SONG. "PCAOB International Inspections and Audit Quality." *The Accounting Review*, 92.5 (2017): 143-166.

LAURION, H.,, A. LAWRENCE, and J.P. RYANS. "US Audit Partner Rotations." *The Accounting Review*, 92.3 (2017): 209-237.

LAWRENCE, A., S. SIRIVIRIYAKUL, AND SLOAN "Who's the Fairest of Them All? Evidence from Closed-end Funds." *The Accounting Review*, 91(2016): 207-227.

LENNOX, C. S. "Did the PCAOB's Restrictions on Auditors' Tax services Improve Audit Quality? *The A"ccounting Review*, 91 (2016): 1493-1512.

LI, L., B. QI., G. TIAN, AND G. ZHANG. "The Contagion Effect of Low-Quality Audits at the Level of Individual Auditors." *The Accounting Review,* 92.1 (2017): 137-163.

LIN, J., AND Y. WANG. "The R&D Premium and Takeover Risk." *The Accounting Review,* 91 (2016): 955-971.

LINDSAY, R. M. "Publication System Biases Associated With the Statistical Testing Paradigm." *Contemporary Accounting Research*, 11 (1994): 33-57.

LOURENCO, S. M. “Monetary Incentives, Feedback, and Recognition—Complements or Substitutes? Evidence from a Field Experiment in a Retail Services Company.” *The Accounting Review*, 91 (2016): 279-297.

MEYER, K., A. WITTELOOSTUIIN, AND S. BEUGELSDIJK “What is a *p*? Reassessing Best Practices for Conducting and Reporting Hypothesis-Testing Research.” *Journal of International Business Studies* 48 (2017): 535-551.

NELSON, M., C. PROELL, AND A. RANDEL. “Team-Oriented Leadership and Auditors' Willingness to Raise Audit Issues.” *The Accounting Review*, 91 (2016): 1781-1805.

NESSA, M. “Repatriation Tax Costs and U.S. Multinational Companies’ Shareholder Payouts.” *The Accounting Review,* 92.4 (2017): 191-216.

NICKERSON, R. S. “Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy.” *Psychological Methods*, 5 (2000), 241-301.

OHLSON, J. A. “Accounting Research and Common Sense.” *Abacus*, 51 (2015): 525-535.

PATATOUKAS, P.N., AND J.K. THOMAS. “More Evidence of Bias in Differential Timeliness Estimates of Conditional Conservatism.” *The Accounting Review* 86 (2011): 1765-1793.

PATATOUKAS, P.N., AND J.K. THOMAS. “Placebo Tests of Conditional Conservatism.” *The Accounting Review*, 91 (2016): 625-648.

ROBINSON, L., B. STOMBERG, AND E. TOWERY. “One Size Does Not Fit All: How the Uniform Rules of FIN 48 Affect the Relevance of Income Tax Accounting.” *The Accounting Review*, 91 (2016): 1195-1217.

ROTHMAN, K.J. (2016) Disengaging from statistical significance. *The American Statistician,* supplemental materials to *ASA Statement on P-Values and Statistical Significance*, 70 (2016).

SCHROEDER, J.H., AND M.L. SHEPARDSON. “Do SOX 404 Control Audits and Management Assessments Improve Overall Internal Control System Quality?” *The Accounting Review*, 91 (2016): 1513-1541.

STONE, D. “The New “Statistics” and Nullifying the Null: Twelve Actions for Improving Quantitative Accounting Research Quality and Integrity.” *Accounting Horizons,* 32 (2018): 105-120.

TOWERY, E.M. “Unintended Consequences of Linking Tax Return Disclosures to Financial Reporting for Income Taxes: Evidence from Schedule UTP.” *The Accounting Review*, 92.1 (2017): 201-226.

WASSERSTEIN, R.L., AND N.A. LAZAR. “The ASA’s Statement on *P*-values: Context, Process, and Purpose.” *The American Statistician*, 70 (2016): 129-133.

WASSERSTEIN, R.L., A.L. SCHRIM, AND N.A. LAZAR. “Moving to a World Beyond “p<.05.” *The American Statistician*, 73.sup 1 (2019): 1-19.

WIECZYNSKA, M. (2016). The “Big” Consequences of IFRS: How and When Does the Adoption of IFRS Benefit Global Accounting Firms?” *The Accounting Review*, 91 (2016): 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)	Managerial	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)[31]	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)	Audit	Archival	3	2	0
Totals			63	43	21
% Articles Non-Zero				82.9%	48.6%

[31] 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[32]	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[33]	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

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

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

		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 Naive 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
Kelly et al.	"We *do not find significant* effects of reward type"	GC
Khan et al.	"firms with good ratings on immaterial sustainability issues *do not significantly outperform* firms with poor ratings on the same issues"	GC

Article	Null Outcome Statement	Type
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
Lin & Wang	1. "*but not to* innovation efficiency"	PC
	2. "*but not the* innovation efficiency premium"	PC
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

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.

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

Table 5

Candidate Overly Large Confidence Interval Null Outcomes

Article	Article Provided PC Interpretation for Null Outcome	Confidence Interval Measure	2 Std. error Conf. Interval	
			LB	UB
Cannon and Bedard	Level 3 is not associated with auditor specialist use. (p. 106)	Impact of presence of Level 3 assets on likelihood that a valuation specialist is used.	-93%	+235%
Chen et al.	"U.S. RM firms do not differ from matched U.S. IPO firms in the likelihood of restatements." (p. 1373)	Likelihood that a U.S, Reverse Merger Firm reports an accounting irregularity relative to the likelihood that a U.S. IPO Firm reports an accounting irregularity.	-76%	+586%
Farrell et al.	"Online workers' honesty does not differ from that of students..." (p. 94)	Percentage point difference (maximum possible is 100) in honesty rates between online subjects and student subjects in the "modified trust contract" setting.	-20	+7
Laurion et al.	"….there is no change in the frequency of restatements following the partner rotation."(p. 210)	Restatement likelihood changes after partner rotations.	-23%	+157%
Lin & Wang	"a firm's innovation efficiency… is not related to its likelihood of becoming a takeover target." (p. 957)	Sensitivity of the relation between innovation efficiency and stock returns to a one standard deviation shift in takeover probability.	-540%	+913%

This table lists illustrative instances where, based on provided information and discussion, imputed confidence intervals around reported null outcomes interpreted with precisely conclusive (PC) language appear to so large as to suggest that the null outcome possesses little descriptive relevance. Appendix C contains detailed information on the generation and evaluation of each of these confidence interval determinations.

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 that convey the inherent uncertainty associated with the null outcome. 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.[34]	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-

[34] 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 compatible 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 hiring 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 CI 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 than the new outsider CEO fee effect, in context it does not seem at all compatible 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 CI 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 CIs 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 is 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 CI lower bound is -2.67 and the upper bound is +1.21. Or, in terms of likelihoods between -93% and +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 CI 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 error CIs 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 CIs 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 CI 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 CI 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 indicate that the examined evidence is incompatible with 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) CIs. 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 compatible 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 as having little economic significance, consistent with the generally conclusive interpretations provided by for it by the authors.
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 error CIs 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 error CIs 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 (workers) 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 participants". 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 error CIs 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 CIs documented here are representative of the sort of CIs 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	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. 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 CIs, 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 error CI 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.
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 error CI is: LB = -0.0124, UB = +0.0044. The accrual 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 compatible with the inference that any sort of accruals quality effect that is present, particularly downside effect, is reliably small.
Guenther et al.	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 error CIs 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

<table>
<tr><td></td><td>percentage of: (1) the mean of future volatility; and, (2) the standard deviation of future volatility.

<table><tr><td></td><td colspan="2">% of Mean Future Volatility</td><td colspan="2">% of S.D. of Future Volatility</td></tr><tr><td></td><td>LB</td><td>UB</td><td>LB</td><td>UB</td></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>
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.

(Our analysis here only produces CIs for a subset of the question-relevant null hypotheses examined in the study. Several of the other tests lead to opposite direction rejections, which represent stronger evidence in favor of the relation between volatility and tax avoidance being non-positive.)</td></tr>
<tr><td>Hall</td><td>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 definitions), 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 error CIs 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 the evidence does not 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</td></tr>
</table>

	(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.).
Henry and Leone	Insufficient information is provided for the insignificant change in R-square with respect to the Li machine learning metric.
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 CI 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 its 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 CI 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	As discussed in the text, Kim and Klein obtain a null outcome when examining the impact of the board composition rule change on market return. The estimated average return effect 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. The standard error for the per event date average is around 0.066, or 6.6 basis points. We 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 error CI for the overall effect has a lower bound of -67.2 basis points and an upper bound of 144 basis points.[35] 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 compatible 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 a large number of affected firms, a viable

	descriptive inference here is that it is unlikely that the net impact of the rule change was extraordinarily positive. Moreover, the examined evidence is most certainly not of itself a descriptively sound basis for claiming (as is done repeatedly in the press release provided for the article, AAA, 11/1/2017) that the market assigned no, or even little, benefit to the rule change. 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 error CIs 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. 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.
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. 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 error CIs 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 CIs.
Laurion et al.	Laurion et al. report the null outcome for the presence of a relation between misstatement likelihood changes 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 error CI 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 CI 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.* The implied standard error for the 0.146 interaction estimate is .2865. The resulting two standard error CI is LB = -.427, UB = +.719. These values imply that a one standard deviation shift in takeover probability is associated with shifts in the magnitude of the relation between innovations and return effects of between -.068 and +.115. Expressed as percentages of the innovation efficiency effect these amounts to -540% and +913%. 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	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 error CI 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 CI by these the midpoint of these two values, 24, yields a CI 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 compatible with the possibility that FEED flips the sign of the MONEY*EXP effect.
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 CI 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 CI 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 CI 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 compatible with assertions that the impact of repatriation costs on repatriation levels is small.
Patatoukas & Thomas	Discussed in Body of Article.
Robinson et al.	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 error of .0058. Consequently, the pertinent two standard error CI 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%. That is, a hypothesis that FIN 48 reduced the SETTLEIND effect that existed prior to its implementation by as much as 85% is compatible with the examined evidence here. So, while the evidence is compatible 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.
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 CI 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 eight articles (**bolded authors)** at least one set of estimated bounds are compatible with the underlying effect being, in our judgement, plausibly thought of as "small" or inconsequential from a minimum practical significant distance perspective. In six articles (*italicized authors*) the information reported in the article was insufficient for the determination of meaningful CIs.