The "Big N" Audit Quality Kerfuffle

July, 2020

by

William M. Cready
Adolf Enthoven Professor of Accountng
Naveen Jindal School of Management
The University of Texas at Dallas
Cready@utdallas.edu

This paper has benefited from comments provided by an anonymous reviewer and workshop participants at the University of North Texas. All overly harsh language used is the fault of the author.

The “Big N” Audit Quality Kerfuffle

In a highly influential analysis, Lawrence, Minutti-Meza, and Zhang (2011), LMZ henceforth, report that statistically significant relations between a firm’s choice of a Big N auditor and three audit quality metrics (discretionary accruals, cost equity capital, and analyst forecast accuracy) turn “insignificant” after application of matching (propensity score and size) designs. LMZ, however, in interpreting these outcomes mistakenly identify the difference between statistically significant and statistically insignificant as significant (Gelman and Stern, 2006). This analysis re-examines the LMZ evidence descriptively. It finds that little descriptive support exists in the LMZ evidence for conclusive assertions regarding the “insignificance” of audit quality proxy level differences between Big N and non-Big N auditors. Nor does its evidence provide a reliable basis for thinking that propensity score matching based assessment of these differences produces substantially closer to zero inferences about them relative to inferences obtained from existent (inclusive of LMZ provided estimates) conventional non-matching design based multiple regression assessments. Indeed, the LMZ evidence is most appropriately interpreted as providing broad robustness support for the insights provided by such models.

The "Big N" Audit Quality Kerfuffle

1. Introduction

In 2016, The American Statistical Association (ASA) issued the "*ASA Statement on Statistical Significance and P-Values*" (Wasserstein and Lazar, 2016), consisting of six principles addressing the conduct and interpretation of tests of statistical significance and P values. The *Statement's* introduction makes the further point that a key motivating factor for its promulgation is that "it (the *p*-value) is commonly misused and misinterpreted."[1] An observation that Cready, Liu, and Wang (2020) confirms broadly holds for articles published in leading accounting research journals. The *ASA Statement* also recommends the use of more descriptive approaches to data analysis that better address the uncertainty associated with statistical estimation. However, as Cready, He, Liu, Shao, Wang, and Zhang (2019) document, the accounting literature rarely provides such descriptive assessments, even in settings that demand them.

This article presents a salient illustration of how such inadequate interpretative engagement coupled with overreliance on tests of statistical significance assessment taints the broader literature and, by extension, impedes production of accounting knowledge. It takes the form of a comprehensive descriptive re-examination of a widely cited study by Lawrence, Minutti-Meza, and Zhang (2011), LMZ henceforth, that addresses the sensitivity of differences in three proxy measures of audit quality (discretionary accruals levels, implied cost of capital, and analyst forecast accuracy) between Big N and non-Big N auditors to alternative research design specifications. LMZ "find" that favorable relations between the use of a Big N auditor and several audit quality outcome proxies lack statistical significance

[1] That is, the *Statement's* objective is to clarify how to appropriately employ significance testing given that such testing is adopted as a basis for conducting an analysis, as opposed to addressing the far more commonly encountered debate regarding the merits, or lack thereof, of significance testing as a basis for advancing knowledge.

in matched sample designs where client size is a key matching characteristic. LMZ, as well as the subsequent literature that references it, interpret this "finding" as demonstrating the "insignificance" of these Big N effects. Moreover, because such effects do possess statistical significance in complementary multiple regression designs that include client size and other client characteristic variables LMZ further suggest that this "insignificance" outcome demonstrates the inadequacies of such designs. Specifically, that they produce severely overstated estimates of Big N quality effects due to inadequate control of client characteristics.

The problematic nature of LMZ's interpretations of its evidence stem directly from its inferring the general insignificance or absence of estimated effects from statistically "insignificant" outcomes and nothing else. "Statistical insignificance" is not a *per se* reliable basis for thinking an underlying effect is unimportant (i.e., insignificant in a general sense) any more than a statistically significant outcome is a basis for thinking that an underlying effect is important (Wasserstein, Schrim and Lazar, 2019). In contrast, descriptive assessment targeted at locating sets of evidence-compatible parameter values coupled with substantive engagement with the notion of how large such values need to be in order to be taken as meaningful can provide such insights.[2] Such assessment can support inferences that the examined evidence: (1) is exclusively compatible with meaningful parameter value conjectures; (2) is exclusively compatible with inconsequential parameter value conjectures; or, (3) is compatible with both meaningful and inconsequential parameter value conjectures (i.e., the examination has not reliably resolved the existing uncertainty regarding the importance of the studied relation).

This re-analysis of the LMZ evidence demonstrates that there is little descriptive support for taking the LMZ evidence as a well-founded basis for assertions that the Big N effects for its studied quality proxies are inconsequential or insignificant. Nor is there much descriptive support for claims that

[2] Dyckman and Zeff (2014) and Cready et al. (2019) find that inadequate, bordering on non-existent, descriptive engagement with evidence is widespread in the accounting literature.

LMZ's matching designs identify substantially smaller Big N effects relative to those collectively identified in relevant multiple regression examinations. Confidence intervals for LMZ matching design effect estimates generally encompass arguably meaningful effect values. Indeed, they commonly identify values that exceed companion "significant" estimates produced by LMZ's non-matching designs. It is also the case that LMZ matching design Big N effect confidence intervals commonly completely overlap companion non-matching design effect intervals. Hence, a more appropriate view of the LMZ evidence is that it provides mostly confirmatory robustness support for the Big N effect insights produced by appropriately specified non-matching based multiple variable regression designs. Not that it somehow presents a compelling basis for thinking that existing multiple variable regression based evidence of Big N quality differences are materially attributable to client characteristics.

2. The LMZ Analysis

The LMZ analysis examines differences in three audit-quality proxies between Big N and non-Big N audit clients. The three studied proxies are: (1) absolute unexplained discretionary accruals (ADA) where discretionary accruals are measured on a performance matched basis following Kothari, Leone, and Wasley (2005); (2) implied cost of capital levels (RPEG) based on Easton (2004); and, (3) analyst forecast accuracy levels (ACCY) as developed in Lang and Lundholm (1996). A number of prior (to LMZ) studies, including Becker, DeFond, Jiambalvo, and Subramanyam (1998), Francis and Krishnan (1999) and Butler, Leone, and Willenborg (2004) assess audit quality by measuring discretionary accrual levels. However, none of them employ performance-matching in measuring them.[3] Hence, LMZ's ADA measure is arguably in and of itself a material design innovation relative to the existing literature. RPEG

[3] The merits of using performance matching in this setting are actually questionable since, by construction, it involves differencing the unexplained accruals of two firms while only controlling for the client characteristics of one of them. Indeed, firm pairs are commonly going to be audited by the same auditor type, in which case differencing them arguably removes the Big N effect of interest from the ADA measure.

and ACCY, on the other hand each correspond to measures employed in assessing audit quality. Khurana and Raman (2004) introduce the RPEG to assess cross-country differences in Big N audit quality while Behn, Choi, and Kang (2008) introduce using ACCY as a way of linking audit quality with analyst forecasting properties.

LMZ examine Big N versus non-Big N differences in these three metrics: (1) unconditionally (in a univariate regression structure) on the full sample of available observations; (2) using multiple variable regression models that incorporate client characteristics on the full sample of available observations (henceforth identified as FULL models); (3) various matching designs, the most prominent of which use propensity score based criterion as a basis for restricting the set of examined observations (henceforth identified as PSM models). The LMZ analysis proceeds by documenting statistically significant differences in the unconditional and FULL regression assessments, and a lack of such statistical significance for difference estimates produced from design that, in particular, match observations based on client size. Outcomes from these matching design estimations informs LMZ's claim that "we find that the effects of Big 4 auditors are insignificantly different from those of non-Big 4 auditors". They also, when coupled with the presence of Big N "statistical significance" in non-matching designs, are foundational to LMZ's broader assertion that its evidence "suggest(s) that differences in these proxies between Big 4 and non-Big 4 auditors largely reflect client characteristics."

LMZ's analysis and associated interpretation, as described above, suffers from two serious deficiencies that severely compromise its ability to provide evidence relevant to its titled objective of addressing the question: "Can Big 4 versus non-Big 4 differences in audit-quality proxies be attributed to client characteristics?"[4] First, it advances supposedly evidence-based assertions about differences in cross-design Big N effect estimates without ever taking into account what these differences are. That is,

[4] As the title poses a question, a reasonable presumption is that what follows endeavors to answer it, preferably in a convincing manner.

while LMZ clearly argue (per title) that they are presenting evidence supportive of substantial Big N effect estimate shrinkages that are attributable to superior matching design achieved control of client characteristics, nowhere does the analysis directly report or discuss their magnitudes. A far more on point design would focus directly on just how big these shrinkages are. Evidence of substantial shrinkages indicates that yes, client characteristics play an important arguably underappreciated role in explaining observed Big N audit quality associations. Evidence broadly compatible with the absence of any such substantial decline, on the other hand, suggests that there is not a reliable basis in the examined evidence for thinking that inferior control of client characteristics is a salient issue in interpreting existent Big N quality effect evidence.[5]

LMZ's second deficiency concerns its failure to adhere to foundational principles governing the appropriate interpretation of statistical significance and *p*-values as promulgated in the "*ASA Statement on Statistical Significance and P-Values*" (Wasserstein and Lazar, 2016). LMZ's inferential structure relies on the presence or absence of "statistical significance" as a basis for inferring whether underlying studied effects of interest are important or unimportant. However, as the *ASA Statement* makes quite clear, a small *p*-value does not imply that an underlying effect is large or meaningful (Principles 5 and 6 of *The ASA Statement*). Nor does a large *p*-value (i.e., non-significance) imply that the underlying effect of interest is absent or small (Principles 2, 5, and 6 of the *ASA Statement*). Consequently, there is nothing in LMZ's inferential structure that precludes it from routinely assigning "insignificant" assessments when underlying effects of interest are truly meaningful or, for that matter, from assigning "significant" assessments when underlying effects are of trivial magnitude. Hence, the fact that a non-matching design produced estimate is statistically significant while a companion matching design produced estimate lacks

[5] It is important to point out that a third outcome is also possible—the evidence is ambiguous. It is compatible with the difference being either consequential or inconsequential. In fact, as shall become apparent in the analyses that follow, the LMZ evidence largely falls into this third "no definitive conclusion possible land."

statistical significance, the lynchpin of LMZ's interpretive structure, does not say much of anything at all about the nature of the underlying difference between them.[6]

Despite these aforementioned deficiencies, the audit quality literature commonly references LMZ as an authoritative basis for questioning the presence of (meaningful) differences in quality between Big N and non-Big N auditors. DeFond, Erkens, and Zhang (2017a) indicate that the broader literatures interpretation of LMZ is that it "casts serious doubt on the existence of a Big N effect" and that "the absence of a Big N effect not only overturns a large literature, but also questions our basic understanding of fundamental drivers of audit quality." Articles also commonly identify LMZ as a basis for thinking that observed Big N quality differences are largely attributable to differences in client characteristics. At a more general level, articles outside of the audit quality domain commonly advance LMZ as a compelling illustration of the how propensity score matching (PSM) designs, in particular, produce meaningful consequential estimation improvements relative to those obtained from conventional non-matching based multiple regression methods.

Audit literature studies specifically addressing Big N audit quality also commonly identify LMZ as a basis for motivating further study of the issue. These examinations mostly take the form of evaluating alternative audit quality proxies, identifying settings amenable to clearer measurement of audit quality impacts, and assessing the generalizability of the LMZ "findings" to other populations or settings. Studies typically only question LMZ findings in terms of how generalizable they are to the studied quality measures or setting. They do not, as a rule directly engage with the veracity of LMZ's interpretation of its evidence.

[6] Gelman and Stern (2006) provide a widely cited comprehensive critique of using statistical significance and insignificance as a basis for assessing difference. See Greenland, Senn, Rothman, Carlin, Poole, Goodman, and Altman (2016) for a general presentation of ways by which researchers misinterpret and misapply statistical significance.

DeFond et al. (2017a, 2017b) is a noteworthy exception to the literature's unquestioned acceptance of the fundamental integrity of the LMZ analysis. These two studies directly question the robustness of LMZ's "statistical insignificance" assessments. In replications of LMZ's PSM analyses they show that PSM designs comparable to those chosen by LMZ typically produce statistically significant outcomes. Hence, they essentially argue that LMZ's "insignificance" assessments lack "statistical insignificance" replicability. Such non-replicability certainly casts doubt on the compatibility with the LMZ evidence of conjectures that non-Big N audits are of higher quality, as reflected in LMZ's chosen proxies, than Big N audits. It does not, as is clear from relevant *ASA Statement* principles, shed much insight at all on the far more salient issues of whether LMZ matching designs identify meaningful or trivial quality effects or whether these matching designs identify substantially smaller Big N effects than those identified by full sample multiple regression designs.

It is also of some relevance to point out that LMZ does broadly qualify its Big N "insignificance" assertions. However, the nature of these qualifications pertain to whether Big N insignificance generalizes to other measures, populations, and research designs. LMZ most certainly does not self-identify its chosen inferential structure as being a thoroughly unsuitable approach to assessing design-choice determined differences in Big N effects; or, equivalently, to attributing such differences to client characteristics (Gelman and Stern, 2006; Wasserstein and Lazar, 2016; Wasserstein et al., 2019; Amrhein, Greenland, and McShane; 2019). Moreover, as the analysis that follows demonstrates, most of LMZ interpretations do not hold up to the light of better suited examinations of its evidence.

3. Descriptive Assessments of the LMZ Evidence

LMZ's assertions of substantial attenuation of estimated Big N quality effects in its matching design estimations flow from an exclusive reliance on statistical significance and insignificance

determinations. LMZ never directly identifies the actual Big N effect reductions achieved by matching designs relative to either its own proposed non-matching design estimates or estimates from the then available literature. Consequently, these attenuation assertions exist in something of a descriptive vacuum. They have no grounding whatsoever in the actual Big N effect magnitudes and magnitude changes identified by the relevant empirical evidence. The examination that follows subjects these statistical significance-based assertions from LMZ to substantive descriptive assessment. Specifically, it evaluates the degree to which LMZ's interpretations of its evidence comport with what this evidence is really saying about the importance of client characteristics, particularly those pertaining to client size, in assessments of differences between Big N and non-Big N audit quality proxies.

3.1 Confidence Interval Effect Size Assessments

Table 1 provides descriptive enhancement of the LMZ evidence in the form of 95% confidence intervals (CIs) for the Big N versus non-Big N quality proxy (multiple independent variable regression based) differences it does report together with their partial correlation analogs. CIs are particularly relevant for the conduct of descriptive assessment of evidence as they identify ranges of evidence-compatible values for studied measures. Such ranges explicitly incorporate precision quality of underlying evidence implications for effect magnitudes. Amrhein et al. (2019), among many others, advance CIs as a particularly salient alternative to the reliance on statistical significance in conducting productive assessments of empirical evidence. Partial correlations are commonly employed as scale free measures of effect sizes. Meta analyses, in particular, use them to integrate estimates obtained from collections of studies employing diverse methods, models, and measures.[7] At a very general level the

[7] Following Aloe and Thompson (2013), partial correlations (r_p) are determined from study reported t values as $r_p = t/(t^2 + n)$ where t is the reported t statistic for the estimated Big N effect and n is the associated sample size. (In theory n should be reduced by the number of regression model parameters, but the exact value of this number is unclear in most of the studies we examine and, given the relevant sample sizes involved is of little practical consequence.) The associated variance, required for determining relevant confidence intervals, is similarly calculated as $var(r_p) = (1 - r_p^2)^2/n$.

partial correlation measures the incremental explanatory potency of a variable for a dependent variable of interest benchmarked against the overall unexplained background variation present in the dependent variable. Hence, it is scale free approach to assessing effect size that is amenable to the conduct of cross-study, cross-design, and cross-metric integrations and comparisons of evidence.

Panel A provides this assessment analyses for the LMZ provided FULL design "replications" of existent literature Big N effect value findings. The LMZ FULL designs employ all available usable data and employ multiple independent variables as a means of controlling for differences in client-characteristics between Big N and non-Big N audit clients. In terms of directional significance, these findings are fully consistent with those from the prior literature that LMZ reference. The ADA and RPEG Big N effect estimates, each multiplied by -1 to align directional presentation with higher values corresponding to higher Big N audit quality proxy levels, and associated CI bounds are all positive. The ACCY effect estimate, which does not require directional transformation, and its associated bounds are likewise positive.

The practical implications of these estimates as demonstrating the consequentiality of the Big N auditor choice vary by proxy. The ADA estimates provide the greatest support for such consequentiality. The estimated partial correlation associated of 3.15% is over double the size of the companion RPEG and LMZ correlations. Its partial correlation lower bound value of 2.42% is well clear of zero, unlike the vanishingly small 0.11% and 0.23% bounds obtained for the RPEG and ACCY measures. The reported point estimate of 0.0179 indicates that choosing a Big N auditor reduces unexplained absolute accruals by 1.79% of assets, a value that is considerably larger than the .3% to .5% materiality threshold advanced by DeFond, Erkins, and Zhang (2017b) as a meaningful significance threshold in this setting. Indeed, the identified lower bound of 1.38% is similarly well clear of this materiality threshold.

The RPEG measure has a natural economically meaningful basis point interpretation. The 37 basis point panel A Big N difference estimate suggests that a firm with a billion dollar market capitalization might expect to reap a $3.7 million dollar payback from a one period switch to a Big N auditor. The 71 basis point upper bound indicates that conjectures that the basis point reduction benefit exceeds $7 million dollars is compatible with the evidence. On the other hand, however, the CI lower bound for this basis point payback is a mere $300,000. Hence, the evidence here is compatible with a rather wide range of basis point effects, some of which are clearly quite meaningful (particularly from a percent of audit fee perspective) and others not so much.

The point estimates for the ACCY measure are 0.0042 and 1.4%. The partial correlation estimate is the more readily interpretable of the two, as it reflects the reduction in forecast error associated with the use of a Big N auditor. Hence, based on the associated CI bounds, the evidence here is compatible with Big N associated forecast error reductions of between 0.23 % and 2.57%. The 0.23% lower bound is arguably indistinguishable from 0. The upper bound of 2.57% is underwhelming, at best. From a behavioral psychology perspective, for instance, such a correlation would be deemed inconsequential. Such correlations are not nearly large enough to be noticed by a careful observer (Cohen, 1992). Hence, they have no practical importance for explaining observed behavior. In this context, the analogous argument is that reducing unexplained forecast variation by less than 3% is noticeable to no one and hence is inconsequential. Of course, if one accepts this perspective of the inconsequential ACCY effect sizes based on LMZ's FULL design evidence, then the only thing the subsequent PSM assessment can achieve is its confirmation. Alternatively, since partial correlations are scale free metrics one can evaluate the ACCY correlations based on the sorts of partial correlation magnitudes that are implicitly identified as meaningful in the ADA and RPEG settings. For instance, the overlap between the ACCY and RPEG partial effect CIs is extensive. As the actual cost of capital benefits associated with most of

the points found within the RPEG interval appear meaningful, such overlap suggests that most of the points within the ACCY CIs may be similarly meaningful.

Panel B shifts the focus from the audit quality proxy estimates reported in LMZ to estimates from the existing literature referenced by LMZ. LMZ's own descriptive engagement with this prior literature is quite limited. It draws on these studies as evidencing the presence of "statistically significant" evidence of favorable Big N quality impacts for the proxies it examines. While it also cites this literature in its identification of relevant control variables, nowhere does it engage with what is reported in terms of Big N effect magnitudes. Nor is there any engagement with the underlying estimation precisions associated with them. The specific estimates evaluated here are: ADA_{BLW}-- Butler, Leone, and Willenborg (2004), BLW henceforth the estimated Big N effect on absolute discretionary accrual levels from table 5, equation 2b, of Bulter, Leone, and Willenborg LW; $RPEG_{KR}$--the estimated Big N effect on implied cost of capital from table 3, U.S. model, KR; and $ACCY_{BCK}$, the estimated Big N effect on analyst forecast accuracy from table 4, model 5, of BCK.[8]

The most striking feature of the relevant literature estimates is the complete absence of correspondence between the ADA and ACCY CIs and those obtained by LMZ's "replications" of these analyses. The CI for the BLW ADA difference estimate is 0.002, with an associated CI of 0.0016 to 0.0024. This CI identifies a substantially smaller, less consequential, set of ADA difference estimates than is identified by LMZ's comparable FULL design. Possible sources of this distinct downward shift

[8] Ideally, the chosen models here would exactly correspond to the multiple regression specifications employed in LMZ. However, there are no such instances. The LMZ specifications, while similar to those found in the existing literature, are each unique to LMZ. In selecting models from the literature I do, however, rely on LMZ's own identifications of source models for the models it employs. LMZ only identify a single prior literature cost of capital (KR) and a single prior literature forecast accuracy (BCK). Hence, I use estimates from these two studies here. For the discretionary accrual proxy multiple studies are identified, but only the BLM analysis employs absolute discretionary accruals, the LMZ measure, in a multiple regression specification. In choosing among alternative Big N effect estimates provided by various design variations within each study I considered only estimates from specifications where the Big N effect appears only once (i.e., it is not interacted with other variables). I then narrowed the selection to a single estimate based on number of observations employed and number of included control variables (more being preferred to less in both cases).

in estimated sets of evidence compatible differences, that is duplicated in the partial correlation CIs as well are: (1) LMZ use absolute performance matched discretionary accruals while BLM simply use discretionary accruals; (2) BLM include the square of ROA in addition to ROA and book/market as additional control variables; and, (3) LMZ include industry and year fixed effects. Irrespective of source, the clear message here is that the degree to which one can take the BLM evidence as supportive of the notion that the existing literature is identifying meaningful Big N quality differences based on multi-independent variable non-matching designs is far from clear. In fact, Lawrence, Minutti-Meza, and Zhang (2017), in their response to DeFond et al.'s (2017a) statistical significance criticism of their analysis, implicitly adopt the perspective that ADA differences in the neighborhood of .2% of assets lack economic significance.[9]

Another notable feature of the BLW ADA analysis is that while the quality difference estimate CIs shift toward zero the partial correlation estimate CIs shift away from zero. The partial correlation CI upper bound is 7.03% while the lower bound is 4.66%. The reason for this divergence is that the partial correlation reflects explained variation relative to background variation. LMZ measures ADA as the absolute value of the difference between a firm's discretionary accruals and the discretionary accruals of a matched firm. The matched firm substantial extraneous variations into the dependent variable. *Ceteris paribus*, such extra variation mechanically reduces the partial correlation, ceteris paribus. Similarly, BLW do not employ cluster-adjusted standard errors. Hence, the level of extraneous variation is understated relative to that which is likely imputed under cluster-adjusting. Such understatements ignore relevant extraneous variation, the incorporation of which would reduce partial correlation estimates.

[9] The original LMZ article does not provide meaningfulness assessments of any of its effect estimates or, for that matter, cross-method differences in such estimates. Their appearance at the very end of the Lawrence et al. versus DeFond et al. discussions is rather ironic given that effect magnitude assessment is actually fundamental to any sort of meaningful interpretation of the LMZ evidence. In this respect, it is fair to say that LMZ's collective engagement with the issue ends at the point where it should truly have begun.

The BCK ACCY CI assessments of the Big N difference also suggest the presence of more sizable relative forecast accuracy associations than those obtained from the LMZ replication. The difference estimate of .0315 is roughly 8 times the size of LMZ's replication estimate. The partial correlation estimate of 5.85% is roughly 4 times the size of its LMZ corollary. The BCK CIs are 0.0191 to 0.0436 for the difference estimate and 3.39% to 7.46% for the partial correlation. Zero overlap exists between the sets of evidence-compatible values identified by the associated CIs for these estimates and their panel A LMZ CI corollaries. The LMZ ACCY baseline full sample regression model identifies an entirely distinct set of evidence-compatible effect value conjectures relative to those identified by the BCK ACCY replication exercise. Possible design difference explanations for this location shift include: (1) LMZ employ fixed industry effects while BCK do not; (2) BCK include controls for selection bias and auditor specialization that LMZ do not; (3) BCK trim observations with extremely small ACCY values (less than -1.5) while LMZ Winsorize ACCY observations.[10]

In contrast to the ADA and ACCY non-replication outcomes, KR's RPEG difference estimate of 30 basis points is relatively close to the Panel A 37 basis point estimate. The partial correlation estimate of 2.46% is not that far removed from the 1.35% value reported in panel A. The associated CIs overlap their panel A corollaries by 56% and 68% respectively. Moreover, the 56% overlap of for the Big N difference estimate is deceptive in that the 49 to 11 basis point interval for it is actually located fully within the larger panel A CI. It differs simply because its estimate is more precise than the one obtained from the LMZ replication exercise. Hence, LMZ's FULL design exercise, unlike its ADA and ACCY FULL model exercises, seems accurately characterized as a successful replication of existent literature evidence.

[10]BCK trim observations with extremely small ACCY values (less than -1.5) while LMZ Winsorize ACCY observations. The sizable difference in ACCY standard deviation estimates between the two designs, 0.0437 versus 0.113, is consistent with this differential measurement of ACCY being important. LMZ also identify Winsorizing as a necessary condition for their PSM ACCY "insignificance" outcomes.

Panel C introduces the LMZ propensity score matching estimates into the mix.[11] As the PSM estimates necessarily employ substantially reduced sample sizes, they tend to produce comparatively noisier estimates. Hence, associated CIs tend to be wider than those obtained for the companion estimates reported in panels A and B. Each of the PSM based proxy difference estimates is directionally consistent with its panel A and B counterparts and smaller, in absolute terms, than its non-PSM counterparts in panels A and B. However, the ADA and RPEG PSM partial correlation estimates of 2.08% and 2.86% are actually larger than their LMZ FULL design counterparts. That is, the evidence provides weak support for the Big N variable explaining more of the dependent variable variation in PSM designs than it does in FULL designs. It is also the case that the CI upper bound values for all three measurses indentify plausibly meaningful effect sizes. Specifically, the ADA upper bound of 1.85% clearly exceeds the .3% materiality threshold advanced by DeFond et al. for assessing importance, the meaningfulness of the RPEG upper bound of 62 basis points is self-evident, and the ACCY partial correlation upper bound of 6.51% is contextually large. That said, however, it is also true these same CIs encompass effect size conjectures compatible with Big N auditors being associated with lower rather than higher quality levels.

The other key insight from the panel stems for the two CI overlap columns. The first of these reflects the extent to which each of the reported PSM CIs overlaps its companion panel A estimate's CI. The second reflects the extent to which each such CI overlaps its companion panel B estimate's CI. Little or no overlap supports the possible presence of a meaningful design driven difference in identified Big N effect. Extensive overlap largely precludes the possibility of clear identification of such a difference. In and of itself, however, such overlap does not rule out the possibility that meaningful a difference is

[11] LMZ provide client-size matching design based analyses to supplement their PSM evidence. The outcomes for these analyses is quite similar to those obtained from the more prominent PSM analyses. Hence, in the interests of conciseness, the examinations here only address LMZ's PSM evidence.

truly present. Given this interpretive perspective, the widespread presence of high percentage overlaps for the RPEG and ACCY metrics, half of them hitting the 100% maximum, effectively precludes the possibility of determining whether such a difference is actually present based on the available evidence. Ascertaining whether a case exists for viewing this evidence as broadly compatible with the absence of meaningful underlying differences requires further analysis.

The panel C ADA overlaps, on the other hand, provides some support for the conjecture that PSM designs identify smaller audit quality proxy differences than those identified by multiple variable regression designs. There is no overlap at all in three of the four comparisons. However, the fourth comparison, concerning the extent to which the PSM ADA CI overlaps the BLW ADA CI, indicates an overlap of 100%. The difference between the two Big N effect point estimates is tiny (.02% of client asset value). Moreover, the corresponding non-overlap in the partial correlation CIs here is, as discussed previously, attributable to key design differences that have nothing to do with PSM. These differences reduce the level of extraneous ADA variation in the BLW design relative to the LMZ designs, thereby inflating its partial correlation estimates (but not its point value difference estimates) relative to those produced by the LMZ FULL and LMZ PSM designs. Hence, it is these design differences, not the PSM design choices, that drive this absence of overlap. Therefore, based on this set of evidence as examined thus far, about all that can be clearly inferred is that reliable support exists for the presence of PSM design determined differences in Big N relative to non-big N quality proxy level estimates provided by LMZ's PSM versus FULL designs.

The scale free nature of the various table 1 partial correlation values means that they are amenable to collective cross-metric cross-design evaluation based on the assumption that explanation of dependent variable variation is a relevant basis of comparative understanding. Forest plots are a conventional visual approach to providing such collective effect size insights. Figure 1 presents a forest plot of the nine

partial correlation CIs presented in table 1. Fundamentally, the LMZ argument is that there should be two forests here. One consisting of the three PSM CIs clustered very close to the zero y axis and the other, consisting of the six non-PSM CIs located well to the right of bot zero and the PSM forest. This parting of the woods is not what one sees in this plot. Rather, it seems far more like one big forest with a PSM prevalence on the near-zero axis side. Figure 2 repeats the exercise using 1 standard error CIs. The one forest takeaway remains intact.

Each plot line also identifies associated semi-partial correlation point estimate locations. The semi-partial correlation reflects the correlation between the unconditional variation in the independent variable and the associated incremental explained variation in the dependent variable. Fundamentally, relative to the partial correlation, it penalizes the reported correlation for the degree to which other independent variables in a model explain the independent variable of interest. Semi-partial correlations are always equal to or smaller, in absolute terms, than partial correlations with the degree of reduction reflecting the saliency of these other variables (i.e., "client characteristics") for the studied variable's effect size. In the vast majority of intervals these semi-partial values are quite close to their partial correlation analogs. Hence, the descriptive evidence here does not seem very sensitive to the fact that matching designs by design place less emphasis on independent control variables in addressing dependent variable variation.

3.2. Effect Value Difference Analyses

The prior analysis, following the approach taken by LMZ, evaluate the effect value evidence provided by FULL and PSM designs independently. Apart from CI overlaps, it does not directly assess cross design type differences in effect estimates. One of LMZ's core assertions, however, is that client size oriented matching designs produce substantially smaller estimates than those obtained from non-

matching designs. Tables 2 and 3 address this shrinkage in effect value impact based on direct examination of differences in effect estimates between FULL and PSM designs.

Table 2 reports the differences between the LMZ PSM estimates and the existent literature (panel A) and LMZ (panel B) FULL estimates. Three estimates are provided for each difference: (1) the unscaled estimated difference; (2) the difference as a percentage of the addressed quality measure's estimated standard deviation (from LMZ FULL design sample descriptive statistics);[12] and, (3) the percentage decline (in absolute terms) associated with the PSM estimate relative to the full sample (non-PSM) estimate. The percentage decline measure is of particular relevance to LMZ's conclusions that differences between PSM and FULL design estimates "largely reflect" (abstract) or are "likely attributable to (inadequate control of) client characteristics" (p. 273, parenthetical insertion mine). If one takes the chosen adverbs "largely" and "likely" at face value then these descriptions imply the presence of effect value reduction in excess of 50%. In fact, Lawrence et al. (2017) adopt just such an approach to assessing their evidence, arguing that the DeFond et al. (2017b) replication evidence identifies a "61% to 83% reduction in the economic magnitude of the estimated Big 4 (ADA) effect." Four of the six differences, however, fall short of this 50% threshold. Indeed, in terms of unscaled magnitude, each of these differences is smaller than its associated "insignificant" (per LMZ) PSM estimate.

Table 3 provides 95% CI bounds for the cross-design effect determined reductions in Big N effect values measured from the perspective of the amount by which PSM designs reduce absolute Big N effect estimates toward zero relative to companion FULL design effect estimates. The pseudo standard error values used to generate these CI are calculated as:

$$SE_{DIFF} = (SE_{FULL}^2 + SE\_PSM^2)^{**.5} \qquad (1) \quad .$$

[12] The standard deviation values, per LMZ, are: 0.1398 for ADA, 492bp for RPEG, and 0.0437 for ACCY.

where SE_{FULL} is the SE for a given full (non-matching multiple regression design) model estimate and SE_{PSM} is the SE estimate for the companion LMZ PSM model estimate. The calculation of SE_{DIFF} follows from the well-known expression for the variance of the sum of two independent random variables. It is the same calculation commonly employed for determining SE values in two sample t-test analyses. They are identified as pseudo standard errors because the non-PSM and PSM quality effect estimates employ overlapping data. Consequently, the measurement error in a given non-PSM estimate co-varies with the error in the paired PSM estimate. Such covariation attenuates the impact of measurement error in the difference. The SE estimates produced by (1) assume independence and so ignore this beneficial attenuation in measurement error. Consequently, SE_{DIFF} values overstate expected error levels associated with the effect value difference estimates. However, this overstatement should not be that great since the PSM design sample sizes are far smaller than companion non-PSM sample sizes. That is, the vast majority of observations used to produce non-PSM estimates are independent of the data used to produce companion PSM estimates.

The pseudo standard error values, reported in the first column of table 5, (inversely) reflect the precision level associated with each of the table 2 cross-design effect value reduction estimates. Hence, the evidence nominally favors taking the BLW based reduction of -0.0002 as the more precise of the two ADA reduction assessments, the KR based -9 basis point estimate as the more precise of the two RPEG reduction assessments, and the LMZ based estimate of -0.0011 as the more precise of the two ACCY reduction assessments. The ratio of each estimated cross-design Big N effect estimate reduction to its associated standard error also, of course, constitutes a (conservatively determined) t-statistic. These t-values for the ADA_{LMZ} and $ACCY_{BCK}$ reduction estimates are -5.03 (-0.0161/.0032) and -4.45 (-0.0285/0.0064). Conventionally, such magnitudes qualify these two reductions for “statistical

significance" identification. The t-statistics (not tabulated) for the other four differences, however, are all quite small. Hence, they lack "statistical significance" at conventional levels.

The table 3 ADA_{LMZ} and $ACCY_{BCK}$ CIs also provide robust support for conjectures that LMZ PSM designs consistently identify substantially reduced Big N effects relative to those obtained from these particular FULL designs. The evidence is broadly compatible with conjectures that the LMZ PSM design identified ADA and ACCY effects are each less than half the size of the respective upper bound values for the LMZ FULL design ADA estimate and the BCK ACCY estimate. However, it is important to recognize that these consistently meaningful declines in Big N effect estimates are specific to these two FULL designs. The remaining four decline assessments do not support conjectures that PSM designs consistently identify substantially smaller Big N effects than FULL designs.[13]

The ranges of evidence compatible reductions in estimated Big N effects identified by table 3's remaining four CIs are all broadly compatible with the effect insights provided in table 1. The ADA_{BLW}, $RPEG_{LMZ}$, $RPEG_{KR}$, and $ACCY_{LMZ}$ CIs all identify sizable positive and negative reduction (i.e., increase) conjectures as being compatible with the evidence. Indeed, when scaled by FULL design Big N effect estimates the CI bounds equal or exceed 100% in both directions in all cases. An important high level implication of such extreme CI widths is that the evidence here simply lacks the levels of precision necessary to reliably evaluate whether or not PSM designs identify meaningfully smaller Big N quality effect estimates relative to those identified by FULL designs.

[13]It is also the case that the lower bound for the difference between the BLW and LMZ FULL design identified ADA effects of 0.0118, not tabulated, is actually larger than the tabulated 0.0097 lower bound for the LMZ PSM achieved ADA reduction. Similarly, the lower bound for the difference in the BCK and LMZ FULL design ACCY effects, not tabulated, is 0.0149, which is comparable with the tabulated 0.0159 lower bound for the reduction in effect value for the PSM design. That is, the sorts of effect reductions identified for PSM designs with respect to the LMZ ADA and BCK ACCY are hardly unique to such designs, they are seen in the companion FULL design evidence as well.

3.3. Statistical Power Assessment

Fundamentally, LMZ's interpretation of its evidence employs statistical significance as a divining rod for separating "significant" effects from "insignificant" effects. A key underlying validity criterion for such an approach is a well-founded expectation that the utilized design reliably returns a significant test value when the underlying effect is truly consequential in magnitude. Absent such an expectation, there is no basis for taking "statistically insignificant" outcomes as providing any sort or relevant insight regarding the meaningfulness of whatever effect is actually present.

A critical component of any focused assessment of a given research design's reliability in returning statistically significant outcomes when the tested null hypothesis is (materially) false is identifying what sorts of departures from the null value(s) are large enough to be considered meaningful. Commonly, however, such identifications can be challenging and subjective. In this particular case, however, the LMZ inferential structure provides guidance in that it implicitly identifies FULL design effect value estimates identified in the prior literature and its own analyses as being of meaningful magnitudes. Hence, it is fair to say that LMZ implicitly identify Big N effects as small as a .002 decline in absolute abnormal discretionary accrual levels (BLM's "statistically significant" ADA estimate), a 30 basis point decline in cost of capital (KR's "statistically significant" RPEG estimate) and a 0.0042 increase in forecast accuracy (LMZ's "statistically significant" ACCY estimate) as being of meaningful importance. To view them otherwise implies that the LMZ analysis simply sets out to demonstrate that less powerful PSM designs attenuate the "statistical significance" of already known to be inconsequential effects to the point that they lack statistical significance.[14]

Given this background, table 4 provides likelihoods that underlying Big N effects that equal or exceed available FULL design estimates produce "statistically significant" outcomes under alternative

[14] An outcome that is readily achievable in any setting by simply reducing sample size.

standard error assumptions. These standard errors either directly correspond with, or are sample size adjusted variations of, the design and study specific standard errors reported in table 2. "Statistically significant" likelihoods are determined by first assuming that the true underlying Big N quality effect for a given quality proxy equals a given prior study determined estimate (i.e., meaningful effect magnitude). Under this true value of the underlying parameter assumption, each of the tabulated values reflects the likelihood that a research design providing a specified level of measurement precision (i.e., standard error level) produces a statistically significant (.05 level) outcome in a test of the null hypothesis that the quality impact of interest is completely absent (i.e., it equals zero). The underlying process involves first dividing the assumed underlying effect (*F_E*) by the relevant standard error (SE). This ratio is asymptotically distributed N(*F_E/SE*, 1). The proportion of this distribution that lies above the relevant 1.96 or below the relevant -1.96 threshold corresponds to the likelihood that a test of a conventional zero effect null hypothesis produces a $p<.05$ (two-tailed) rejection outcome.[15]

The table 4 panel A analysis provides likelihoods using LMZ FULL design difference estimates as *F_E* values while the panel B analysis provides them using prior literature Big N difference estimates as *F_E* values. For each effect "statistical significance" attainment likelihoods are provided for three relevant SE values: (1) the identified FULL analysis SE for the estimated Big N effect; (2) this same FULL analysis SE adjusted to reflect the PSM analysis sample size; (3) the LMZ SE for the PSM estimate of the relevant Big N effect.

[15] LMZ report that they conduct bootstrap analyses of randomly determined 50% Big 4 and 50% non-Big4 sub-subsamples with sample sizes set equal to those employed in the PSM designs. They report that "All reported inferences are robust to these alternative explanations." The implication here seems to be that such reduced sample size designs produced "statistically significant" effect estimates with some degree of reliability. The problem with this perspective, however, is that it is quite difficult to reconcile it with the reported FULL sample RPEG and ACCY Big N estimate standard errors. In both cases, the original FULL analysis estimate barely clears the "statistical significance" threshold. It is simply not plausible to think that sub-samples a fraction the size routinely produce outcomes that clear the "statistical significance" threshold absent rather inexplicable shifts from LMZ reported effect estimates and associated estimation precisions.

The FULL SE analyses address the baseline replicability in the form of likelihoods that hypothetical exact replications of each FULL analysis directionally replicate the "statistically significant" verdicts obtained by LMZ. In panel A this likelihood exceeds 99% for ADA, but is only 57% for the RPEG proxy and to 65% for the ACCY proxy. In panel B it exceeds 99% for the ADA and ACCY metrics and is a relatively robust 87% for the RPEG proxy. Hence, hypothetical exact replications of the prior literature analyses all exhibit high likelihoods of returning "statistically significant" verdicts. In terms of the LMZ provided "replications" of these prior literature analyses, only the ADA FULL analysis exhibits a compelling likelihood of a rejection outcome in an exact replication.

One important SE relevant difference between the various FULL and PSM designs is sample size. The FULL designs employ anywhere from 3 to 9 times as many observations as their companion PSM designs. Structurally, under the null hypothesis that matching is an inconsequential design modification the PSM designs differ from the FULL designs only in that they employ smaller sample sizes. Hence, from the perspective of identifying, ex ante, whether a given design possesses enough power to detect underlying effects of interest these sample size reductions are quite relevant. These reductions, in conjunction with the FULL analysis outcomes, are factors LMZ could have utilized to conduct an a priori assessment of the power sufficiency of their PSM designs. In this regard, the 10% and 11% likelihood values for the $RPEG_{LMZ}$ and $ACCY_{LMZ}$ designs indicate that they clearly do not provide, in expectation, the necessary power to reliably identify underlying meaningful effects in the data in the sorts of sample sizes employed in LMZ's associated PSM designs.

The final set of percentages reported in table 4 provide an alternative ex post perspective regarding the power of the LMZ PSM designs to detect meaningful underlying Big N quality impacts. Each of these percentages specifically address the likelihood that an Big N effect assessment with the relevant LMZ PSM design's estimated precision returns a "statistically significant" verdict given that

the underlying effect truly equals the given implicitly identified as meaningful FULL design estimate. Across both panel these estimates range from 12% to over 99%. When the set is restricted to those with smaller assumed underlying Big N effects (i.e., the bolded values) the expected rejection rates are 12% for ADA, 30% for RPEG, and 55% for ACCY. These sorts of rejection likelihoods are incompatible with a belief that the LMZ PSM designs reliably produce rejection outcomes for Big N effects that the LMZ design itself implicitly identifies as being of meaningful magnitude.

4. Conclusion

As the renowned statistician John Tukey observed, the collective noun for a group of statisticians is a quarrel. Statisticians are particularly quarrelsome when it comes to null hypothesis significance testing. It is a mistake, however, to think that such widespread, often vehement disagreement implies that every aspect of significance testing is in question or is "a matter of opinion." (Cready, 2019) There are core principles for conducting and interpreting tests of statistical significance about which statisticians broadly agree and which are fundamental to the integrity of test of significance based empirical assessments. Failure to heed them is a recipe for methodologically bankrupt analyses and transparently erroneous inferences.

The LMZ analysis strays from acceptable statistical significance testing practices when it equates failure to reject null hypotheses as in and of itself demonstrating that unobserved underlying differences in audit quality metrics between Big N and non-Big N auditors are truly either non-existent or of inconsequential magnitude. A test of statistical significance or P value does not speak to the truth of the tested null hypothesis per Principles 2 and 6 of the *ASA Statement on Statistical Significance and P*

Values (Wasserstein and Lazar, 2016).[16] Nor, per Principle 5 of the *ASA Statement*, do *p*-values provide reliable insights about the magnitudes of underlying effect sizes. In other words, the LMZ analysis, as presented, is constitutionally incapable of providing relevant insights about Big N effect importance or unimportance or, for that matter, insights about the consequentiality of client size matching design impacts on Big N effect estimates.

The descriptive analyses presented here directly address the inadequacies in LMZ's Big N effect assessment design. The most salient insight from these analyses is that the LMZ evidence is broadly compatible with conjectures that its matching designs identify Big N effects similar to and even a good deal larger than supposedly "significant" effects identified by non-matching based designs. Consequently, the LMZ evidence provides no clear insights whatsoever about either the absence of meaningful Big N quality effects or the existence of matching design identified reductions in the estimated magnitudes of such effects. That is, LMZ's methodologically unfounded interpretations of its evidence do not faithfully represent what its evidence indicates is true regarding client characteristics as an underlying explanation for observed differences between Big N and non-Big N audit quality proxy levels.

Finally, the subsequent literature's response to LMZ's misrepresentations illustrates the acquisition of knowledge consequences of such methodological malpractice when left uncorrected. According to the *Social Science Citation Index*, as of December 2019, nearly 300 articles cite LMZ with over 60 of these citations attributable to year 2019 publications. It is the 7th most cited article published by the *The Accounting Review* over the past 10 years, the second most highly cited from its publication year-2011. Moreover, based on a reading of a fair number of these articles, they adopt LMZ's

[16]Amrhein et al. (2019) label the use of the statistical insignificance outcome from a test of a zero effect or difference null hypothesis as a basis for making an empirical case that the underlying effect is non-existent or unimportant as "absurd" and "ludicrous."

misrepresentations uncritically. There is no instance of a study indicating that the bulk of LMZ's evidence is actually broadly compatible with Big N quality effects identified in the pre-LMZ literature. There are, however, frequent references to LMZ as a compelling basis for attributing prior evidence of Big N audit quality effects to inadequate control of client characteristics and non-linearity in the relation between client size and relevant audit quality measures.[17] Some articles further argue that the LMZ analysis evidences the non-existence of Big N quality effects. Studies also commonly reference LMZ as a compelling illustration in support of PSM design based estimates being vastly preferable to comparatively unreliable non-matching design based estimates. Hence, the literature has not responded to LMZ's arguably transparent misrepresentations of its evidence with deserved dismissal, skepticism, or criticism. Rather, these misrepresentations appear to be central to the literature's current understandings of both Big N audit quality and the saliency of propensity score matching analyses as a device for identifying and correcting spurious inferences attributable to unknown model specification deficiencies.

[17]Articles also sometimes link the LMZ evidence to the relevance of PSM designs for controlling for endogenous selection biases. However, the idea that PSM designs provide selection bias robustness in this setting is unlikely. The LMZ PSM design works by excluding client firms from the sample that have little or no ability to choose between a Big N auditor and a non-Big N auditor due to them being either too small for a Big N auditor to be a sensible choice or too large for a non-Big N auditor to be a sensible choice. That is, it excludes firms where largely exogenous factors (to the empirical issue at hand) dictate auditor choice while retaining firms that are able to freely choose their auditor type. In other words, the PSM design intensifies the saliency of any underlying endogenous selection effects by restricting the sample to only those client firms where the ability to freely choose auditor type exists.

References

Aloe, A., and C. Thompson. 2013. A synthesis of partial effect sizes. *Journal of the Society for Social Work and Research* 4(4), 390-545.

Amrhein, A., S. Greenland, and B. McShane. 2019. Scientists rise up against statistical significance. *Nature* 567, 305-307.

Amrhein, A., D. Trafimow, and S. Greenland, 2019. Inferential vs. descriptive statistics: There is no replication crisis if we don't expect replication," *The American Statistician* 73.sup1, 262-270.

Becker, C., M. DeFond, J. Jiambalvo, and K.R. Subramanyam. 1998. The effect of audit quality on earnings management. *Contemporary Accounting Research* 15(1): 1-24.

Behn, B., J-H Choi, and T. Kang. 2008. Audit quality and properties of analyst earnings forecasts. *The Accounting Review* 83(2): 327-349.

Butler, M. A. Leone, and M. Willenborg. 2004. An empirical analysis of auditor reporting and its association with abnormal accruals. *Journal of Accounting and Economics* 37: 139-165.

Cohen, J., 1992. A power primer. *Quantitative Methods in Psychology* 112(1): 155-159.

Cready, W. 2019. Complacency at the gates. *Significance* 16 (4): 18-19.

Cready, W., J. He, W. Liu, C. Shao, D. Wang, & Y. Zhang. 2019 Is there a confidence interval for that? A critical examination of null outcome reporting in accounting research. Working paper (August).

Cready, W., Liu, B., & Y. Zhang. 2020. A content based assessment of the relative quality of leading accounting journals. Working paper (January).

DeFond, M., D. Erkens, and J. Zhang. 2017a. Do client characteristic really drive the Big N audit quality effect? New evidence from propensity score matching. *Management Science* 63(11): 3628-3649.

DeFond, M., D. Erkens, and J. Zhang. 2017b The Big N effect persists after matching on client characteristics: A response to Lawrence, Minutti-Meza, and Zhang (2017), *Management Science* 63(11): 3652-3653.

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

Easton, P. 2004. PE ratios, PEG ratios, and estimating the implied rate of return on equity capital. *The Accounting Review* 79 (1): 73-95.

Francis, J., and J. Krishnan. 1999. Accounting accruals and auditor reporting conservatism. *Contemporary Accounting Research* 16 (1): 135-165.

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

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, Z. 2016. Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31, 337-350.

Khurana, I., and K. Raman. 2004. Litigation risk and the financial reporting credibility of Big 4 versus non-Big 4 auditors. *The Accounting Review* 79(2): 473-495.

Kothari, S.P., A. Leone, and C. Wasley. 2005. Performance matched discretionary accrual measures. *Journal of Accounting and Economics* 39(1): 163-197.

Lang, M., and R. Lundholm. 1996. Corporate disclosure policy and analyst behavior. *The Accounting Review* 71 (4): 467-492.

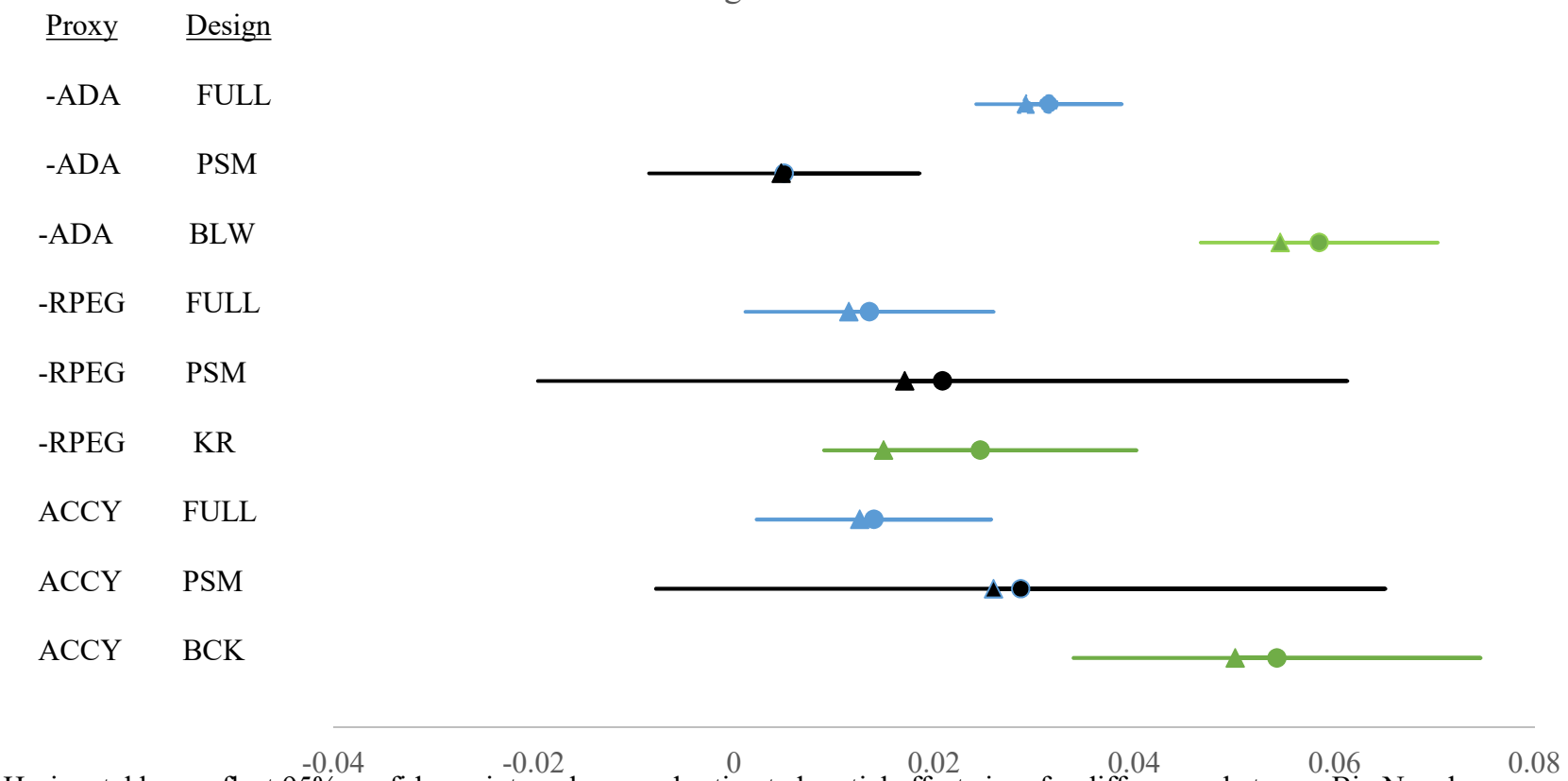
Lawrence, A., M. Minutti-Meza, and P. Zhang. 2011. Can Big 4 differences in audit-quality proxies be attributed to client characteristics? *The Accounting Review* 86(1): 259-286.

Lawrence, A., M. Minutti-Meza, and P. Zhang. 2017. The importance of client size in the estimation of the Big 4 Effect: A comment on DeFond, Erkens, and Zhang (2017)," *Management Science* 63 (11): 3650-3652.

Wasserstein, R.L., and N.A. Lazar. 2016. The ASA's Statement on *P*-values: Context, Process, and Purpose. *The American Statistician*, 70: 129-133.

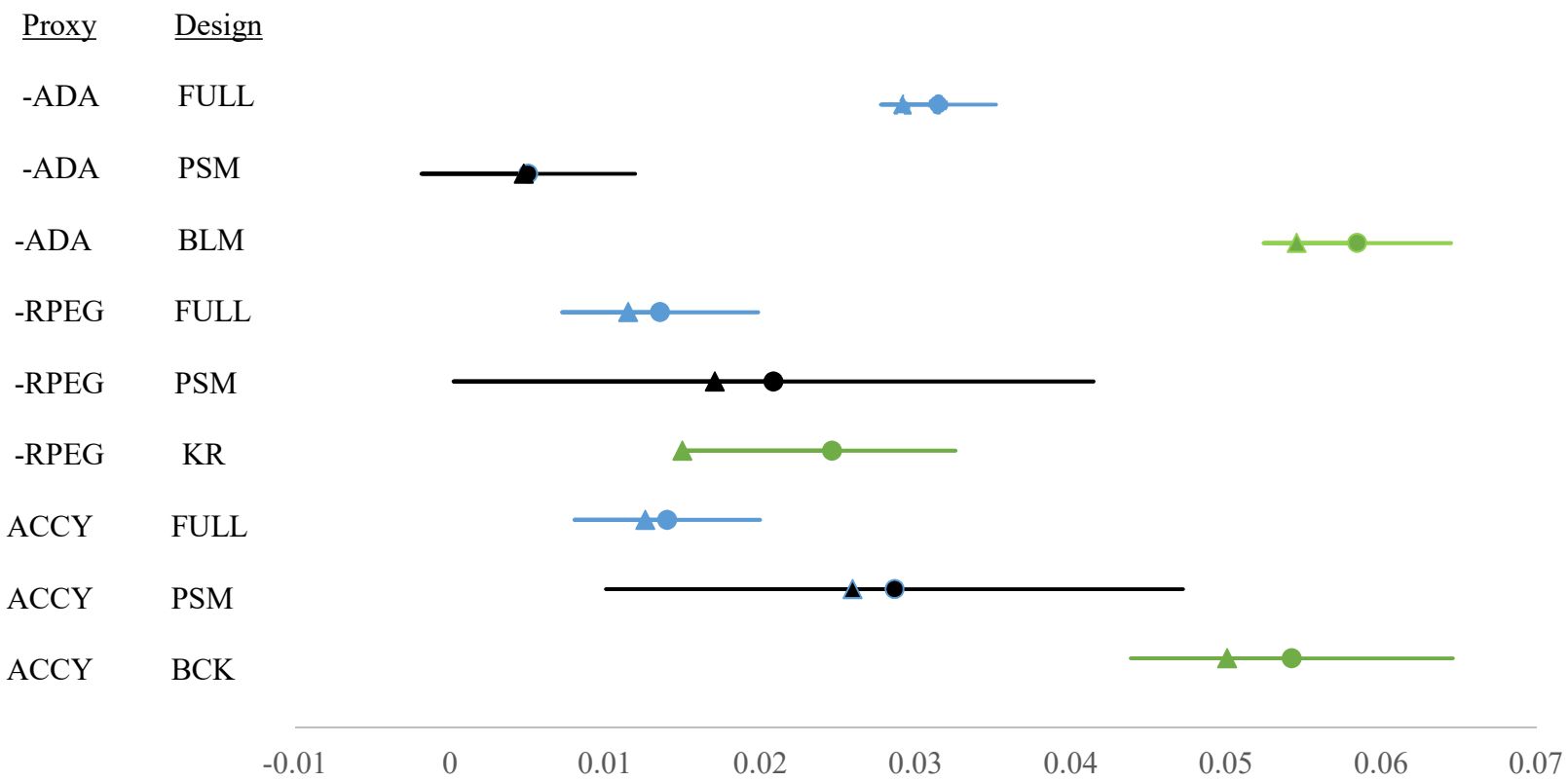
Wasserstein, R.L., A.L. Schrim, and N.A. Lazar. 2019. Moving to a World Beyond "p<.05." *The American Statistician* 73.sup 1: 1-19.

Figure 1
Effect Size Confidence Intervals For Audit Quality Proxy Differences Between Big N and non-Big N Auditors



Horizontal bars reflect 95% confidence intervals around estimated partial effect sizes for differences between Big N and non-Big N audit quality proxy levels. The three proxies examined are negative absolute discretionary accruals (-ADA), negative implied cost of equity capital (-ACCY) and analyst forecast accuracy (ACCY). In all cases positive values favor higher Big N auditor quality. FULL and PSM designs reference Lawrence et al. (2011) full sample and propensity score matched sub-sample multiple control variable regression estimates. The remaining three design designations reference existent literature estimates from designs Lawrence et al. argue are comparable to the Lawrence et al. FULL design analyses. Circle points indicate estimated partial correlations (on which the reported confidence intervals are centered) while triangle points identify semi-partial correlation estimates.

Figure 2
One Standard Deviation Effect Size Confidence Intervals for Audit Qualtiy
Proxy Differences Between Big N and Non-Big N Auditors



Horizontal bars reflect on standard error confidence intervals around estimated partial effect sizes for differences between Big N and non-Big N audit quality proxy levels. The three proxies examined are negative absolute discretionary accruals (-ADA), negative implied cost of equity capital (-ACCY) and analyst forecast accuracy (ACCY). In all cases positive values favor higher Big N auditor quality. FULL and PSM designs reference Lawrence et al. (2011) full sample and propensity score matched sub-sample multiple control variable regression estimates. The remaining three design designations reference existent literature estimates from designs Lawrence et al. argue are comparable to the Lawrence et al. FULL design analyses. Circle points indicate estimated partial correlations (on which the reported confidence intervals are centered) while triangle points identify semi-partial correlation estimates.

Table 1

Descriptive Assessments of Estimated Differences in Big N and non-Big N Audit Quality Proxies Reported by LMZ and in the Existent Literature (circa 2010) Based on Multi-Variable Regression Models

This table reports Big N quality estimates and associated 95% confidence intervals (CIs) from selected original propensity score matching (PSM) multiple regression analyses of absolute unexpected discretionary accruals (ADA), implied cost of equity capital (RPEG), and analyst forecast accuracy (ACCY). Panel A reports unscaled effect values while panel B reports effect values scaled by estimated standard deviations of the relevant quality proxy measure based on full sample standard deviation values provided by Lawrence et al. (LMZ). CI overlaps indicate the percentage of each quality metric's effect values contained in prior literature (table 1) or Lawrence et al. (table 2) FULL design CIs that are also contained in the tabulated PSM CI for the metric.

Panel B: LMZ FULL Sample Big N Effects and Associated Confidence Intervals

Quality Metric	Effect Type	Point Estimate	LB	UB
-1*ADA_{F_MZ}	Est.	0.0179	0.0138	0.0220
	Corr.	0.0315	2.42%	3.87%
-1*$RPEG_{F\_LMZ}$	Est.	37bp	3bp	71bp
	Corr.	1.35%	0.11%	2.59%
$ACCY_{F\_MZ}$	Est.	0.0042	0.0007	0.0077
	Corr.	1.40%	0.23%	2.57%

Panel B: Prior Literature Big N Effects and Associated 95% Confidence Intervals

Quality Metric	Effect Type	Point Estimate	LB	UB	P. Lit. CI Overlap
-1*ADA_{BLW}	Est.	0.002	0.0016	0.0024	0%
	Corr.	5.85%	4.66%	7.03%	0%
-1*$RPEG_{KR}$	Est.	30bp	11bp	49bp	100%
	Corr.	2.46%	0.90%	4.02%	68%
$ACCY_{BCK}$	Est.	0.0315	0.0191	0.0436	0%
	Corr.	5.43%	3.39%	7.46%	0%

Panel C: Estimated Big N Effects and Associated 95% Confidence Intervals

Quality Metric	Effect Type	Point Estimate	LB	UB	LMZ FULL CI Overlap	P. Lit CI Overlap
-1*ADA_{P_LMZ}	Est.	0.0018	-0.0030	0.0066	0%	100%
	Corr.	0.50%	0.85%	1.85%	0%	0%
-1*$RPEG_{P\_LMZ}$	Est.	21bps	-20bps	62bps	87%	100%
	Corr.	2.08%	-1.96%	6.13%	100%	100%
$ACCY_{P\_LMZ}$	Est.	0.0031	-0.0009	0.0070	90%	0%
	Corr.	2.86%	-0.78%	6.51%	100%	77%

Table 2
Comparative Analyses of Differences Between Multiple Regression and
PSM Big N Effect Estimates

This table reports differences in Big N quality effect estimates for absolute discretionary accrual (ADA), implied cost of capital (RPEG), and analyst forecast accuracy (ACCY) levels based on multiple variable regression specifications where no matching restrictions are imposed on the included set of observations (FULL) and specifications where inclusion is based on propensity score matching (PSM) restrictions. The first two columns report the indicated FULL design Big N effect estimate along with the assocated PSM design estimate. Standard error estimates, as imputed from reported article information, are provided in parentheses for each estimate. Standard errors small enough to identify the associated effect estimate as "statistically significant" at the .05 level are **bolded**. The final three columns report differences in the two effect value estimates, differences as a percentage of the relevant quality metric's standard deviation, and the percentage decline in the absolute magnitude of the PSM estimate relative to the paired FULL estimate. Panel A reports these statistics for the FULL multiple regression "replications" provided by LMZ while Panel B reports them for selected original FULL multiple regression analyses by BLW (Butler et al., 2004), KR (Khurana and Raman, 2004), and BCK (Behn et al., 2008).

Panel A: Differences Between LMZ FULL Estimates and. LMZ PSM Estimates					
FULL Design Audit Quality Proxy	FULL Estimate (s.e.)	PSM Estimate (s.e.)	Difference	Diff./SD	% Change
-1*ADA_{F_LMZ}	0.0179 **(.0021)**	0.0018 (.0025)	-0.0161	-11.77%	-90%
-1*$RPEG_{F\_LMZ}$	37bps **(17bps)**	21bps (21bps)	-0.0016	-3.58%	-43%
$ACCY_{LMZ}$	0.0042 **(.0018)**	0.0031 (.0020)	-0.0011	-2.52%	-26%
Panel B: Differences Between Prior Literature FULL and LMZ PSM Estimates					
FULL Design Audit Quality Proxy	FULL Estimate (s.e.)	PSM Estimate (s.e.)	Difference	Diff./SD	% Decline
-1*ADA_{BLW}	0.002 **(.0002)**	0.0018 (.0025)	-0.0002	-0.14%	-10%
-1*$RPEG_{KR}$	30bps **(10bps)**	21bps (21bps)	-9bps	-1.83%	-30%
$ACCY_{BCK}$	0.0316 **(.0061)**	0.0031 (.0020)	-0.0285	-65.22%	-90%

Table 3

Confidence Intervals for PSM Determined Estimated Effect Value Declines

This table reports estimated standard errors (SEs) and associated 95% confidence intervals (CIs) for differences between LMZ "replication" FULL and LMZ PSM Big N quality effect estimates for absolute discretionary accrual (ADA), implied cost of capital (RPEG), and analyst forecast accuracy (ACCY). Standard errors for cross-design differences in effect estimates are calculated as $SE_{DIFF} = (SE_{FULL}^2 + SE\_PSM^2)^{**.5}$, which holds under the conservative assumption of independence between the estimation errors in the FULL and PSM estimates. **Bolded** SE_{DIFF} values indicate that the value is small enough to identify the estimated reduction as "statistically significant" at the .05 level. All differences are measured in terms in the form of effect value reductions obtained from PSM estimates where negative values indicate absolute effect value increases.

FULL Design Audit Quality Proxy	Std. Error	Estimated Decline in Big N Effect Estimate (+ value indicate effect size increase)			Decline Values Scaled by Absolute FULL Design Difference Estimates (+ value indicate effect size increase)		
		95% CI L. Bound	Estimated Reduction	95% CI U. Bound	95% CI L. Bound	Estimated Reduction	95% CI U. Bound
-1*ADA_{F_LMZ}	**.0032**	-0.0225	-0.0161	-0.0097	-126%	-90%	-54%
-1*ADA_{BLW}	.0025	-0.0047	-0.0002	0.0051	-255%	-10%	235%
-1*$RPEG_{F\_LMZ}$	27bps	-69bps	-16bps	37bps	-186%	-43%	100%
-1*$RPEG_{KR}$	22bps	-53bps	-9bps	36bps	-177%	-30%	120%
$ACCY_{F\_LMZ}$	.0027	-.0064	-0.0011	0.0042	-152%	-26%	100%
$ACCY_{BCK}$	**.0064**	-.0410	-0.0285	-0.0161	-129%	-90%	-51%

Table 4

Likelihoods of Obtaining a Directional t-Value of 1.96 or More For Alternative Levels of Estimation Precision and Underlying Effect Parameter Values

This table reports likelihoods of obtaining a directional (favoring the existence of a favorable Big N quality effect) t-value of 1.96 or higher for salient standard error (SE) values when the unknown underlying true effect value equals the LMZ FULL design estimate. Considered SEs are: (1) LMZ's FULL and prior literature design SEs, SE_F, which corresponds to the expected SE for an hypothetical exact replication of the given FULL analysis on a new entirely independent equivalent sample size drawn from the same population; (2) SE_F adjusted to the smaller sample sizes employed in the LMZ PSM designs, (3) LMZ's PSM design SE, SE_P. **Bolded** entries identify values based on the smaller (in absolute terms) of the two effect value estimates considered for each audit quality proxy.

Panel A: Likelihoods Based on LMZ FULL Design Estimates

Audit Quality Proxy	FULL Design Effect	Design Standard Error Assumption		
		SE_F is the FULL Design SE from LMZ	SE_N Adjusts SE_F for LMZ PSM Design Sample Sizes	SE_P is the PSM Design SE from LMZ
		t approx.~ $N(ES_F/SE_F, 1)$	t approx.~$N(ES_F/SE_N, 1)$	t approx.~$N(ES_F/SE_P, 1)$
$-1*ADA_{F\_LMZ}$	0.0179	>99%	>99%	>99%
$-1*RPEG_{F\_LMZ}$	37bps	57%	10%	43%
$ACCY_{LMZ}$	**0.0042**	**65%**	**11%**	**55%**

Panel B: Likelihoods Based on Prior Literature Estimates

Audit Quality Proxy	FULL Design Effect	Design Standard Error Assumption		
		SE_F is the FULL Design SE from Relevant Prior Design	SE_N Adjusts SE_F for LMZ PSM Design Sample Sizes	SE_P is the PSM Design Standard Error from LMZ
		t approx.~ $N(ES_F/SE_F, 1)$	t approx.~$N(ES_F/SE_N, 1)$	t approx.~$N(ES_F/SE_P, 1)$
$-1*ADA_{BLW}$	**0.002**	**>99%**	**>99%**	**12%**
$-1*RPEG_{KR}$	**30bps**	**87%**	**21%**	**30%**
$ACCY_{BCK}$	0.0316	>99%	86%	>99%