A Study in “insignificance”: The “Big N” Audit Quality Kerfuffle

June, 2023

by

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

This paper has benefited from comments provided by Robbie Moon, an anonymous reviewer and workshop participants at the University of North Texas and the 2020 EAA Workshop on Audit Quality.

A Study in “insignificance”: The “Big N” Audit Quality Kerfuffle

In a highly influential analysis, Lawrence, Minutti-Meza, and Zhang (2011), LMZ henceforth, undertake modified replications of prior studies addressing relation between the use of a Big N auditor and various “accounting quality” measures/consequences. They interpret failures to replicate statistical significance outcomes after imposition of matching design sample constraints as reflecting Big N effect “insignificance”. This outcome leads it to claim that existing findings largely reflect difference in client attributes. This study, in contrast, approaches the LMZ setting as a matter better addressed from an estimation perspective. That is, it addresses where the LMZ evidence locates studied Big N effects. It shows that there is little substantive difference, beyond lower precisions of matching based estimates, between LMZ’s matching based estimates and those produced by non-matching designs. More generally, the analysis illustrates how misunderstanding of statistical significance and insignificance outcomes and neglecting statistical inference fundamentals broadly compromises accounting research integrity.

A Study in “insignificance”: The “Big N” Audit Quality Kerfuffle

1. Introduction

As the preamble of the American Statistical Associaton’s *ASA Statement on Statistical Significance and p-values* (Wasserstein and Lazar, 2016), identified as *The ASA Statement* henceforth) observes, misunderstanding of statistical significance (i.e., p-value) assessments of evidence is a widespread problem that often severely compromises the inferential value of empirical analyses employing it.[1] Key to this state of affairs is non-recognition of how closely the outcome of such an assessment is wedded to its tested null hypothesis (or model or belief). All that a statistical significance test possibly achieves is the identification of evidence that is incompatible with this tested null hypothesis. While “falsifications” of such null may provide indirect support for some other “alternative” hypothesis(es), “statistical significance” pertains to the null, not these alternatives. Moreover, and of direct relevance to this study in “insignificance, failed falsification efforts (i.e., “statistical insignificance: outcomes) say far less. How does failing to get a ticket possibly make the case that one did not speed?

This analysis explores consequences of such misunderstanding of statistical significance, particularly when pursued to the exclusion of broader estimation driven understandings of empirical evidence in stochastic settings.[2] It concerns a kerfuffle that arose, and based on the current relevant literature is still with us, about whether the very largest external audit firms (i.e., Big N auditors), on average, provide higher quality audits than non-big-N auditor competitors. Initial studies on the matter produced “statistical significance” outcomes across a number of “audit quality” associated metrics. A

[1] Importantly, the *ASA 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.

[2] See Greenland, Senn, Rothman, Carlin, Poole, Goodman, and Altman (2016) for a general presentation of ways by which researchers misinterpret and misapply statistical significance.

subsequent widely cited study by Lawrence, Minutti-Meza, and Zhang (2011), identified as LMZ henceforth, challenges these "significance" assessments. It reports that these same quality relations turn "statistically insignificant" when using samples matched on propensity for engaging a big N auditor. Based on this "insignificance", they argue that observed Big N effects are likely due to client characteristics, not auditor quality. [3] DeFond et al. (2017a), in turn, take issue with these insignificance "findings", but not because of the irrelevance of statistical insignificance to the resolution of such a question. Rather, they argue that LMZ's insignficance "findings" are not robust. They depend on the specific choices LMZ make in estimating propensity scores. Other, equally reasonable approaches to determining propensity scores commonly return statistically significant "findings."

In contrast to the over interpretation and misrepresentation of tests of "statistical insignificance" outcomes that pervade this kerfuffle, this analysis approaches matters from the perspective of estimation. Hence, it focuses on what examined empirical evidence suggests about Big N effect locations, not the p-values produced from test of null hypothesis probes of such evidence.[4] Descriptively, LMZ's favored matching designs do indicate compatibility with conjecturers that studied Big N effects are zero or even negative (i.e., that non-Big-N auditors provide higher quality) while considered multiple regression based designs strongly disfavor such locations. However, these possibilities must be balanced against the fact that these same matching designs also indicate similar compatibility with conjectures that the Big N effect is actually larger than full sample regression based Big N effect point estimates and, in a sometimes even upper (confidence interval based) coefficient bounds. Collectively, the descriptive case for LMZ matching designs identifying substantively smaller Big N effects than those obtained from

[3] 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."

[4] Dyckman and Zeff (2014) and Cready et al. (2022) find that inadequate, bordering on non-existent, descriptive engagement with evidence is widespread in the accounting literature.

multiple regression based non-matching designs is very weak. In fact, the LMZ evidence in its totality would likely fail a preponderance of the evidence test as applied by courts of law. As such designs also produce noisier estimates based on non-representative samples, there is little to recommend them and rather solid reasons to avoid using them in the audit quality settings studied by LMZ.

2. The LMZ Analysis

While the Big N effect kerfuffle encompasses numerous articles the LMZ analysis lies at the center of things. Hence, it is important to understand both what it does and the inferences that do and do not follow from its "findings."

LMZ 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), BLW henceforth, assess audit quality by measuring discretionary accrual levels. However, none of them employ performance-matching. Hence, LMZ's ADA measure is arguably in and of itself a material design innovation relative to the existing literature. RPEG and ACCY, on the other hand each correspond to measures employed in assessing audit quality. Khurana and Raman (2004), KR henceforth, introduce the RPEG to assess cross-country differences in Big N audit quality while Behn, Choi, and Kang (2008), BCK henceforth, introduce using ACCY as a way of linking audit quality with analyst forecasting properties.

LMZ use three different approaches for estimating Big N versus non-Big N differences in studied audit quality proxies: (1) unconditionally (i.e., difference in means) 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)' and, (3) various matching designs, the most prominent of which use propensity score based criterion (henceforth identified as PSM models). LMZ interpret the outcomes of these various Big N effect estimations through a statistical significance lens. It deems the assorted full sample estimates as "significant" as the evidences is incompatible with them being, on average negative or zero. It deems the PSM estimates insignificant since their analysis fails to identify compelling incompatibility with the proposition that they are negative or zero. Hence, in LMZ's view, the Big N effect is very small, arguably non-existent, and the reason that prior examinations failed to identify such smallness or non-existence rests in their inferiority (relative to PSM designs) in controlling for client characteristics.

LMZ's interpretations of its "findings" suffer from two serious deficiencies reflecting far from uncommon misunderstanding of statistical significance assessments of evidence. First, it does not take into account that statistical significance solely pertains to the tested null hypothesis. Here the tested null hypotheses are that various Big N quality effects equal zero. Consequently, LMZ's statistical significance outcomes identify incompatibilities of the evidence with such zero effect conjectures. By construction, they do not speak to the incompatibility of evidence with a conjecture that the effect is present but really close to zero (see Principle 5 of the *ASA Statement*). Indeed, given sufficient power, one should expect low p-value outcomes whenever the specified null is untrue regardless of how close to being true it may be. Second, and more severely, it equates failures to identify "statistical significance" with absence or inconsequentiality of an effect. All such failures signify is a compatibility between the evidence and such a belief, recognizing that the evidence is similarly compatible with contrarian to the null beliefs.[5] In other words, absence of statistical significance, is not a basis for thinking effects are absent or inconsequential. Nor, to the likely surprise of the many adherents to the centrality of "falsification" testing as a basis for asserting absence of evidence about things, is it a basis for thinking that the examined evidence is silent on the matter.

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 et al. (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

[5] Gelman and Stern (2006) provide a widely cited comprehensive critique of using statistical significance and insignificance as a basis for assessing difference. More pointedly, Amrhein et al. (2019) labels such absence or inconsequentialiy attributions as "absurd" and "ludicrous."

examined quality measures or empirical settings. They do not, as a rule, directly engage with the veracity of LMZ's interpretation of its evidence.

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. Their replications of LMZ's PSM analyses indicate that PSM designs comparable to those chosen by LMZ typically produce statistically significant outcomes. They 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. As is clear from relevant *ASA Statement* principles, it does not 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. Statistical significance of no-effect-at-all null hypotheses is not suited to such magnitude of effect or difference assessments (Cready, 2022).

It is also of some relevance to point out that LMZ does 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 estimation centered examinations of its evidence.

3. Estimation Based Assessment of the LMZ Evidence

Estimation assessment of evidence concerns rigorous determination of what examined evidence suggests or reveals about the location of an unknown effect, difference, likelihood, association, or relationship of interest (henceforth collectively referred to as effects of interest). As such, it differs from null hypothesis significance based inference. Such inference focuses on incompatibility of evidence with a known location or difference of interest (typically zero in the accounting field). This distinction is important because while significance testing requires estimates, estimation assessment most certainly does not require significance testing. Indeed, as the remainder of this section illustrates, statistical significance ideas mostly confuse estimation assessments of evidence.

3.1 Big N Effect Point Estimates

Point value estimates of effects of interest are the obvious starting place for locating unknown effects of interest. Such estimates commonly identify expected values of the underlying unknown parameter of interest under the assumptions of the underlying research design. However, they are only expectations, not certainties. From an estimation perspective, they simply provide a rough, arguably unbiased, idea of an effect of interest's general location.

Table 1 provides point estimates for the Big N effect on ADA, RPEG, and ACCY from relevant prior literature, LMZ FULL replications of prior literature, and LMZ PSM multiple regression models. All of the estimates are positive. Hence, at the best point estimate level, the evidence suggests higher, on average, Big N audit quality. That said, the PSM estimates are smaller than their non-PSM companions. For ADA the BKL estimate is .0020, which is 11% higher than the PSM estimate of .0018. For RPEG, the KR estimate is 30 basis points, which is 44% higher than the PSM estimate of 21bp. For ACCY the LMZ FULL replication estimate is .0042, which is 35% higher than the PSM estimate of .0031. Hence, at the point estimate level there is support for the idea that PSM models identify smaller

Big N effects. That said, in the cases of RPEG and ACCY the PSM point value estimates are in and of themselves possibly consequential. In the case of RPEG a 21 basis point reduction in cost of capital translates into to $2.1 million in benefit for a one billion market capitalization firm, a value that seems likely to materially impact such a firm's thinking regarding paying a Big N audit premium. In the case of ACCY, mean ACCY for the PSM equals -0.162 (table 7 of LMZ). By construction, ACCY is capped at 0. Hence, the 0.0031 estimate accounts for nearly 20% of the on average improvement in ACCY that is possibly attainable for this sample.

On the other hand, the BLW ADA estimate is 0.002, which is a scant .2% of total assets. This value is noteworthy since Lawrence, Minutti-Meza, and Zhang (2017), in responding to the DeFond et al. (2017a) criticism of their original analysis, argue that ADA differences in the neighborhood of .2% of assets lack economic significance. Hence, one perspective of this estimate is that it indicates that the prior literature evidence already reveals that Big N impacts on ADA lack substantive importance. Thus, all a PSM reassessment can accomplish is the re-demonstration of this insight. Another perspective, however, one advanced by DeFond et al. (2017b), is that .2% of asset deviations are often of material importance in audit materiality assessments.

3.2 Standard Errors

While point estimates provide locational insight, they do so with error. Moreover, depending on circumstances, the degree of error attached to such estimates can be considerable. Serious estimation exercises consider such error in their identifications of effect locations. The sample variance based standard error is one widely recognized measure of the amount of error attached to an estimate. The lower the standard error, the more precise the estimate. Hence, when choosing among competing point estimates of a parameter of interest one generally prefers the one with the lowest standard error. Why, for instance, would one ever prefer an estimate obtained from a random sample of 10 over one obtained from a random sample of 1000?

Table 2 provides the standard error associated with each of the Big N effect estimates reported in table 1. These values reveal that the prior literature ADA and RPEG estimates are far more precise than the LMZ estimates. The standard errors of the LMZ ADA estimates are ten times the size of the BKL estimate's 0.0002 error. As LMZ employ larger sample sizes than BKL, the most likely driver of much of this difference is LMZ's use of performance adjusted absolute discretionary accruals rather than just absolute discretionary accruals. As the adjustment used to determine performance adjusted absolute accruals itself exhibits variation it necessarily increases unexplained variation in ADA. That is, it makes things noisier which, in turn, increases standard errors. The standard errors of the LMZ RPEG replication estimates are also 70% to 110% larger than the KR estimate's 10bp error. Hence, under *ceteris paribus* conditions, the LMZ ADA and RPEG estimates are of decidedly inferior inferential quality than the prior literature estimates that they seek to displace. Both of LMZ's ACCY estimates, on the other hand, fare better. Their standard errors are roughly a third the size of the BCK ACCY estimate. Also, of relevance to LMZ's (implicit) contrasts of its FULL and PSM estimates, each of the PSM estimate standard errors is between 10% and 20% larger than its companion FULL estimate. That is, *ceteris paribus*, holding initial sample size constant they produce are noisier and so more prone to high *p*-value outcomes, just the sorts of outcomes that LMZ "find" for them.

3.3 Interval Identification

Fundamentally, the fact that statistics does not determine parameters with certainty implies that relevant estimation focus more on intervals than on solitary point values. That is, given the impossibility of determining the precise value of an underlying parameter of interest, the more feasible objective should be identifying what examined evidences suggests about its general location. Confidence intervals are one widely recognized approach for inferring such locations. Simply put, a confidence interval identifies a range of most evidence compatible candidate values for a parameter. Importantly, there is no guarantee that this range contains the true parameter value, only that the likelihood that it fails to do so meets a specified tolerance for being completely wrong level.

One readily implementable approach to determining confidence intervals employs parameter point estimates in conjunctions with their associated standard errors. The analysis that follow provide so-constructed 99.7% (three standard error), 95%, and 68% (one standard error) confidence intervals for the Big N effect estimates reported in table 1. I employ a range of confidence levels to reflect the fact that tolerance for misidentification of an underlying parameter's location is situational. When the issues in play prioritize the identification of every possible location within reason, as might be true in defining the search area for a lost hiker in the wilderness, then a three or more standard error confidence interval seems in order. If the objective is the identification of a relatively more likely set of locations, as would likely be true in determining which locations to examine first in the search for our lost hiker or is roughly the case in the determination of hurricane track forecasting cones (see https://www.nhc.noaa.gov/aboutcone.shtml), then a one standard error confidence interval would likely suit. Finally, of course, a 95% confidence interval aligns with the widespread acceptance of 5% error rates as a key null hypothesis testing inferential dividing line. That said, its locational identification attractiveness is unclear. It mis-locates 5% of the time, so it is not providing an effectively exhaustive identification of possible locations.[6] Nor does it provide clarity about where the evidence particularly locates the unknown effect of interest.

[6] A way of putting the attractiveness of the 95% confidence interval as the "right" choice for scholarly research publications is the implications of such a choice for the overall "correctness" of a scholarly journals content. Would the idea that one in every 20 so reported intervals being completely wrong be an acceptable level of false identifications for a high quality journal?

ADA Confidence Interval Analysis

Table 3 provides confidence interval guided locations for the Big N effect on ADA identified by the prior literature BLM full sample multiple regression estimation, the LMZ FULL replication and the LMZ propensity score matching estimation. The BLM intervals reflect where the pre-LMZ evidence places the Big N ADA effect's location. In the extreme, this evidence places it between 0.0014 and 0.0026 (99.7% confidence interval) while the more targeted but also more error prone 68% confidence interval places it between 0.0018 and 0.0022. Despite its relative width, the 99.7% interval is of particular interest because of what it does not contain. Specifically, the 0.0179 estimate produced by LMZ's FULL replication of BLM. This failure suggests that while LMZ's FULL analysis successfully replicates BLM's statistical significance determined inference that the evidence is incompatible with the LMZ effect being less than or equal to zero, its point estimate does not locate it anywhere near where the BLM analysis suggests it ought to be found. The narrower 68% interval is also of interest, but for the opposite reason. The LMZ PSM point estimate of 0.0018 lies at the very edge of this interval, thereby identifying it as at least borderline compatible with a comparatively narrow BKL identification of the ADA effect's location. In other words, it indicates that LMZ's PSM point estimate is unsurprising in light of the BLM evidence.

The LMZ FULL replication intervals raise further concerns about the lack of conformity between where this replication effort places the LMZ effect relative to where the BLM multiple regression places it. The 99.7% confidence interval places the ADA effect between 0.00116 and 0.0242, a range of values that dwarfs the 0.0014 to 0.0026 range identified by the BLM analysis. A possible explanation for this divergence, one that I revisit in a later section, is that BLM and LMZ differ in how they measure ADA. BLM measure it as absolute discretionary accruals while BLM measure it as absolute performance adjusted discretionary accruals.

The LMZ PSM 99.7% interval locates the ADA effect as falling somewhere between -0.0057 and 0.0093. Hence, it is compatible with conjectures that the ADA effect is essentially zero, zero, or even that it favors higher non-Big N audit quality. This range also, however, fully encompasses the BKL 99.7% confidence interval values. So, it is similarly supportive of conjectures that the ADA effect falls within the BKL identified intervals. What it clearly does not support is conformity with the LMZ FULL multiple regression based confidence intervals. The PSM design clearly identifies a range of substantially smaller performance matched ADA effects. The PSM 68% interval largely reinforces these inferences. Similarly, the 95% interval indicates that significant test assessments of null hypotheses that the effect is larger than 0.0026 (the upper bound of the BKL 99.7% confidence interval), zero or less than zero would produce "insignificant" outcomes. On the other hand, the evidence is not compatible with a null hypothesis that the effect equals or exceeds 0.0116 (the lower bound of the ADA FULL 99.7% confidence interval).

RPEG Confidence Interval Analysis

Table 4 provides confidence intervals for the estimated Big N reduction in RPEG. The 99.7% interval for the KR multiple regression estimation places the effect somewhere between 0 and 60 basis points. Hence, it admits to the remote possibility that the RPEG effect equals or nearly equals zero. That said, it is certainly true that the more error tolerant 68% interval sets the lower bound at a rather substantial 20 basis points (a 20 basis point reduction in cost of capital translates into two million dollars for a one billion dollar market cap firm). Similarly, the 95% "statistical significance interval" places the effect as falling between 4 and 20 basis points. The noisier LMZ FULL regression three standard error confidence interval lower bound is -14bp, thereby identifying an on average increase in RPEG as a possibility. Here again, however, the more error tolerant one standard error interval lower bound identifies a 20 basis point reduction benefit minimum. As each of the LMZ FULL intervals fully overlaps its companion KR interval value it is clearly the case that they provide broad, albeit noisier, support for the KR interval identification of the RPEG effect's location.

The 99.7% confidence interval for RPEG based on LMZ's PSM analysis places the effect as falling somewhere between -42 and 84 basis points. While this range certainly entertains the possibility that the RPEG effect is zero or even leans in favor of non-Big N audits, its more relevant property is that it fully overlaps the KR determined intervals. While there is certainly a strong case here for thinking that PSM designs produce noisier estimates, there is no such clarity in the evidence regarding the assertion that PSM designs identify a meaningfully smaller RPEG effect relative to the effects identified by the KR and LMZ full sample multiple regression designs. Rather, taken as an entirety and after making allowance for differences in precision, the evidence here portrays a surprisingly (to the many authors who have cited LMZ at least) consistent picture of the Big N RPEG effect's location.

ACCY Confidence Interval Analysis

Table 5 provides relevant confidence intervals for the Big N improvement in ACCY. The 99.7% interval based on the BCK pre-LMZ analysis broadly places this effect between 0.0133 and 0.0499. The LMZ multiple regression replication of BCK, however, places the ACCY effect elsewhere. The FULL design based 99.7% interval's upper bound of 0.0096 places the ACCY effect well below the BCK interval. Hence, again LMZ "replicate" statistical significance but clearly fail to replicate where the evidence places the underlying effect's general location. In this case, the likely explanation is time period differences between the two samples. The LMZ analysis draws observations from as early as 1988 (when forecasts are available for a small number of widely known firms) and as late as 2006 (when forecasts coverable is widespread) while BCK limits its analysis to the 1996-2001 time period. Consistent with such a time period effect, the BCK sample ACCY mean is -0.030, reflecting a far higher baseline lack of accuracy than the LMZ -0.0122 sample mean. In fact, lower bounds on the mean Big N improvement in ACCY of 0.0133 to 0.0255 from the BCK analysis exceed this this mean.

The PSM 99.7% interval's identifies the ACCY effect as ranging between -0.0029 (i.e., Big N audits lead to lower accuracy) and 0.0091. This interval overlaps 95% of the far more relevant LMZ FULL 99.7% interval with the difference attributable to the PSM interval entertaining somewhat stronger adverse Big N audit impacts on ACCY. The far more targeted 68% interval contains only positive values, locating the ACCY improvement between 0.0011 and 0.0051. Collectively, the evidence suggests broad compatibility between where the FULL and PSM designs locate the ACCY effect. The primary difference being that the higher precision provided by the FULL design provides "statistical significance" clarity regarding the incompatibility of this evidence with the conjecture that the ACCY effect is possibly zero while the PSM design's imprecision leaves it short of such clarity.

3.4 More on the ADA Measurement Matter

While LMZ follow the prior literature with respect to their measurement of Big N effects on RPEG and ACCY its ADA analysis does not. It replaces absolute discretionary accruals with absolute performance adjusted discretionary accruals. Crucially, performance matching determines the discretionary accruals of a studied firm by differencing it from the discretionary accruals of "the closest ROA firm in the same two-digit SIC code" (LMZ, fn 11). Consequently, a firm with completely unknown characteristics apart from industry membership and current period ROA (identified as the PM firm henceforth) determines a substantial portion of the variation in LMZ's ADA measure. So,

$$\mathrm{ADA}_i = \mathrm{b0} + \mathbf{b^{*}ADA\_Determinants}_{\mathbf{i}} + \mathbf{b^{*}ADA\_Determinants}_{\mathbf{pm,i}} \qquad (1)$$

Where **ADA_Determinants** identifies a vector of relevant determinants of ADA inclusive of BigN auditor choice and the i subscript references a studied sample firm and the pm,i subscript identifies the firm that is performance matched to sample firm i (identified as the PM firm henceforth). The two sets of determinants are additive. Addition happens here despite the fact that they are differenced in the determination of ADA because ADA is an absolute value.(not a difference in absolute values). Apart from co-variance connections, it responds positively (negatively) to variation (suppression of variation) in its individual components irrespective of their respective signs. Positive co-variances (e.g., the directional performances of the paired firms) will reduce ADA, but the LMZ analysis does not address the use of directional performance matched accruals. Hence, without loss of generality, I ignore such co-variance offsets or amplifications here in the interest of parsimony.

An obvious consequence of ignoring ADA's PM determinants, as LMZ's FULL replication does, is an increase in estimation error—it leaves explainable variation in ADA unexplained. More importantly, as the PM determinants are conditional on performance it also follows that each correlates with its studied sample companion. For instance, when a sample firm's $ROA_{i,t}$ is high both it and its performance match, relative to random assignment, have higher likelihoods of: being large firms; having high ROA values (by construction in t and by expectation in t-1); having Big N auditors, having high leverage, having low current ratios. Based on the LMZ evidence, ADA varies inversely, both unconditionally and incremental to each other, with all of these variables. Hence, apart from the current ratio, higher values of these variables push the unconditional correlation between performance (i.e., ROA_t) and ADA lower (more negative). Consequently, omitting one or more of them from the PM vector transfers their ADA impacts to ROA_t. Further omission of ROA_t (which is the case in LMZ) transfers this downward impact to the parameter estimates of those variables that happen to be included (e.g., Size, Big N auditor choice, past performance, leverage). Hence, LMZ's omission of the PM firm's characteristics contributes to the difference between the LMZ and BKL full sample multiple regression based Big N effect estimates. The extent of this contribution depends upon how much of this omitted variables bias lands on the Big N variable as opposed to other included variables such as lagged ROA and firm size. If the overall bias is sizable and Big N performs particularly well relative to other control variables at picking it up, then ADA performance matching likely accounts for much of the upward bias in the LMZ FULL Big N estimate.

The LMZ size matched Big N effect estimates provide some relevant insights about the extent to which performance matching impacts the LMZ FULL Big N effect estimate. This analysis finds that a mean difference in ADA between Big N and non-Big N audited firms of -0.0030. A value that aligns with the BKL non-performance matched ADA estimates. More importantly, including the complete set of LMZ control variables barely moves it. It drops to -0.0026. Hence, this matching design is highly effective. It seemingly controls for both the direct omission of measured correlated determinants as well as performance matching bias produced add-on effects. Seemingly, this outcome identifies firm size, not audit type, as the performance matching bias's primary conduit. However, closer consideration of how a choice to match on size interacts with the underlying distributions of Big N and non-Big N firms suggests a different possibility. Non-Big N firms are rarely large. Consequently, matching on size effectively divides the sample between small firms and large firms in much the same way that an indicator variable for firms being of above or below average size divides it. If division by size is primarily responsible for the robustness and replicative properties of this design's Big N estimates then we are looking for a variable that similarly isolates large firms in a dichotomous fashion as the likely conduit of performance matching bias in LMZ's full design. That variable is surely a firm's non-use of a Big N auditor. The scarcity of non-Big N audited large firms, after all, is what produces the matching design's widespread removal of large firms from the analysis.

3.5 Partial Correlation "Effect Size" Assessment

The presented evidence thus far focuses on regression parameter values estimates of Big N associated quality improvements inclusive of associated estimation uncertainty levels. The descriptive saliency of such parameter value based inference depends on the contextual meaningfulness of such parameters. For instance, the RPEG parameter is contextually understandable in that it identifies an implied cost of capital benefit from choosing to employ a Big N auditor over a non-Big N auditor. Along the same lines, the ACCY improvement estimates suggest a meaningful increase in accuracy. However, there is no clear way to map this improvement to decision driven value-relevant outcomes. So, it is arguably something of a black box. In contrast, the economic meaningfulness assessment of the reduction in ADA is debatable. The BKL ADA estimate of 0.002 corresponds to .2% of total assets. Is this a meaningful improvement? Interested scholars disagree. Lawrence et al. (2017), in responding to the DeFond et al. (2017a) criticism of their original analysis, argue that ADA differences in the neighborhood of .2% of assets lack economic significance.[7] DeFond et al. (2017b) suggests that from an audit materiality threshold .2% of total assets value variations in account balances is important.

Assessing the extent to which a studied effect of interest influences or explains variation in dependent variables of interest is a commonly advanced alternative or supplemental approach to assessing consequence (e.g., in the accounting literature, Johannesson, Ohlson, and Zhai, 2022). Importantly, the saliency of such "effect size" assessment strategies directly depend on the stand-alone research importance of explaining dependent variable variation. Independent variables are important in this paradigm when they provide noticeable improvements in dependent variable outcomes. Hence, for instance, Cohen's (1988, 1992) seminal minimum effect size thresholds stem from the idea of a studied behavioral intervention's impact on observed variation in subject behavior being noticeable to a careful

[7] This appeal to economic inconsequentiality is the first time the LMZ authors seriously engage the actual values of the Big N effect estimates they report.

observer of such subject behavior.[8] In other words, effect size saliency increases with the extent to which the matter of primary research interest is determining or explaining dependent variable variation. Unfortunately, accounting research efforts commonly focus on the importance of some novel independent variable with the dependent variables playing a distinctly supporting role in things. LMZ is a case in point. Its objective concerns audit quality variation by auditor type. It is not at all concerned with better explaining variation in ADA, RPEG, or ACCY apart from what it suggests about audit quality differences. Consequently, while this section's effect size assessments provide a supplemental perspective of the LMZ and related study evidence, there is no reason to view them as a particularly compelling approach to understanding this evidence.

I use partial correlations as measures of Big N quality effect sizes. A number of options exist for measuring effect size, including the rather clunky one decile or standard deviation movements in independent variable impacts encountered in more estimation engaged accounting research efforts. My use of partial correlations follows from their widespread acceptance in the broader meta-analysis literature and the fact that alternative measures are commonly variations (e.g., standardized regression coefficients) on such correlations. Fundamentally, a 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

[8] The behavioral psychology literature, which commonly employs effect size assessment measures (e.g., Cohen's *D*), generally views the explanation of less than 10% of relevant dependent variable variation as trivial. Such explanatory levels, following broad guidance provided by Cohen (1988, 1992) regarding "large," "medium," and, "small" effect size attribution, do not explain even remotely observable variations in subject behaviors. However, as Cohen (1988) cautions, his interpretive structure defines "the terms 'small,' 'medium,' and 'large' relative, not only to each other, but to the area of behavioral science … (including) the content and research method being employed." Indeed, to reinforce Cohen's point here, were this 10% standard adopted by the accounting literature it would be fair to say that most accounting research findings would be designated as "trivial."

assessing effect size that is amenable to the conduct of cross-study, cross-design, and cross-metric integrations and comparisons of evidence.[9]

Table 6 reports ADA, RPEG, and ACCY partial correlation estimates and associated 95% confidence interval bounds for the pre-LMZ BKL, KR, and BCK analyses (panel A), their LMZ FULL design replications (panel B), and the LMZ PSM design analyses (panel C). The pre-LMZ literature's correlation estimates range between 2.46% for RPEG and 5.85% for ADA. Such values are not large in an absolute sense. Yet, if one were to go to the pages of any of our research journals and determine partial correlation values for reported estimates of interest one would, in my casual experience with such determinations, find such values to fall on the high side of things. In fact, the Johannesson, Ohlson, and Zhai (2022) evidence suggests that partial correlations in the neighborhood of one to two percent are truly a norm for our field.   The .90% to 4.02% RPEG confidence interval broadly locates its correlation closer to zero relative to where the ADA and ACCY locate their correlations. The ADA and ACCY upper bounds of 7.03% and 7.46% indicate that the evidence does not rule out the possibility (following a statistical significance interpretive perspective) of these correlations exceeding 7%.

The LMZ FULL replication of BKL places the ADA partial correlation between 2.42% and 3.87%. This interval is well below where the BKL evidence places it (lower bound is 4.66%). Hence, whereas the ADA parameter value intervals place the LMZ ADA effect completely above the pre-LMZ identification of the ADA coefficient's location, the opposite is true for partial correlations. The likely explanation for this reversal is the added estimation noise injected into the ADA measure when it is performance adjusted. Partial correlations, unlike coefficient estimates, respond (toward zero) to

[9] 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.

increases in dependent variable measurement error. The RPEG interval of 0.11% to 2.59% overlaps 68% of the KR RPEG interval, supportive of the LMZ RPEG design being a successful replication. As was true for the coefficient intervals, the ACCY interval of 0.23% to 2.57% locates the correlation below where the BCK analysis places it.

The most striking feature of the PSM design correlations is that the RPEG and ACCY point estimates of 2.08 % and 2.86% exceed their companion FULL sample design point estimates of 1.35% and 1.40%. Hence, a preponderance of the evidence standard here favors the conclusion that relative to highly comparable (in terms of studied population) non-matching multiple regression designs, PSM designs identify larger, not smaller, RPEG and ACCY levels of explained variation. As the RPEG and ACCY intervals also extensively (mostly entirely) overlap the companion pre-LMZ and FULL intervals, the evidence here is a woefully inadequate basis for attributing unique explanatory relevance of any sort to the PSM RPEG and ACCY estimations. On the other hand, there is a superficial case that the PSM design identifies a substantially smaller ADA correlation (-0.85% to +1.85%) than that identified for the BLM analysis (4.66% to 7.03%). The superficiality here stems from the additional measurement error in LMZ's performance matched ADA measure. As this measurement error increase is sizable, this evidence is not a sound basis for identifying the reduction as PSM driven.

Partial correlation based forest plots (Cummings 2012) provide collective understanding of where evidence from diverse designs and measures place phenomena of interest. Figure 1 presents a forest plot of the nine 95% confidence intervals reported in table 6. Collectively, the nine intervals locate a generic improvement in explained audit quality variation associated with Big N audits that leans very much to the side of harmony over discord. The PSM intervals lean to the zero side of things while the pre-LMZ intervals lean to the upside, but that is about it. Yes, the PSM intervals do stretch over the zero line, but how surprising is this given that they are, in relative terms, quite wide? In fact, the RPEG and ACCY PSM intervals are so wide that they pretty much encompass the entire forest.

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 is not that sensitive to the fact that matching designs place less emphasis on independent control variables in addressing dependent variable variation.

4. Conclusion

In reaching its conclusions regarding the inferential consequences of propensity score matching assessments of evidence and Big N effect "insignificance" the LMZ analysis relies upon significance/insignificance contrasts. In doing so they pay no substantive attention to what the actual evidence they examine, as summarized by Big N effect point estimates and confidence intervals, reveals about the veracity (or lack thereof) of such claims. This evidence, as examined here, indicates that there is little basis for drawing substantive distinctions between multiple regression full sample based "statistically significant" outcomes and propensity score matching "statistically insignificant" outcomes. Both approaches place Big N effects in similar, extensively overlapping in confidence interval assessments, locations. They do differ in terms the respective levels of precision they bring to the data.

Propensity score matching estimates are less precise. Pursuing imprecision, however, is not a pathway leading to better understandings of evidence. Hence, the LMZ PSM based analysis says little beyond the identification of what is likely a sample time period driven decline in the degree to which Big N audits improve analyst forecast accuracy that is just as readily observed in non-matching based assessments of the data.

The analysis also reveals the centrality of direct engagement with estimates to sound replication analysis (Cready, 2022). LMZ attempt to replicate relevant Big N effect findings from the prior literature as a means of better connecting their analysis with that literature. While they do replicate the statistical significance outcomes from this literature, their fixation on statistical significance leads them to miss the fact that their estimates place two of the three big N effects in very different locations. That is, at the far more fundamental parameter value location level, these are failed, not successful, replications. Moreover, directly reflecting how blinding fixation on statistical significance can be, had LMZ recognized these failures for what they were and investigated they would likely have discovered the highly problematic nature of their decision to depart from prior literature by employing performance matched accruals.

The blindness of LMZ's authors to the truly insignificant nature of their "findings" is possibly understandable. In my experience embedding oneself too tightly into the pursuit of an "in one's mind" research objective often brings with it tunnel vision about the bigger picture, particularly when picture doesn't align with objective. Of considerably more interest is the apparent widespread failure of LMZ's reviewers, editors, and citers, the last group surpass one thousand per Google Scholar (126 in 2022), at ascertaining LMZ's insignificance.[10] It seems that we, as a discipline, have much to learn (accept?) about the constitutional limitations of null hypothesis significance testing. Statistical significance targets

[10] See Cready 2019, 2022) and Cready et al. (2020) for further evidence of how the field misunderstands statistical significance assessments of evidence.

incompatibility of evidence with a specific (null hypothesis) conjecture about what produced it. Nothing more. It is not, in particular, a ready basis for drawing broad (prior altering) insights about the truth or inconsequentiality of tested conjectures. Given this demonstrated misunderstanding on these matters, our discipline's journals would do well to discourage hyping of statistical insignificance outcomes in their pages. Indeed, an outright ban applied at the desk rejection assessment stage might even be in order. They would also do well to actively welcome substantive engagement with the estimation side of inference, recognizing that estimation is often a messy business that does not lend itself to the sorts of (commonly fake) dramatic insights provided by test of hypothesis outcomes.

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.

Cohen, J., 1988. *Statistical Power Analysis for the Behavioral Sciences*. Routledge. ISBN 978-1-134-74270-7.

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

Cready, W. 2022. Accounting research's 'flat earth' problem. *Accounting, Economics, and Law: A Convivium*. https://doi.org/10.1515/ael-2021-0045

Cready, W., J. He, W. Liu, C. Shao, D. Wang, & Y. Zhang. 2022. Is there a confidence interval for that? A critical examination of null outcome reporting in accounting research. *Behavioral research in Accounting* 34(1): 43-72..

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

Cumming, G. 2012. Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. New Your. Routledge.

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.

Johannesson, Erik and Ohlson, James A. and Zhai, Sophia Weihuan, The Explanatory Power of Explanatory Variables (Oct 23, 2022). Available at SSRN: https://ssrn.com/abstract=3622743 or http://dx.doi.org/10.2139/ssrn.3622743

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.

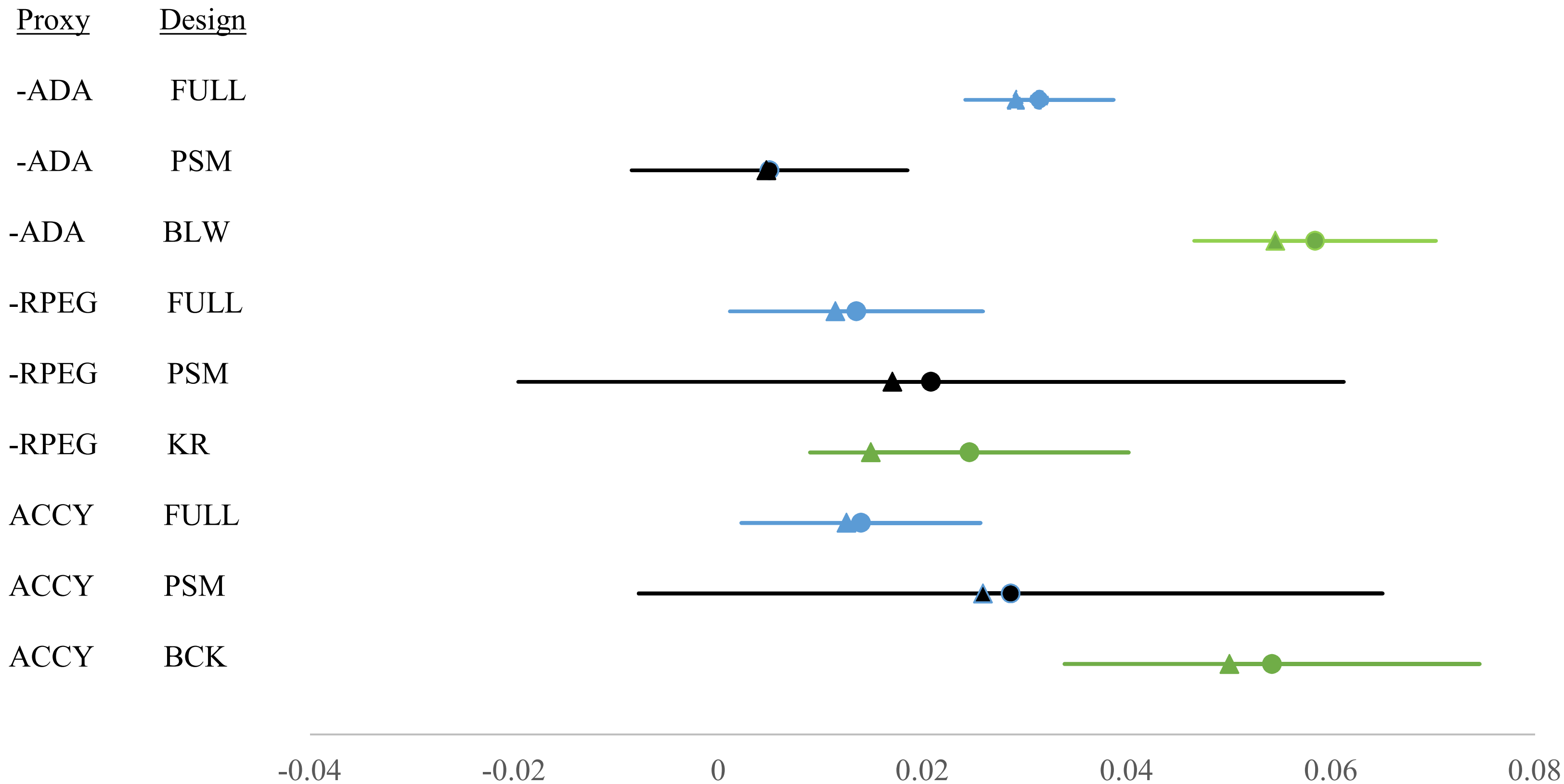
Lawrence, A., M. Minutti-Meza, and P. Zhang. 2011. Can Big 4 differences in audit-quality proxies be attributed to client characteristics? *The Accounting Review* 86(1): 259-286.

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 (-RPEG) 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

Big N Effect Point Estimates

This table reports Big N quality difference estimates from pre-LMZ studies by Butler et al. (2004), Khurana and Raman (2004), and Behn et al. (2008), LMZ's replications of those analyses (FULL estimates), and LMZ's propensity score matching design estimates. The relevant quality proxies being absolute (performance matched in the case of LMZ estimates) discretionary accruals (ADA), implied cost of equity capital (RPEG), and analyst forecast accuracy (ACCY).

Quality Proxy	Direction of Big N Effect	Pre-LMZ Estimate	LMZ FULL Estimate	LMZ PSM Estimate
ADA	Lower	0.002	0.0179	0.0018
RPEG	Lower	30bp	37bp	21bp
ACCY	Higher	0.0315	0.0042	0.0031

Table 2

Precisions of Big N Effect Estimates

This table reports standard error values for Big N quality difference estimates from pre-LMZ studies by Butler et al. (2004), Khurana and Raman (2004), and Behn et al. (2008), LMZ's replications of those analyses (FULL estimates), and LMZ's propensity score matching (PSM) design estimates. The relevant quality proxies being absolute (performance matched in the case of LMZ estimates) discretionary accruals (ADA), implied cost of equity capital (RPEG), and analyst forecast accuracy (ACCY).

Quality Proxy	Pre-LMZ Estimate	LMZ FULL Estimate	LMZ PSM Estimate
ADA	0.0002	0.0021	0.0025
RPEG	10bp	17bp	21bp
ACCY	0.0061	0.0018	0.0020

Table 3

ADA Location

This table reports 99.7%, 95%, and 67% confidence interval identifications of the Big N ADA effect's location for the Butler et al., LMZ FULL replication of Butler et al., and LMZ PSM designs.

Analysis	99.7% Confidence Interval		95% Confidence Interval		67% Confidence Interval	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound
BKL	0.0014	0.0026	0.0016	0.0024	0.0018	0.0022
LMZ FULL	0.0116	0.0242	0.0138	0.022	0.0158	0.0200
Overlap of BKL	0%		0%		0%	
LMZ PSM	-0.0057	0.0093	-0.0031	0.0067	-0.0007	0.0043
Overlap of BKL	100%		100%		100%	
Overlap of FULL	0%		0%		0%	

Table 4

RPEG Big N Effect Location

This table reports 99.7%, 95%, and 67% confidence interval identifications of the Big N RPEG effect's location for the Khurana and Raman (KR), LMZ FULL replication of KR, and LMZ PSM designs.

Analysis	99.7% Confidence Interval		95% Confidence Interval		67% Confidence Interval	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound
KR	0 bp	60 bp	11bp	49 bp	20 bp	40 bp
LMZ FULL	-14 bp	88 bp	46 bp	70 bp	20 bp	54 bp
Overlap of KR	100%		100%		100%	
LMZ PSM	-42 bp	84 bp	-19 bp	61 bp	0 bp	42 bp
Overlap of KR	100%		100%		100%	
Overlap of FULL	96%		87%		66%	

Table 5

ACCY Big N Effect Location

This table reports 99.7%, 95%, and 67% confidence interval identifications of the Big N RPEG effect's location for the Behn et al. (BCK), LMZ FULL replication of BCK, and LMZ PSM designs.

Analysis	99.7% Confidence Interval		95% Confidence Interval		67% Confidence Interval	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound
BCK	0.01333	0.0499	0.0195	0.0437	0.0255	0.0377
LMZ FULL	-0.0008	0.0096	0.0007	0.0077	0.0024	0.0060
Overlap of BCK	0%		0%		0%	
LMZ PSM	-0.0029	0.0091	-0.0080	0.0070	0.0011	0.0051
Overlap of BCK	0%		0%		0%	
Overlap of FULL	95%		89%		76%	

Table 6

Effect Size Assessments

This table reports Big N quality effect partial correlation estimates and associated 95% confidence intervals (CIs) for ADA, RPEG, and ACCY. Panel A reports prior literature based correlations and intervals for analyses by BKL, KR, and BCK. Panel B reports correlations and intervals for LMZ's FULL design replications of these prior studies. Panel C reports correlations and intervals for LMZ's PSM analyses

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

Quality Metric	Point Estimate	LB	UB
ADA	5.85%	4.66%	7.03%
RPEG	2.46%	0.90%	4.02%
ACCY	5.43%	3.39%	7.46%

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

Quality Metric	Point Estimate	LB	UB	Overlap of PRIOR CI
ADA	3.15%	2.42%	3.87%	0%
RPEG	1.35%	0.11%	2.59%	68%
ACCY	1.40%	0.23%	2.57%	0%

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

Quality Metric	Point Estimate	LB	UB	Overlap of PRIOR CI	Overlap of FULL CI
ADA	0.50%	-0.85%	1.85%	0%	0%
RPEG	2.08%	-1.96%	6.13%	100%	100%
ACCY	2.86%	-0.78%	6.51%	77%	100%

Overall Average Partial Correlation: 2.79%