Do Client Characteristics Really Drive the Big N Audit Quality Effect?

Mark DeFond,[*] David H. Erkens,[*] Jieying Zhang[**]

[*]*University of Southern California*

*****University of Texas at Dallas***

November 20, 2014

ABSTRACT

A large auditing literature concludes that Big N auditors provide higher audit quality than non-Big N auditors. An unresolved question, however, is whether self-selection drives this "Big N effect." Recently, a high profile study concludes that Propensity Score Matching (PSM) on client characteristics causes the Big N effect to disappear. We conjecture that this finding may be affected by PSM's sensitivity to its design choices or by the particular set of audit quality measures used in the analysis. To investigate, we examine 3,000 random combinations of three basic PSM design choices and several commonly used audit quality measures. We find that the results are sensitive to the design choices as well as the audit quality measures, with the majority of models finding a Big N effect. We also find a Big N effect using an alternative matching procedure, Coarsened Exact Matching. Overall, our findings suggest that it is premature to conclude that client characteristics drive the Big N effect.

Acknowledgments: This study has benefited from helpful comments by workshop participants at the University of Arizona, Boston College, the Chinese University of Hong Kong, Hong Kong University of Science and Technology, National University of Singapore, Nanyang Technological University, Stanford University, Temple University, the University of Kansas, the University of Rochester, the University of Southern California, and the University of Texas at Dallas.

Do Client Characteristics Really Drive the Big N Audit Quality Effect?

1. Introduction

The auditing literature generally concludes that Big N auditors are associated with higher audit quality than non-Big N auditors, often referred to as the "Big N effect." An unresolved question, however, is whether the Big N effect is driven by the pairing of Big N auditors with higher quality clients. Recent research provides support for this self-selection explanation by finding that Propensity Score Matching (PSM) on observable client characteristics causes the Big N effect to disappear (Lawrence, Minutti-Meza, and Zhang [2011], hereafter, LMZ). This highly cited paper casts serious doubt on the existence of a Big N effect.[1] The absence of a Big N effect not only overturns a large literature, but also questions our basic understanding of the fundamental drivers of audit quality, as captured by the strong incentives and competencies that characterize Big N auditors. The purpose of this study is to reexamine whether client characteristics really drive the Big N effect.

We conjecture that the absence of a Big N effect in LMZ is explained by the inherent sensitivity of PSM to its design choices. PSM requires several design choices, including sampling with or without replacement, the closeness of the match (known as "pruning"), the number of control firms matched to each treatment firm, and the non-linear terms included in the propensity score model. All of these choices affect the composition of the matched sample and hence potentially have large effects on the resulting conclusions. For example, matching with replacement increases the possible matches compared to matching without replacement, and a higher level of pruning reduces the possible matches. However, there is little theoretical guidance in choosing among the alternative designs, making the choices somewhat arbitrary.

[1] LMZ is the second most highly cited of the 71 articles published in *The Accounting Review* in 2011 according to Google Scholar. As of November 17, 2014, LMZ has 152 Google Scholar cites, compared to the average and median of 41 and 30, respectively.

A primary objective of matching is to minimize differences in the matched covariates in order to reduce bias in the resulting estimates (Stuart [2010]). However, several design choice combinations may minimize covariate imbalances equally well, and it is unclear which combination is best. Another objective is to preserve sample size in order to reduce variance in the resulting estimates. However, most PSM design choices that reduce bias also increase variance, and again it is unclear which combination of choices result in the optimal tradeoff. For example, tightening the closeness of the match is expected to lower bias by improving the covariate balance, but increase variance by reducing the sample size. [2] Accordingly, we perform iterative analyses that employ a large set of reasonable design choices, and plot the distribution of the resulting Big N treatment effects. This allows us to observe where the bulk of the evidence lies, and to assess how likely it is to find a Big N effect without subjectively judging which combination of design choices is best.

We also conjecture that the absence of a Big N effect in LMZ is explained by the nature of the audit quality proxies examined by LMZ, which consist of discretionary accruals (DAC), analyst forecast accuracy, and the cost of equity. While DAC is frequently used to measure audit quality, analyst forecast accuracy and the cost of equity are infrequently used, probably because the effect of audit quality on these measures is particularly indirect. Following DeFond and Zhang [2014], we examine five measures of audit quality that are more commonly used in the literature: absolute and signed DAC, restatements, going-concern opinions (GCs), and audit fees. Collectively, these proxies capture both outputs and inputs of the audit process, include both egregious audit failures as well as mild “within GAAP” manipulations, and consist of both discrete and continuous measures. Because they capture complementary dimensions of audit quality, we expect the collective inferences from these proxies to be more informative than the proxies used in LMZ. In addition,

[2] We discuss the bias-variance tradeoff for each of the PSM design choices in more detail in Section 2.

because these proxies are more conventionally used in the literature, they make our results more comparable to the findings in the existing Big N literature.

We begin by replicating LMZ's analysis that uses DAC to measure audit quality, employing their PSM design choices. Consistent with LMZ, we also fail to find a Big N effect. To investigate this finding's sensitivity to PSM's design choices, we repeat the analysis using 3,000 PSM models, each employing a random combination of three basic design choices: (1) the treatment to control ratio (i.e., the number of control firms matched to each treatment firm), (2) the level of pruning (i.e., dropping the worst matches), and (3) the non-linear covariate terms included in the propensity score model. This is analogous to running 3,000 sensitivity tests to gauge the robustness of the results to the study's particular set of design choices. We then draw a density plot of the coefficients on the Big N indicator variable from the 3,000 regressions. We find that the magnitude of the Big N coefficient ranges from −3.2% to +5.0%, consistent with the results being sensitive to PSM's design choices and hence model dependent. We also find that 94.5% of the samples have negative Big N coefficients, and that 77.2% of the samples have a significantly negative Big N coefficient. This suggests that if researchers randomly select a combination of design choices, they would fail to find a Big N effect in only a minority of selections.

While LMZ match without replacement of the control firms, we also examine matching with replacement, because replacement allows closer matching, thus reducing bias in the treatment effect. Matching with replacement finds a negative coefficient on the Big N indicator variable in 99.3% of the samples, and a significantly negative coefficient in 95.0% of the samples. In addition, matching with replacement yields a smaller variance in the estimated Big N coefficients compared to matching without replacement. Because matching with replacement lowers variance in our setting as well as bias, we match with replacement in all subsequent analysis. Overall, while the

PSM model used in LMZ finds that Big N auditors are not associated with lower DAC, the majority of our models find a Big N effect.

We then expand our analysis to examine other commonly used audit quality proxies using a more recent time period, 2003-2009. We use this time period because it allows us to take advantage of the Audit Analytics database and focuses on the current post-SOX regime. As with DAC, we repeat our analysis for each of the additional audit quality proxies using 3,000 randomly selected PSM matched samples. This analysis also finds a wide variation in the magnitude of the estimated Big N effects, consistent with the PSM results also being model dependent for these proxies. Further, the signs and significance of the Big N coefficients generally support the Big N effect in a large majority of samples. Specifically, we find that Big N auditors, on average, are associated with lower DAC, fewer restatements, more frequent GCs, and higher audit fees. Moreover, we find that the percentage of samples supporting a significant Big N effect is 91.0% for DAC, 48.3% for signed DAC, 27.6% for restatements, 71.3% for GCs, and 100% for audit fees.[3] Thus, our analysis suggests that the majority of the design choices we examine are likely to find a Big N effect for a more comprehensive set of audit quality proxies.

We also perform several robustness tests. We first examine the robustness of our results to covariate balances. Restricting our analysis to matched samples chosen from the bias-variance frontier, and to matched samples that have insignificant mean covariate differences, we continue to find a Big N effect. We then examine the robustness of our results to additional covariates included in the propensity score model. After randomly adding five additional covariates that are commonly used in the literature pertaining to each audit quality model, we continue to find a Big N effect.

[3] The results are relatively weaker for the signed DAC compared to the absolute DAC, perhaps because signed DAC combines both income-increasing and income-decreasing accruals. The weak results related to restatement are consistent with the inconclusive literature on whether Big N is associated with fewer restatements (e.g., DeFond and Jiambalvo [1991], Archambeault, Dezoort and Hermanson [2008], DeFond, Lim, and Zang [2012]).

Finally, we examine the robustness of our results to an alternative matching procedure, Coarsened Exact Matching (CEM), and continue to find similar results. Thus, even with its inherent sensitivity to its design choices, PSM appears to be a robust matching technique.

We make two contributions to the literature. Our primary contribution is providing new evidence on the unsettled question of whether self-selection drives the Big N effect. While a recent highly cited study suggests that the Big N effect disappears under PSM, we find that the Big N effect persists under a majority of PSM's research design choices and across several commonly used audit quality proxies.[4] Reexamination of this issue is important because the absence of a Big N effect casts serious doubt on our basic understanding of the incentives and competencies that underpin our understanding of audit quality (Watts and Zimmerman [1981]).

We also contribute to the literature by raising the awareness of the sensitivity of PSM to its design choices. In this vein, our paper is closely related to Lennox, Francis, and Wang [2012], which examines the sensitivity of the Heckman procedure to design choices.[5] We emphasize, however, that our intention is not to criticize the econometrics of PSM, nor do we wish to criticize the application of PSM in the accounting literature generally. The latter would require an analysis of multiple studies and distracts from our primary purpose, which is to bring additional evidence on whether client characteristics explain the Big N effect. Rather, we simply wish to shed light on some overlooked features of PSM, a methodology that is becoming rapidly adopted into the mainstream accounting literature. While we examine one specific setting, albeit an interesting one in which a long-standing conclusion in the literature is overturned by the use of PSM, the randomization and graphical analysis used in our paper can be extended to other settings to systematically assess the existence of model dependency.

[4] We note, however, that our replication of LMZ is limited to their primary DAC analysis and does not examine their other audit quality proxies and various sensitivity tests.

[5] Similarly, Tucker [2010] discusses how PSM differs from the Heckman two-step procedure.

2. Motivation and Background

We begin this section by summarizing the existing evidence on the Big N effect and discussing selection bias as an alternative explanation. We then discuss how PSM may address selection bias, and the sensitivity of PSM to its research design choices. Finally, we discuss the audit quality proxies used in LMZ and motivate the use of a broader set of proxies.

2.1 THE BIG N EFFECT – THEORY AND EVIDENCE

Big N auditors are posited to provide higher audit quality than non-Big N auditors because they have greater incentives and more competence in providing high quality audits (Watts and Zimmerman [1981]). Big N auditors' incentives arise from having more reputation capital to protect (DeAngelo [1981]), higher litigation risk, and greater regulatory scrutiny. In addition, Big N auditors' large customer base makes them less financially dependent on any given client, thus reducing their incentives to compromise their independence by conceding to client demands. Big N auditors are expected to be more competent because their large size allows them to attract and retain higher quality audit inputs, such as human resources and expertise (Dopuch and Simunic [1982]). Further, their larger size means that Big N auditors also enjoy larger economies of scale when compared to smaller auditors, making them more efficient in monitoring the level of audit quality they deliver (Watts and Zimmerman [1981]).[6]

Much of the literature finds evidence consistent with Big N auditors providing higher audit quality than non-Big N auditors. For example, Big N auditors are associated with higher quality audit outputs, such as a lower likelihood of fraud (AAERs) (e.g., Lennox and Pittman [2010]), a higher likelihood of GCs (e.g., Chan and Wu [2011]), lower DACs (e.g., Becker, DeFond, Jiambalvo, and Subramanyam [1998], Francis, Maydew, and Sparks [1999], Kim, Chung, and Firth

[6] For example, to the extent there are fixed costs associated with audit firms' internal controls over audit quality, larger audit firms will have an advantage in producing high quality audits (Watts and Zimmerman [1981]).

[2003]), improved management forecasts (Ball, Jayaraman, Shivakumar [2012]), and timelier 8-K filings (Schwartz and Soo [1996]). Big N auditors are also associated with higher quality audit inputs, such as more audit effort as reflected in higher audit fees (e.g., Ireland and Lennox [2002]). In addition, Big N auditors are perceived by the market to provide higher audit quality, as reflected in increased ERCs (Teoh and Wong [1993]) and lower cost of capital (Khurana and Raman [2004], Pittman and Fortin [2004], Mansi, Maxwell, and Miller [2004]). Taken together, this literature examines a wide spectrum of audit quality proxies that are complementary on many dimensions. The evidence from these studies suggests that Big N auditors are associated with proxies that capture both actual and perceived audit quality, both inputs and outputs from the audit process, audit output measures that are both direct (e.g., fraud, GCs) and indirect (e.g., DACs), and audit failures that are both egregiously large (i.e., restatements) and mildly “within-GAAP” (i.e., DAC). Thus, this research provides broad-based and consistent evidence that Big N auditors deliver higher quality audits than non-Big N auditors.

2.2 SELECTION BIAS AS AN ALTERNATIVE EXPLANATION

The selection bias argument suggests that the Big N effect is not due to the quality of the audits delivered by Big N auditors, but is instead due to the inherently high financial reporting quality of Big N clients. Selection bias is an appealing alternative explanation for the Big N effect because Big N auditors have incentives to choose low-risk clients to protect their reputation, lower litigation risk, and reduce regulatory scrutiny. This is consistent with empirical evidence that Big N auditors tend to have less risky clients (e.g., Raghunandan and Rama [1999], Johnstone [2000], Johnstone and Bedard [2004]). The potential for selection bias explaining the Big N effect is exacerbated in studies that use financial reporting quality proxies to capture audit quality, since these measures are a joint product of both audit quality and client firms' innate characteristics.[7] While Big N researchers typically control for client risk characteristics in an OLS setting, the associations between Big N and audit quality may still be spurious if the association between client risk and audit quality is non-linear or not identical for Big N and non-Big N clients.

Several observations, however, challenge the selection bias explanation. First, Big N auditors' higher competency to deliver higher audit quality should persist even for low-risk clients. Holding the client characteristics constant, Big N auditors' higher competency is likely to result in a higher quality audit when compared to smaller auditors. Second, it is difficult to explain why Big N auditors are able to charge a fee premium, an association that is pervasive in the literature, if their clients have inherently higher financial reporting quality.[8] Clients with inherently higher financial reporting quality should reduce the risk of misreporting, thereby reducing, not increasing, audit fees.

It is also notable that selection bias has long been recognized as a threat in the Big N literature, and researchers have used a variety of methods to address this threat. The vast majority of studies find that the Big N effect persists even after applying statistical techniques such as Heckman's two-stage procedures (e.g., Weber and Willenborg [2003], Khurana and Raman [2004], Behn, Choi, and Kang [2008]), two-stage least squares (e.g., Guedhami and Pittman [2006]), two-stage treatment effects models (e.g., Kim, Chung, and Firth [2003]), general matching procedures (e.g., Doogar and Easley [1998]), and change analyses (e.g., Teoh and Wong [1993], Schwartz and Soo [1996], Wang, Wong, and Xia [2008]).[9] A criticism of this literature, however, is that these methods are not always properly implemented (Lennox, Francis, and Wang [2012]). Importantly, a

[7] See DeFond and Zhang [2014] for a more detailed discussion on how audit quality is a function of firms' innate characteristics and reporting systems.

[8] While the premium can derive from a signaling effect, the signaling effect is unlikely to be sustainable if Big N auditors do not actually improve clients' financial reporting quality.

[9] Lennox, Francis, and Wang [2012] provide a thorough discussion on the use of the Heckman procedure to control for selection bias.

recent highly cited paper, LMZ, concludes that propensity score matching on client characteristics eliminates the Big N effect, casting serious doubt on the veracity of the Big N effect.

The results in LMZ are appealing because they support the long-standing concern that selection bias drives the Big N effect, and because they are based on PSM, a statistical technique that has recently gained popularity. However, given the theory and evidence that supports a Big N effect, LMZ's results are also unsettling. Because Big N auditors have starkly stronger incentives and competencies to provide higher quality audits, the bulk of the supply-side audit quality literature over the past thirty plus years is focused on understanding the role of auditors' incentives and competencies in explaining audit quality differentiation. The Big N effect provides the underpinnings for the notion of audit quality differentiation, and has led to the discovery of a variety of factors that affect differentiated audit quality, such as auditor specialization (e.g., Francis, Reichelt, and Wang [2005]), auditor office size (e.g., Francis and Yu [2009]), and individual audit partner quality (e.g., Gul, Wu, and Yang [2013]). The absence of a Big N effect shakes the foundation of this literature by casting doubt on our basic understanding of the fundamental drivers of audit quality. Thus, the purpose of this paper is to re-examine the Big N effect after considering the effects of PSM's design choices and alternative audit quality measures.

2.3 THE USE OF PSM IN ADDRESSING SELECTION BIAS – ADVANTAGES, LIMITATIONS, AND REMEDIES

The underlying idea of PSM is to match Big N clients to non-Big N clients based on a propensity score derived from a set of observable client characteristics. The goal is to minimize the differences in the matching characteristics (referred to as "covariate balancing"), so that these differences do not explain the potential difference in audit quality. Compared to OLS models, PSM is a more robust approach for achieving an unbiased estimate of the treatment effect because it does not require researchers to assume the functional form of the relationship between the outcome and control variables (Armstrong et al. [2010]).

A major limitation of PSM, however, is that it requires several subjective research design choices that affect the choice of matched firms and hence the conclusions from its analysis (King et al. [2011]). These choices include the number of control firms matched to each treatment firm, the closeness of the match, the non-linear terms included in the propensity score construction, and the replacement decision.[10] Because changes in these design choices potentially change the matched observations included in the final matched sample, they are likely to affect the inferences drawn from the PSM analysis. For example, a treatment to control ratio of one-to-three results in a larger sample than a ratio of one-to-one; a higher level of pruning increases the match closeness but reduces the sample size; and adding non-linear terms alters the composition of the matched firms by changing the propensity scores. PSM's sensitivity to these design choices potentially generates a wide range of possibly conflicting inferences. As a result, definitive inferences cannot be drawn until they are "shown to be adequately insensitive to the choice of assumptions" (Leamer [1983]).[11]

Because the number of alternative specifications under PSM is quite large, we employ an iterative technique that randomly assigns several thousand different sets of research design choices and plots the distribution of the resulting Big N treatment effects. This iteration analysis allows us to systematically observe the distribution of the treatment effect assuming a spectrum of possible combinations of research design choices without subjectively judging which particular set of choices is best. Thus, inferences from the distribution of our iterations should be more robust compared to relying on a given set of design choices.

[10] Because researchers often seek to balance the covariates in various ways, these attempts also introduce subjectivity in the research design.

[11] Along these same lines, Leamer [1983] also notes "an inference is not believable if it is fragile, if it can be reversed by minor changes in assumptions. As consumers of research, we correctly reserve judgment on an inference until it stands up to a study of fragility, usually by other researchers advocating opposite opinions."

Our iteration analysis focuses on four PSM design choices. First, we vary the treatment to control ratio (i.e., the number of control firms matched to each treatment firm). While matching more control firms to each treatment firm generally increases bias in the treatment effect, because the second and third closest matched control firms are more distant from the treatment firm, it also reduces variance in the treatment effect by increasing the sample size. Thus, the treatment to control ratio reflects a tradeoff between bias and variance. We vary this choice by alternating between one-to-one, one-to-two, and one-to-three matches.

Second, we vary the closeness of the match. While closer matches reduce bias in the treatment effect by minimizing the propensity score differences between the treatment and control firms, they also increase variance in the treatment effect by reducing the sample size. Thus, the closeness of the match also involves a bias-variance tradeoff. Empirically, the closeness of the match is determined by the chosen caliper, which refers to the maximum allowed difference in the propensity score of the treatment and control firms. While LMZ adopt a caliper difference of 0.03, we use pruning, which incorporates a wide range of caliper choices. Specifically, for each matched sample we rank treatment firms based on their overall match quality (difference in propensity score between the treatment and control firms) and prune a random number of the worst matches.[12]

Third, we vary the choice of non-linear terms included in the propensity score model. Including non-linear terms, such as squares or cubes of covariates, and the simple interactions between covariates, potentially reduces multivariate distributional differences between the covariates of the treatment and control firms. However, including non-linear terms may also increase the covariate imbalance by increasing the number of covariates summarized in the

[12] This procedure results in pruning levels that range from 0% to over 99% of our treatment firms. In untabulated analysis we examined 55 PSM studies published in the top three accounting journals between 2006 and 2014, and find pruning levels ranging from 0% to 96%. The pruning level in LMZ is approximately 82.3%.

propensity score (King et al. [2011]).[13] Thus far, relatively little attention has been paid to these non-linear terms in the accounting literature, since most studies focus primarily on minimizing differences in the means of the covariates instead of the differences in their multivariate distributions.[14] We vary this choice by including random combinations drawn from the set of possible non-linear terms.

Finally, we examine the effect of matching with replacement on the sensitivity of the PSM results. Matching with replacement reduces the bias in the treatment effect, because each treatment firm can be matched to the closest control firm even if that control firm is already matched to another treatment firm.[15] Matching with replacement may also affect the variance, although the direction is difficult to predict. Matching with replacement can increase the variance when an idiosyncratic control firm is used repeatedly. However, it can decrease the variance by increasing the number of matched pairs to be included in the analysis. Matching with replacement is most likely to increase the number of matched pairs when the treatment firms outnumber the control firms, as in our setting. Because replacement is a dichotomous choice, we evaluate the separate effect of replacement on the PSM results by conducting our iterations of the preceding three design choices with and without replacement.

2.4 AUDIT QUALITY PROXIES

LMZ examine three audit quality measures: DAC, analyst forecast accuracy, and cost of equity. While DAC is a commonly used proxy for audit quality, analyst forecast accuracy and the cost of equity are not commonly used, probably because the effects of audit quality on these proxies

[13] As the number of covariates increases, the propensity score tends to perform progressively worse, because balancing on one covariate may lead to imbalances on other covariates.

[14] LMZ address non-linearity by performing additional tests that employ single variable matching on each of the variables that are nonlinear to both Big N and the audit quality proxy.

[15] In addition, while matching without replacement can be sensitive to the order in which control firms are drawn (because the order determines which control firms are available for subsequent matches), matching with replacement is not sensitive to the order.

is less direct when compared with the other commonly used measures. Following DeFond and Zhang [2014] we use five audit quality measures, each of which captures complementary dimensions of audit quality: restatements, GCs, absolute and signed DACs, and audit fees. Compared to the other proxies, restatements are a relatively direct output-based measure of audit quality because they indicate the auditor signed off on materially misstated financial statements. GC modified audit opinions are also a relatively direct output-based measure of audit quality because failure to report a GC when one is warranted means the auditor issued the wrong audit opinion, an output of the audit process directly under the auditor's control. However, a limitation of restatements and GCs is that they do not capture subtle variations in quality, and occur relatively rarely. In contrast to these measures, DAC is a relatively less direct output-based audit quality measure, because the auditor's influence on accruals is likely to be more limited than his or her influence over preventing material misstatements or issuing a GC opinion. DAC is used to capture evidence of opportunistic earnings management, and its use as an audit quality proxy rests on the assumption that high quality auditing constrains earnings management. We use both absolute and signed DAC because it is unclear which one is a better proxy for audit quality (Lennox, Wu, and Zhang [2014]). Audit fees are used to proxy for audit quality because they capture the auditor's effort level, which is an input of the audit process that is linked to higher audit quality (Caramanis and Lennox [2008]). Both DAC and audit fees are continuous measures that potentially capture subtle variations in audit quality. However, DAC is subject to relatively high measurement error (Kothari, Leone, and Wasley [2005]), and audit fees cannot be unambiguously interpreted as audit quality because they are potentially confounded by fee premia and audit efficiency.

Overall, these five audit quality proxies capture both outputs and inputs of the audit process, represent both relatively direct and indirect measures of audit quality, capture both egregious audit failures as well as mild “within GAAP” manipulations, and consist of both discrete and continuous measures. As a result, they complement one another other in portraying a comprehensive picture of audit quality.

3. Main Analysis

3.1 REPLICATING LMZ

We begin by replicating the DAC analysis in LMZ, which covers the period 1988-2006.[16] The descriptive statistics reported in Panel A of Table 1 are generally consistent with those in Table 1 of LMZ. Following LMZ, we first estimate the propensity score using a logistic model that regresses a Big N dummy on five client characteristics (size, ROA, leverage, current ratio, and asset turnover). Panel B of Table 1 presents the regression results. We find that the clients of Big N auditors, when compared to the clients of non-Big N auditors, are significantly larger (as captured by both assets and market capitalization) and more profitable (as captured by ROA). Big N clients also have lower asset turnover, lower leverage, and smaller current ratios. These significant differences suggest that selection bias may exist, motivating the use of matching methods. We also examine the accuracy of the propensity score model in discriminating between Big N and non-Big N observations, using the area under the ROC (Receiver Operating Characteristic) curve (hereafter, AUC) (Hosmer and Lemeshow [2000]). We find that the propensity score model has an AUC of 0.858, which is above the acceptable threshold of 0.7 used in prior studies. Thus, this model seems to discriminate well between Big N and non-Big N observations.

[16] Our sample is larger than LMZ’s because we only delete missing observations for variables that are used in our analysis. We also winsorize the variables used in the analysis (e.g., ROA), while LMZ winsorize the raw variables (e.g., they separately winsorize net income and total assets before computing ROA). In addition, following Kothari et al. [2005], we require at least 10 observations in an industry to calculate discretionary accruals and drop observations with absolute total accruals larger than total assets, while LMZ do not impose these requirements. We thank the authors of LMZ for sharing their programs, which allowed us to identify these differences.

Following LMZ, we match one Big N client to one non-Big N client without replacement, and retain matches that are within a caliper width of 0.03. Panel C of Table 1 reports three covariate balance metrics. The first is the simple mean difference of each of the five covariates, also reported by LMZ, which suggests that matching narrows the differences between Big N and non-Big N clients to a large extent. For example, while size remains significantly different between Big N and non-Big N clients, the difference drops from 2.551 before matching to -0.213 after matching. The second covariate balance metric we examine is the absolute standardized percentage bias, |%bias|, which scales the mean difference by the square root of the average treatment and control sample variances (Rosenbaum and Rubin [1985]). We find that size and ROA have the highest absolute standardized percentage bias, consistent with significant differences remaining between the treatment and control samples for these variables.

The third covariate balance metric we examine is L_1, the difference between the histogram of covariates, calculated as the absolute difference between the relative empirical frequency of the treated observations for a particular bin minus the relative empirical frequency of the control observations in that bin, summed over all bins (Iacus et al. [2011]). Details on the calculation are provided in Table 1. The advantage of this measure is that it takes all dimensions of covariates' distribution into account, not just the first moment of their univariate distributions, which is what our first two traditionally used imbalance metrics do. L_1 varies between 0 and 1, where 0 indicates identical treatment and control distributions, and 1 indicates no overlap in the distributions. Comparing L_1 for each individual covariate indicates that it is closer to 0 than to 1, consistent with the treatment and control firms having similar univariate distributions. However, when we compare the multivariate histogram of the treatment and control firms, we find L_1 equals 0.597, suggesting that the treatment and control firms have very different multivariate distributions. This is because balance in the multivariate distribution requires a similar distribution across all covariates, and thus is always harder to achieve. As a result, multivariate L_1 is not directly comparable to univariate L_1, but comparable to other multivariate L_1 from alternative PSM specifications. Overall, the imbalance metrics suggest a reasonably high match quality.

We replicate the DAC result in LMZ by regressing the absolute value of discretionary accruals on the Big N dummy and the five matching variables as control variables. Consistent with LMZ, we find an insignificant coefficient on Big N of –0.002, indicating that Big N auditors are not associated with lower absolute discretionary accruals after matching on client characteristics.

3.2 PSM SENSITIVITY TO RESEARCH DESIGN CHOICES

We evaluate the sensitivity of the Big N effect to common PSM research design choices by examining the results of 3,000 matched samples constructed as follows. We first take 3,000 random draws from the full set of the nonlinear terms from the five covariates (i.e., five squared terms, five cubed terms, and ten simple interactions). We restrict our analysis to the five covariates used in LMZ because these are commonly used determinants of Big N (DeFond and Zhang [2014]), and because it makes our findings easier to compare with LMZ's.[17] We then add the randomly selected non-linear terms to the five primary covariates, and estimate the propensity score model. We then match one, two or three control firms to each treatment firm without replacement, randomly allocating them proportionally across the 3,000 PSM models. For each of the resulting 3,000 matched samples we then prune a random number of the treatment firms that have the worst overall match based on their propensity score differences, resulting in pruning levels that range from 0% to over 99% of our treatment firms. This procedure results in 3,000 matched samples that reflect random variations in the treatment-to-control ratio, match closeness, and non-linear terms. To

[17] In section 4.2 we relax this restriction and examine the effect of varying the covariates included in the propensity score model.

ensure that our matched samples have sufficient statistical power to detect a Big N effect if one exists, we follow prior literature by using Cohen's *d* to assure our sample is large enough to detect a moderate effect size with a 20% chance of making a Type II error.[18, 19]

We then regress absolute DAC on the Big N Dummy and the five covariates as control variables for the 3,000 PSM matched samples. We report summary statistics of the estimated Big N treatment effect from the 3,000 regressions in the first row of Table 2, and report a density plot of the estimated Big N treatment effect using a solid line in Figure 1. The horizontal axis in Figure 1 represents the value of the Big N coefficients we obtain from the 3,000 regressions, and the vertical axis represents the corresponding density values.

Table 2 and Figure 1 indicate that the estimated Big N treatment effect (i.e., the coefficient on the Big N Dummy) ranges from -0.032 to 0.050. This is consistent with PSM being sensitive to the research design choices we vary in our analysis. However, the large majority of the coefficients have negative values, consistent with Big N being associated with smaller absolute DAC under most of the research design choices likely to be made using PSM. The first row of Table 2 reports that for matches without replacement, the mean and median treatment effect are respectively -0.008 and -0.009. Table 2 also reports PCT Sign, which captures the percentage of negative Big N coefficients, which equals 94.5% for matches without replacement. This indicates that for 3,000 randomly selected design choices, 94.5% are in the direction that supports the Big N effect. Table 2 also reports PCT Significant, which indicates the percentage of significantly negative Big N coefficients, which equals 77.2%. This suggests that if a researcher were to randomly select a single design choice, there is a 77.2% chance of finding a significant Big N effect.

[18] Other studies that use Cohen's *d* to assess sample power include Kim et al. [2012], Libby et al. [2004], Kadous et al. [2011], Cheng et al. [2012], Huelsbeck et al. [2011], Naiker et al. [2012], and Glover et al. [1997].

[19] An alternative way to assure power is to require matched samples to retain at least 1% of the treatment firms. Using this alternative power restriction, we find qualitatively similar results.

We note that we do not use the t-statistics from the 3,000 Big N coefficients to impute the significance level of the estimated Big N effect. Sample dependency across the matched samples prevents us from doing so. Rather, we graphically portray the Big N coefficients in order to evaluate where the bulk of the evidence lies. We also use the proportion of significant t-statistics to infer the chances of researchers finding a Big N effect had they randomly chosen a set of design choices. Basing our inferences on the collective evidence from 3,000 samples is analogous to running 3,000 robustness tests that assess whether the results are model dependent, or whether they are driven by the underlying data. A large proportion of significant Big N coefficients that are consistent with a Big N effect suggests the effect exists, and a large variation in the Big N coefficients suggests the PSM results are model dependent.

We also examine the replacement decision. The second row of Table 2 reports the results of replicating LMZ with replacement, with the Big N treatment effects plotted using a dotted line in Figure 1.[20] Table 2 shows that the mean and median Big N treatment effect is -0.005. PCT Sign, the percentage of coefficients that are in the direction of the Big N effect, increases to 99.3% with replacement. In addition, the percentage of significantly negative coefficients on Big N increases to 95.0% with replacement. Moreover, Figure 1 and Table 2 show that the coefficient distribution for matches with replacement is tighter. For example, matching with replacement reduces the range of coefficients to 0.048, compared to 0.082 for matching without replacement. [21] Consistent with the smaller range, the standard deviation for matches with replacement is 0.001, substantially lower than 0.005, the standard deviation for matches without replacement. This suggests that matching with replacement reduces the variance of the treatment effect by increasing the sample size when

[20] The number of treatment firms retained ranges from 57 to 19,544 for matching without replacement and from 65 to 78,698 for matching with replacement.

[21] We infer the range from the minimum and maximum value of the Big N treatment effects. For example, the range from matching with replacement is the distance between the minimum (-0.032) and the maximum (0.016).

the number of treatment firms significantly outnumbers the number of control firms. Because matching with replacement lowers both bias and variance in our setting, we match with replacement in all of our remaining analyses.

Overall, Figure 1 and Table 2 indicate that both the sign and significance of the Big N coefficients support the Big N effect in a majority of the PSM matched samples. The large variation in the Big N coefficient suggests the results from PSM are model dependent. Thus, while LMZ concludes that Big N auditors are not associated with lower DAC using PSM, this conclusion appears to arise only under a minority of design choices.

3.3 ANALYSIS OF FOUR ADDITIONAL AUDIT QUALITY PROXIES

To explore whether the absence of a Big N effect in LMZ is explained by the choice of audit quality proxies they examine, we examine four additional commonly used audit quality proxies: signed DAC, earnings restatements, GCs, and audit fees. We also restrict our analysis to 2003-2009 to take advantage of the expanded set of auditing variables available in the Audit Analytics database (e.g., the identity of auditors, audit fees, GCs, and restatements), and to focus on the current post-SOX regime.[22] Because we examine a different time period, we also re-examine the absolute DAC measure used in LMZ, bringing the number of audit quality proxies we examine to five. We expect Audit Analytics to capture more accurate information than COMPUSTAT on the auditor's identity and audit opinion, given that it caters to auditing-related research. We use COMPUSTAT data to measure discretionary accruals and client characteristics, and CRSP data to compute return-related control variables. We end the sample in 2009 to allow time for restatements to be discovered and reported.

Because each audit quality proxy and the control variables used in modeling those proxies impose different data requirements, we have three samples. Audit fees and its control variables

[22] AuditAnalytics does not begin coverage until 2000, and the coverage in its initial years is limited.

impose the least data requirements, yielding the largest sample with 19,984 observations. The sample for DAC (both absolute and signed) and restatements reduces to 14,491 observations. The sample for GCs further reduces to 5,088 observations by including only financially distressed firms defined by either negative income or negative operating cash flows (DeFond et al. [2002]).

Appendix A provides definitions of all the variables used in our analysis, and Appendix B reports descriptive statistics for the three samples. For each sample, we first present the descriptive statistics on the audit quality proxies. Panel A of Appendix B reports that clients of Big N auditors on average have lower absolute and signed discretionary accruals. This is consistent with the prior findings on Big N auditors constraining their clients' discretion in reporting accruals. We do not find a difference in the incidence of restatements between Big N and non-Big N clients. This may be due to the absence of controlling for restatement risk in making univariate comparisons, because Big N clients are likely to have lower restatement risk than non-Big N clients prior to being audited. Panel B of Appendix B also does not find a difference in the frequency of GCs between distressed Big N and non-Big N clients, which may be due to the absence of controlling for client characteristics associated with GCs. Panel C of Appendix B, however, indicates that Big N auditors charge higher audit fees than non-Big N auditors, although this could reflect client size effects.

Descriptive statistics on the matching variables used to identify our matched control firms are reported in Appendix B. Panel A shows that Big N clients are significantly different from the non-Big N clients on all five matching variables. Specifically, Big N clients have larger size and higher ROA, but lower current ratio, lower asset turnover ratio, and higher leverage. The comparisons in Panel B and C are similar. Although these contrasts do not suggest that Big N clients are necessarily less risky, they are consistent with a selection bias.

For each of our five audit quality proxies, we follow the same iteration procedure to create 3,000 matched samples with a random combination of treatment to control ratios, match closeness, and nonlinear terms, where all matching is done with replacement. We then regress each audit quality proxy on the Big N dummy variable and the control variables typically used in the literature for the respective model. We rerun the absolute DAC regressions in Figure 1 after including the control variables used in Francis and Michas [2013], and include these same variables in our analysis of restatements.[23] For GC, we use the control variables from DeFond et al. [2002], and for audit fees we use the control variables from Chaney et al. [2004] and Francis et al. [2005]. These control variables are defined in Appendix A, and their descriptive statistics are provided in Appendix B in the panel that pertains to their respective audit quality proxies.

Figure 2 plots the estimated Big N treatment effects for the five audit quality proxies. We find a wide variation in the magnitude of the Big N coefficients (or average marginal effects for dichotomous audit quality proxies), consistent with the PSM results being sensitive to the design choice, and hence model dependent. Table 3 provides statistics on the distribution of the estimated Big N effects, and reports that the mean and median estimated Big N effects are negative for DAC (both absolute and signed) and restatements; and positive for GCs and audit fees.[24] This indicates that on average, Big N auditors are associated with lower DAC, fewer restatements, more frequent GCs, and higher audit fees. PCT Sign ranges from 94.3% for signed DAC to 100% for audit fees, indicating that a majority of the research design choices are also in a direction that supports the Big N effect. PCT Significant is 91.0% for absolute DAC, 48.3% for signed DAC, 27.6% for restatements, 71.3% for GCs, and 100% for audit fees. Thus, the statistical significance of the

[23] For restatements, we further include absolute DAC as a control variable because prior literature documents a relation between large accruals and restatements (Jones, Krishnan, and Melendrez [2008]).

[24] The number of treatment firms retained ranges from 53 to 10,942 for absolute DAC, from 51 to 10,938 for signed DAC, from 129 to 10,944 for restatements, from 99 to 3,368 for GCs, and from 53 to 13,155 for audit fees.

majority of the Big N coefficients supports a Big N effect. We note that the results are relatively weaker for the signed DAC compared to the absolute DAC. This may be because signed DAC combines both income-increasing and income-decreasing accruals, and the reversal of past earnings management. We also note that direct evidence on the inverse association between Big N and restatements is absent in the literature.[25] Further, studies that include a Big N control variable in modeling restatements find only weak evidence that Big N auditors, on average, are associated with fewer restatements (DeFond and Jiambalvo [1991], Archambeault, Dezoort and Hermanson [2008], DeFond, Lim, and Zang [2012]). While Francis, Michas and Yu [2014] find that Big N auditors are associated with fewer restatements, the association is only found among the largest quartile of auditor offices. Therefore, it is not that surprising to find weaker evidence supporting a Big N effect for restatements. In summary, our analysis of a more comprehensive set of audit quality proxies during the post-SOX period finds that the lack of a Big N effect in LMZ can be explained by both PSM's sensitivity to its design choices and the choice of audit quality measures, with the majority of design choices finding a Big N effect for the commonly used audit quality measures.

3.4 MATCH QUALITY

An advantage of analyzing all 3,000 matched samples is that it takes into account a wide variety of design choices. A disadvantage, however, is that some of the matched samples may be poorly balanced, which may bias towards finding a Big N effect. Thus, in this section we examine the matched samples that have balanced covariates. We adopt two approaches in this analysis. The first approach considers the optimal set of design choices from a bias-variance tradeoff perspective. As discussed previously, there is a tradeoff between the covariate balance and the matched sample size. Given this tradeoff, for a given sample size researchers should select matches with the lowest

[25] However, Lennox and Pittman [2010] find that Big N auditors are associated with fewer frauds using PSM, consistent with the Big N effect.

imbalance (which minimizes bias); and for a given degree of imbalance researchers should select matches with the largest sample size (which minimizes variance). These solutions are located on the bias-variance frontier and dominate all other matching solutions in terms of the bias-variance tradeoff. As such, they represent the optimal design choices (King et al. [2011]). Figure 3 graphically portrays the efficient frontier using the matched samples generated for the absolute DAC analysis. From the 3,000 matched samples, we create 100 percentile-ranked groups based on the number of treatment firms in each matched sample. For each group, the sample with the lowest L_1 imbalance metric is on the bias-variance frontier, as represented by the "+" in Figure 3.[26] We then repeat our analysis for each of the 100 samples, and Table 4 Panel A presents the statistics on the Big N treatment effects for our five proxies. The results indicate that across all audit quality proxies the direction (PCT Sign) and statistical significance of the Big N treatment effects (PCT Significant) generally support a Big N effect.[27]

In the second approach we examine two samples with insignificant differences in the means of the covariates (as opposed to the lowest multivariate metric L_1). We examine the differences in the means because while the multivariate metric L_1 is theoretically the best measure of imbalance, it is not commonly used in the accounting literature. In addition, because LMZ find that size is the primary driver of the observed Big N effect, we separately examine a sample with insignificant differences in the mean of size, and a sample with insignificant differences in the mean of all five covariates. Thus, we begin by rerunning our iteration analysis to generate another 3,000 matched samples that not only reflect a random combination of three design choices but also have

[26] Note that for all of these tests we continue to require the matched sample sizes to be large enough to detect a moderate effect size with a power of 0.80. Thus, we exclude matched samples that have insufficient statistical power to detect moderate effect sizes.

[27] In an untabulated analysis we compare the L_1 imbalance score that results from the PSM specification used in LMZ (Table 1) with the L_1 imbalance scores that result from the PSM specifications reported in Figure 1. We find that LMZ's specification results in greater imbalance than the majority of PSM specifications in Figure 1 (82.5% of matching without and 98.5% of matching with replacement PSM specifications). Thus, LMZ's specification results in a relatively poor covariate balance.

insignificant difference in firm size. Panel B of Table 4 reports these results. We find the mean/median and PCT Sign to be very similar to that in Table 3. We also find slight improvement in PCT Significant for restatements and GC using the balanced samples. We then examine another 3,000 matched samples with all five covariates balanced. Panel C of Table 4 reports that the results are again comparable to those in Table 3, with a small improvement in PCT Significant for restatements (from 0.276 in Table 3 to 0.350 in Table 4 Panel C), but a small deterioration in PCT Significant for both DAC measures. Overall, our analysis of balanced match samples suggests that the Big N effect we find in our primary analysis is not driven by covariate imbalance.

4. Additional Analyses

In this section, we first evaluate the sensitivity of PSM results to each of the individual design choices we examine in our main analysis. Then we examine whether our results are sensitive to the use of the five LMZ covariates by adding additional covariates. Finally, we introduce a new matching method, Coarsened Exact Matching (CEM), to corroborate our PSM analysis.

4.1 EXAMINING THE PSM DESIGN CHOICES INDIVIDUALLY

A natural question from examining the plots in Figures 1 and 2 is whether PSM is relatively more sensitive to any of the three design choices we vary. Thus, we separately examine the effect of the treatment to control ratio, nonlinear terms, and match closeness. We first fix the match closeness by pruning a constant 3% of the treatment firms with the worst overall match quality based on the propensity score. Since the closeness of the match is fixed, the variation in the estimated Big N coefficients is caused only by the treatment to control ratio and the nonlinear terms. We plot the resulting Big N treatment effects in Figure 4 using dotted lines. We find that the treatment to control ratio and nonlinear terms generate a relatively small variation in the estimated coefficients, as indicated by the tight distribution across all five audit quality proxies.

We then hold the non-linear terms constant by excluding them from the construction of the propensity score model. Although including non-linear terms potentially improves balance in the multivariate covariate distribution, it is not uncommon in the accounting literature to exclude them. We plot the resulting Big N treatment effects in Figure 4 using solid lines. Since non-linear terms are excluded, the variation in the estimated Big N effects is caused only by the closeness of the matches and the treatment to control ratio. We find that varying the match closeness and the treatment to control ratio generates a relatively large variation in the estimated Big N effects, as indicated by the fat tails across all five audit quality proxies.

Finally, we examine the separate effect of the treatment to control ratio. We separately plot the estimated Big N effects for one-to-one, one-to-two, and one-to-three matches in the second part of Figure 4. In each plot, we randomly vary the pruning level and the nonlinear terms. Figure 4 indicates that PSM is not very sensitive to variation in the treatment to control ratio. From this and our replacement analysis in Table 2 and Figure 1, we conclude that while each design choice contributes to the sensitivity of PSM, the replacement decision and match closeness create higher sensitivity than the non-linear terms and treatment-to-control ratio.

4.2 ADDING COVARIATES TO THE BIG N SELECTION MODEL

We restrict our main analysis to the five covariates used in LMZ to make our findings more comparable with LMZ’s. However, we acknowledge that the PSM results could also be sensitive to the choice of the first-stage matching variables. To examine this, we add additional matching variables selected from the set of control variables for each audit quality proxy used in the second stage. To gauge the sensitivity of the results to these additional variables, we randomly add five variables to the propensity score construction. Figure 5 and Table 5 report the results from this analysis. Overall we find similar results to those in Figure 2 and Table 3, with slightly higher PCT Significant for Signed DAC. This analysis suggests that adding more matching variables does not drive away the Big N effect.

4.3 USING AN ALTERNATIVE MATCHING PROCEDURE

We also repeat our analysis using an alternative matching procedure, Coarsened Exact Matching (CEM). CEM is an adapted application of conventional exact matching, that matches control firms with treatment firms based on ranges (or strata), rather than exact values of the covariates. By stratifying covariates, CEM alleviates the significant demands that exact matching imposes on the data. CEM also directly matches on the multivariate distributions of the covariates instead of matching on a single scalar (i.e., propensity score). As a result, CEM does not rely on the functional form and discriminative ability of a first-stage propensity score regression, and considers higher moments of the covariate distributions (King et al. [2011]). [28]

The key design choice in CEM is choosing a coarsening range for each covariate. Thus, we repeat the matching to generate 3,000 matched samples, each reflecting a different coarsening choice. To search over a large number of CEM solutions, we sample from the set of possible coarsenings equally spaced between minimum and maximum levels. To ensure that we retain a reasonable number of treatment firms we set the minimum number of bins equal to two and the maximum equal to the number of bins suggested by Doane's formula [1976].[29] Similar to PSM we

[28] Gary King's website provides Stata programs and other software facilitating the implementation of CEM: http://gking.harvard.edu/cem

[29] The Doane [1976] formula specifies the maximum number of bins as $1 + log_2(n) + log_2\left(1 + \frac{|Skewness|}{\sqrt{\frac{6(n-2)}{(n+1)(n+3)}}}\right)$, where n is the number of observations in the unmatched sample, and *skewness* is the sample skewness of the respective covariate.

require the resulting matched samples to be large enough to detect a moderate effect size with a 20% chance of making a Type II error.

We examine the Big N treatment effects using CEM plotted in dotted lines in Figure 6. The solid lines in Figure 6 are replications of PSM plots in Figure 2 for comparison. Figure 6 suggests that CEM yields a smaller variation in the Big N effect compared to PSM across all five proxies. Table 6 reports that under CEM, PCT Sign ranges from 90.7% for signed DAC to 100% for GCs and audit fees; and PCT Significant ranges from 32.8% for signed DAC to 100% for audit fees. Notably, PCT Significant for restatements improves from 27.6% using PSM to 75.6% using CEM. Thus, using CEM, we continue to find that Big N auditors provide higher quality audits than non-Big N auditors.[30] Thus, regardless of its sensitivity to design choices, PSM still provides robust inferences in our setting.[31]

5. Conclusion

We investigate whether client characteristics drive the Big N effect by randomizing a set of common PSM design choices and plotting the distribution of the estimated Big N coefficients for five commonly used audit quality proxies. We find a wide variation in the magnitude of the estimated Big N effects, consistent with the PSM results being model dependent for these proxies. We also find that the signs and significance of the Big N treatment effects support the Big N effect in a large majority of samples. Further, our results are not driven by poor covariate balance or the choice of covariates we match on. While each design choice contributes to the sensitivity of PSM,

[30] We caution, however, CEM has its limitations. In particular, as the number of covariates grows, lack of available matches may make CEM infeasible.

[31] In an untabulated robustness analysis, we also impose a common support requirement (e.g., Caliendo and Kopeinig [2008], Chan, Chen, Chen, and Yu [2014]). That is, we exclude treatment firms that have propensity scores that are higher (or lower) than the maximum (minimum) propensity scores across all the control firms. These treatment firms are unlikely to have high quality matches because they are far away from the common support area. We find the results are qualitatively similar after we impose this requirement.

we find that the replacement decision and match closeness create higher sensitivity than the treatment-to-control ratio and the non-linear terms. Our results are also robust to an alternative matching procedure, CEM. Therefore, our evidence suggests that it is premature to conclude that client characteristics drive the Big N effect.

Our study contributes to the auditing literature by providing new evidence on the unsettled question of whether the Big N effect is driven by self-selection. This is important because the absence of a Big N effect would question the auditing literature's focus on incentives and competency as the fundamental drivers of audit quality. Our study also raises awareness of the sensitivity of PSM to its design choices. We do not prescribe a randomization exercise like the one we present here unless the PSM results contradict those from regression analysis, or if sensitivity tests on the design choices suggest that the PSM results are model dependent. We also note that PSM (and CEM) are designed to control for observable selection biases, not unobservable selection biases. As a result, our contribution is limited to the selection bias arising from observable client characteristics.

References

ALTMAN, E. I. *Corporate Distress: A Complete Guide to Predicting, Avoiding and Dealing with Bankruptcy*. New York: Wiley, 1983.

ARCHAMBEAULT, D.S.; T.F. DEZOORT; AND D.R. HERMANSON. "Audit Committee Incentive Compensation and Accounting Restatements." *Contemporary Accounting Research* 25 (2008): 965-992.

ARMSTRONG, C.S.; A.D. JAGOLINZER; AND D.F. Larcker. "Chief Executive Officer Equity Incentives And Accounting Irregularities." *Journal of Accounting Research* 48 (2010), 225-271.

BALL, R.; S. JAYARAMAN; AND L. SHIVAKUMAR. "Audited Financial Reporting and Voluntary Disclosure as Complements: A Test of the Confirmation Hypothesis." *Journal of Accounting and Economics* 53 (2012): 136-166.

BARTUS, T. "Estimation of Marginal Effects Using Margeff." *The Stata Journal* 5 (2005): 309-329.

BECKER, C.L.; M.L. DEFOND; J. JIAMBALVO; AND K.R. SUBRAMANYAM. "The Effect of Audit Quality on Earnings Management." *Contemporary Accounting Research* 15 (1998): 1-24.

BEHN, B.K.; J. CHOI; AND T. KANG. "Audit Quality and Properties of Analyst Earnings Forecasts." *The Accounting Review* 83 (2008): 327-349.

BURNETT, B.; B. CRIPE; G. MARTIN; AND B. MCALLISTER. "Audit Quality and the Trade-off between Accretive Stock Repurchases and Accrual-based Earnings Management." *The Accounting Review* 87 (2012): 1861-1884.

CALIENDO, M., AND S. KOPEINIG. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of economic surveys* 22 (2008): 31-72.

CARAMANIS, C., AND C. LENNOX. "Audit Effort and Earnings Management." *Journal of Accounting and Economics* 45 (2008): 116-138.

CHAN, L.H.; K.C.W. CHEN; T.Y. CHEN; AND Y. YU. "Substitution between Real and Accruals-based Earnings Management after Voluntary Adoption of Compensation Clawback Provisions." *The Accounting Review* (2014): forthcoming.

CHAN, H.K., AND D. WU. "Aggregate Quasi Rents and Auditor Independence: Evidence from Audit Firm Mergers in China." *Contemporary Accounting Research* 28 (2011): 175-213.

CHANEY, P.K.; D.C. JETER; AND L. SHIVAKUMAR. "Self-Selection of Auditors and Audit Pricing in Private Firms." *The Accounting Review* 79 (2004): 51-72.

CHENG, M. M., AND K.A. HUMPHREYS. "The Differential Improvement Effects of the Strategy Map and Scorecard Perspectives on Managers' Strategic Judgments." *The Accounting Review* 87 (2012): 899-924.

DEANGELO, L. "Auditor Independence, 'Low-Balling' and Disclosure Regulation." *Journal of Accounting and Economics* 3 (1981): 113-127.

DEFOND, M.L., AND J. JIAMBALVO. "Incidence and Circumstances of Accounting Errors." *The Accounting Review* 66 (1991): 643-655.

DEFOND, M.; C.Y. LIM; AND Y. ZANG. "Auditor Conservatism and Auditor-Client Contracting." Working paper, University of Southern California and Singapore Management University, 2012.

DEFOND, M.L.; K. RAGHUNANDAN; AND K.R. SUBRAMANYAM. "Do Non-audit Service Fees Impair Auditor Independence? Evidence from Going Concern Audit Opinions." *Journal of Accounting Research* 40 (2002): 1247-1274.

DEFOND, M., AND J. ZHANG. 2014. "A Review of Archival Auditing Research." *Journal of Accounting and Economics* (2014): forthcoming.

DOANE, D.P. "Aesthetic Frequency Classification." *American Statistician* 30 (1976): 181–183

DOOGAR, R., AND R.F. EASLEY. "Concentration without Differentiation: A New Look at the Determinants of Audit Market Concentration." *Journal of Accounting and Economics* 25 (1998): 235-253.

DOPUCH, N. AND D. SIMUNIC. "Competition in Auditing: An Assessment." In Symposium on Auditing Research IV (1982). Urbana: University of Illinois: 401-450.

FRANCIS, J.R.; E.L. MAYDEW; H.C. SPARKS. "The Role of Big 6 Auditors in the Credible Reporting of Accruals." *Auditing: A Journal of Practice & Theory* 18 (1999): 17-34.

FRANCIS, J.R., AND P.N. MICHAS. "The Contagion Effect of Low-Quality Audits." *The Accounting Review* 88 (2013): 521-552.

Francis, J.R.; K. Reichelt, and D. Wang. "The Pricing of National and City-Specific Reputations for Industry Expertise in the U.S. Audit Market." *The Accounting Review* 80 (2005): 113-136.

FRANCIS, J.R., AND M.D. YU. "Big 4 Office Size and Audit Quality." *The Accounting Review* 84 (2009): 1521-1552.

FRANCIS, J.R.; P.N. MICHAS, AND M.D. YU. "Big Four Office Size and Client Restatements." *Contemporary Accounting Research* (2014): forthcoming.

GLOVER, S.M. "The Influence of Time Pressure and Accountability on Auditors' Processing of Nondiagnostic Information." *Journal of Accounting Research* 35 (1997): 213-226.

GUL, F. A.; D. WU, AND Z. YANG. "Do Individual Auditors Affect Audit Quality? Evidence from Archival Data." *The Accounting Review* 88 (2013): 1993-2023.

GUEDHAMI, O., AND J.A. PITTMAN. "Ownership Concentration in Privatized Firms: The Role of Disclosure Standards, Auditor Choice, and Auditing Infrastructure." *Journal of Accounting Research* 44 (2006): 889-929.

HOSMER, D., AND S. LEMESHOW. "Applied Logistic Regression." Second edition, Wiley Series in Probability and Statistics (2000), New York: John Wiley & Sons Inc.

HUELSBECK, D. P.; K.A. MERCHANT, AND T. SANDINO. 2011. "On Testing Business Models." *The Accounting Review* 86 (2011): 1631-1654.

IACUS, S.M.; G. KING, AND G. PORRO. "Mutivariate Matching Methods that are Monotonic Imbalance Bounding." *Journal of the American Statistical Association* 106 (2011): 345-361.

IRELAND, C. J., AND C.S. LENNOX. "The Large Audit Firm Fee Premium: A Case of Selectivity Bias?" *Journal of Accounting, Auditing and Finance* 17 (2002): 73-91.

JOHNSTONE, K. "Client-Acceptance Decisions: Simultaneous Effects of Client Business Risk, Audit Risk, Auditor Business Risk, and Risk Adaptation." *Auditing: A Journal of Practice & Theory* 19 (2000): 1–26.

JOHNSTONE, K.M., AND J.C. BEDARD. "Audit Firm Portfolio Management Decisions." *Journal of Accounting Research* 42 (2004): 659-690.

JONES, K.L.; G.V. KRISHNAN; AND D. MELENDREZ. "Do Models of Discretionary Accruals Detect Actual Cases of Fraudulent and Restated Earnings? An Empirical Analysis." *Contemporary Accounting Research* 25 (2008): 499-531.

KADOUS, K., AND M. MERCER. "Can Reporting Norms Create a Safe Harbor? Jury Verdicts Against Auditors Under Precise and Imprecise Accounting Standards." *The Accounting Review* 87 (2011): 565-587.

KHURANA, I.K., AND K.K. RAMAN. "Litigation Risk and The Financial Reporting Credibility of Big 4 versus Non-Big 4 Audits: Evidence from Anglo-American Countries." *The Accounting Review* 79 (2004): 473-495.

KIM, J.B.; R. CHUNG, AND M. FIRTH. "Auditor Conservatism, Asymmetric Monitoring, and Earnings Management." *Contemporary Accounting Research* 20 (2003): 323-359.

KIM, Y.; H. LI; AND S. LI. "Does Eliminating the Form 20-F Reconciliation from IFRS to US GAAP Have Capital Market Consequences?" *Journal of Accounting and Economics* 53 (2012): 249-270.

LIBBY, T.; S.E. SALTERIO; AND A. WEBB. "The Balanced Scorecard: the Effects of Assurance and Process Accountability on Managerial Judgment." *The Accounting Review* 79 (2004): 1075-1094.

KING, G.; R. NIELSEN; C. COBERLEY; AND J.E. POPE. "Comparative Effectiveness of Matching Methods for Causal Inference." Working paper, Harvard University, 2011.

KOTHARI, S.P.; A.J. LEONE, AND C.E. WASLEY. "Performance Matched Discretionary Accrual Measures." *Journal of Accounting and Economics* 39 (2005): 163-197.

LAWRENCE, A.; M. MINUTTI-MEZA; AND P. ZHANG. "Can Big 4 versus Non-Big 4 Differences in Audit-Quality Proxies be Attributed to Client Characteristics?" *The Accounting Review* 86 (2011): 259-286.

LEAMER, E.E. "Let's Take the Con Out Of Econometrics." *The American Economic Review* 73 (1983): 31-43.

LENNOX, C.; J.R. FRANCIS; AND Z. WANG. 2012. "Selection Models in Accounting Research." *The Accounting Review* 87 (2012): 589-616.

LENNOX, C., AND J. PITTMAN. "Big Five Audits and Accounting Fraud." *Contemporary Accounting Research* 27 (2010): 209-247.

LENNOX, C., AND J. PITTMAN. "Voluntary Audits versus Mandatory Audits." *The Accounting Review* 86 (2011): 1655-1678.

LENNOX, C.; X. WU; AND T. ZHANG. "How Do Audit Adjustments Affect Measures of Earnings Quality?" Working paper, Nanyang Technological University, 2014.

MANSI, S.A.; W.F. MAXWELL, AND D.P. MILLER. "Does Auditor Quality and Tenure Matter to Investors? Evidence from the Bond Market." *Journal of Accounting Research* 42 (2004): 755-793.

MICHAS, P. "The Importance of Audit Profession Development in Emerging Market Countries." *The Accounting Review* 86 (2011): 1731-1764.

NAIKER, V.; D.S. SHARMA; AND V.D. SHARMA. "Do Former Audit Firm Partners on Audit Committees Procure Greater Nonaudit Services from the Auditor?" *The Accounting Review* 88 (2012): 297-326.

PITTMAN, J.A., AND S. FORTIN. "Auditor Choice and the Cost of Debt Capital for Newly Public Firms." *Journal of Accounting and Economics* 37 (2004): 113-136.

RAGHUNANDAN, K., AND D. RAMA. "Auditor Resignations and the Market for Audit Services." *Auditing: A Journal of Practice & Theory* 18 (1999): 124–134.

ROSENBAUM, P. R., AND D.B. RUBIN. "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score." *The American Statistician* 39 (1985): 33–38.

SCHWARTZ, K.B., AND B.S. SOO. "The Association between Auditor Changes and Reporting Lags." *Contemporary Accounting Research* 13 (1996): 353-370.

STURGES, H. A. "The Choice of a Class Interval." *Journal of the American Statistical Association* 21 (1926): 65–66.

STUART, E.A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25 (2010): 1-21.

TEOH, S., AND T.J. WONG. "Perceived Auditor Quality and the Earnings Response Coefficient." *The Accounting Review* 68 (1993): 346–66.

TUCKER, J.W. "Selection Bias and Econometric Remedies in Accounting and Finance Research." *Journal of Accounting Literature* 29 (2010): 31-57.

WANG, Q.; T.J. WONG, AND L. XIA. "State Ownership, the Institutional Environment, and Auditor Choice: Evidence From China." *Journal of Accounting and Economics* 46 (2008): 112-134.

WATTS, R., AND J. ZIMMERMAN. "Auditors and the Determination of Accounting Standards." Working paper, University of Rochester, 1981.

WEBER, J., AND M. WILLENBORG. "Do Expert Informational Intermediaries Add Value? Evidence from Auditors in Microcap IPOs." *Journal of Accounting Research* 41 (2003): 681-720.

ZHOU, X.H.; N.A. OBUCHOWSKI; AND D.M. OBUCHOWSKI. "Statistical Methods in Diagnostic Medicine." John Wiley & Sons Inc., New York, 2002.

ZMIJEWSKI, M.E. "Methodological Issues Related to the Estimation Of Financial Distress Prediction Models." *Journal of Accounting Research* 22 (1984): 59-82.

Appendix A. Variable Definitions

Variable	*Definition*	*Source*
Audit quality proxies		
Absolute discretionary accruals	Performance-matched absolute discretionary accruals as specified by Kothari et al. [2005].	Compustat
Signed discretionary accruals	Performance-matched discretionary accruals as specified by Kothari et al. [2005].	Compustat
Restatement	Equals one when a firm restates net income downward by more than 10%, and zero otherwise.	AuditAnalytics, Compustat
Going concern opinion	Equals one when the firm's auditor issued a going concern opinion, and zero otherwise.	AuditAnalytics
Audit fees	Natural logarithm of audit fees.	AuditAnalytics
Matching variables		
Log(assets)	Natural logarithm of total assets.	Compustat
Asset turnover ratio	Sales scaled by one-year lagged total assets.	Compustat
Current ratio	Current assets scaled by current liabilities.	Compustat
Leverage	Long-term debt (including long-term debt in current liabilities), scaled by average total assets.	Compustat
ROA	Net income scaled by average total assets.	Compustat
Additional control variables used in regression models		
Accruals_{t-1}	Total accruals scaled by total assets in year *t*-1.	Compustat
Altman z-score	Probability of bankruptcy from Altman [1983].	Compustat
Auditor switch	Equals one when the firm switches auditors in the focal year, and zero otherwise.	AuditAnalytics
December fiscal year-end	Equals one when a firm has a December fiscal year-end, and zero otherwise.	Compustat
Future finance	Equals one when the firm issues equity or debt in the subsequent year, and zero otherwise.	Compustat
Idiosyncratic volatility	The standard deviation of the firm's stock return over the fiscal year.	CRSP
Investments	Short- and long-term investment securities (including cash and cash equivalents) scaled by total assets at year-end.	Compustat
Litigation	Equals one if the firm operates in the following SIC codes: 2833–2836, 3570–3577, 3600–3674, 5200–5961, and 7370), and zero otherwise.	Compustat
Log(firm age)	Natural logarithm of firm age, where firm age equals the number of years a firm's financial data is available from Compustat.	Compustat
Loss	Equals one when a firm's net income is negative, and zero otherwise.	Compustat
Loss_{t-1}	Equals Loss lagged by one year.	Compustat
Market-to-book ratio	Market value of equity scaled by book value of equity.	Compustat
Operating cash flow	Operating cash flow scaled by total assets.	Compustat
Operating CF volatility	The three-year standard deviation of Operating cash flow.	Compustat
PP&E growth	One-year percentage growth in net property, plant and equipment.	Compustat
Quick ratio	Current assets excluding inventories scaled by current liabilities.	Compustat
Reporting lag	Number of days between fiscal year-end and earnings announcement date.	Compustat
Sales growth	One-year percentage growth in sales.	Compustat
Sales volatility	The firm's three-year standard deviation of sales.	Compustat
Sec Office	Indicator variable that captures the closest regional SEC office.	Compustat
Shares issued	Equals one when a firm issued equity in the focal year, and zero otherwise.	Compustat
Stock return	Buy-and-hold stock returns for fiscal year *t*-1.	CRSP
Stock return beta	The firm's beta estimated using a market model over the fiscal year.	CRSP
Stock return volatility	Standard deviation of fiscal year stock returns.	CRSP
Total leverage	Total liabilities scaled by total assets.	Compustat
Total leverage increase	Equals Total leverage_t - $\text{Total leverage}_{t-1}$.	Compustat

Appendix A (*continued*)

Variable	Definition	Source
Zmijevski bankruptcy score	The probability of bankruptcy from Zmijevski [1984].	Compustat
# Foreign segments	Natural logarithm of one plus the Number of foreign operating segments.	Compustat
# Geographic segments	Natural logarithm of one plus the Number of geographic operating segments.	Compustat
# Operating segments	Natural logarithm of one plus the Number of operating segments.	Compustat

Appendix B. Descriptive Statistics

	Big N			*Other audit firms*			*Difference*	
Variables	*Mean*	*Median*	*Std. Dev.*	*Mean*	*Median*	*Std. Dev.*	*Mean*	*Median*
Panel A: Big N (N=10,947) versus Other audit firms (N=3,544): discretionary accruals and restatements								
Audit quality proxies								
Absolute discretionary accruals	0.047	0.031	0.052	0.075	0.049	0.078	-0.028***	-0.017***
Signed discretionary accruals	0.000	0.001	0.068	0.003	0.002	0.101	-0.003**	-0.001
Restatement	0.025	0.000	0.158	0.025	0.000	0.155	0.001	0.000
Matching variables								
Log(assets)	6.564	6.548	1.926	4.057	3.976	1.344	2.507***	2.571***
Asset turnover ratio	0.963	0.859	0.654	1.102	1.009	0.771	-0.140***	-0.150***
Current ratio	2.999	2.165	2.605	3.483	2.521	3.126	-0.484***	-0.356***
Leverage	0.231	0.199	0.220	0.173	0.096	0.215	0.058***	0.103***
ROA	-0.028	0.033	0.217	-0.107	0.001	0.303	0.078***	0.032***
Control variables								
Accruals$_{t-1}$	-0.075	-0.053	0.125	-0.088	-0.051	0.178	0.012***	-0.001*
Altman z-score	1.370	1.696	2.797	0.990	1.888	4.139	0.381***	-0.192**
Litigation	0.272	0.000	0.445	0.303	0.000	0.460	-0.030***	0.000***
Log(assets)	6.564	6.548	1.926	4.057	3.976	1.344	2.507***	2.571***
Loss	0.329	0.000	0.470	0.497	0.000	0.500	-0.168***	0.000***
Market-to-book ratio	2.881	2.112	4.187	2.685	1.764	4.566	0.196**	0.348***
Operating cash flow	0.051	0.083	0.193	-0.019	0.041	0.254	0.070***	0.043***
Operating cash flow volatility	98.267	22.650	222.061	8.167	3.148	16.651	90.100***	19.502***
PP&E growth	9.530	2.444	39.807	12.792	-0.683	57.800	-3.263***	3.127***
Sales growth	15.891	8.477	49.411	16.642	6.383	62.182	-0.752	2.094***
Sales volatility	318.743	65.279	736.033	23.142	6.848	53.644	295.600***	58.431***
Shares issued	0.884	1.000	0.320	0.804	1.000	0.397	0.080***	0.000***
Stock return	0.116	0.031	0.628	0.089	-0.016	0.715	0.027**	0.047***
Stock return volatility	0.132	0.111	0.085	0.168	0.147	0.101	-0.036***	-0.036***
Total leverage	0.507	0.497	0.271	0.425	0.367	0.287	0.081***	0.130***
# Geographic segments	2.738	2.000	2.401	1.897	1.000	2.154	0.841***	1.000***
# Operating segments	2.192	1.000	1.780	1.707	1.000	1.268	0.485***	0.000***
Panel B: Big N (N=3,371) versus Other audit firms (N=1,717): going concern opinions								
Audit quality proxy								
Going concern	0.036	0.000	0.188	0.037	0.000	0.189	-0.001	0.000
Matching variables								
Log(assets)	5.427	5.224	1.721	3.755	3.638	1.304	1.672***	1.586***
Asset turnover ratio	0.692	0.558	0.610	0.905	0.760	0.760	-0.213***	-0.202***
Current ratio	4.120	2.792	3.971	4.036	2.485	4.508	0.084	0.307***
Leverage	0.230	0.136	0.271	0.170	0.077	0.233	0.061***	0.059***
ROA	-0.231	-0.133	0.278	-0.256	-0.130	0.346	0.025***	-0.003
Control variables								
Future finance	0.929	1.000	0.257	0.849	1.000	0.358	0.080***	0.000***
Idiosyncratic volatility	0.161	0.139	0.094	0.168	0.147	0.103	-0.007***	-0.008***
Investments	0.394	0.321	0.317	0.315	0.218	0.293	0.079***	0.102***
Log(assets)	5.427	5.224	1.721	3.755	3.638	1.304	1.672***	1.586***
Log(firm age)	2.522	2.485	0.678	2.613	2.639	0.682	-0.091***	-0.154***
Loss$_{t-1}$	0.713	1.000	0.453	0.708	1.000	0.455	0.005	0.000
Operating cash flow	-0.121	-0.023	0.282	-0.152	-0.048	0.313	0.031***	0.025***
Reporting lag	56.176	55.000	22.221	73.659	75.000	20.593	-17.482***	-20.000***
Stock return	0.178	-0.079	0.980	0.031	-0.175	0.874	0.147***	0.096***
Stock return beta	1.854	1.612	1.882	1.394	1.163	1.902	0.460***	0.449***
Total leverage	0.503	0.444	0.343	0.429	0.368	0.315	0.074***	0.076***

Appendix B (*continued*)

Total leverage increase	0.045	0.025	0.176	0.027	0.019	0.195	0.018***	0.005***
Zmijevski bankruptcy score	0.329	0.144	0.366	0.280	0.083	0.353	0.049***	0.061***
Panel C: Big N (N=13,165) versus Other audit firms (N=6,819): audit fees								
Audit quality proxy								
Audit fees	13.867	13.844	1.167	11.877	11.797	1.088	1.990***	2.047***
Matching variables								
Log(assets)	6.531	6.560	1.985	2.941	3.058	2.041	3.590***	3.502***
Asset turnover ratio	0.948	0.823	0.720	1.114	0.885	1.094	-0.165***	-0.063**
Current ratio	2.995	2.048	2.979	3.093	1.887	3.880	-0.098**	0.161***
Leverage	0.265	0.223	0.299	0.485	0.149	0.986	-0.220***	0.074***
ROA	-0.049	0.030	0.328	-0.725	-0.090	1.674	0.676***	0.120***
Control variables								
Auditor switch	0.028	0.000	0.165	0.203	0.000	0.402	-0.175***	0.000***
Current ratio	2.995	2.048	2.979	3.093	1.887	3.880	-0.098**	0.161***
December fiscal year-end	0.249	0.000	0.433	0.352	0.000	0.477	-0.102***	0.000***
Going concern opinion	0.038	0.000	0.192	0.282	0.000	0.450	-0.244***	0.000***
Leverage	0.265	0.223	0.299	0.485	0.149	0.986	-0.220***	0.074***
Log(assets)	6.531	6.560	1.985	2.941	3.058	2.041	3.590***	3.502***
Loss	0.346	0.000	0.476	0.628	1.000	0.483	-0.282***	-1.000***
Quick ratio	2.486	1.465	2.890	2.502	1.298	3.620	-0.016	0.167***
ROA	-0.049	0.030	0.328	-0.725	-0.090	1.674	0.676***	0.120***
# Foreign segments	0.702	0.693	0.724	0.321	0.000	0.573	0.381***	0.693***
# Operating segments	0.987	0.693	0.549	0.848	0.693	0.376	0.139***	0.000***

This table presents descriptive statistics on all variables used in subsequent analyses with the exception of the replication of LMZ presented in Table 1 and Figure 1. Big N equals one when a firm's annual financial statements are audited by a Big 4 auditor (PricewatershouseCoopers, Delottie & Touche, Ernst & Young, or KPMG), and zero otherwise. All measures are as defined in Appendix A. Panel A provides descriptive statistics on the sample used for absolute discretionary accruals, signed discretionary accruals (discretionary accrual regression models include 2-digit SIC code, year, and closest SEC office indicators), and restatement analyses (restatement regression model includes year and closest SEC office indicators, and absolute discretionary accruals) separately for Big N and non-Big N firms, and panels B and C report descriptive statistics for the going concern opinions and audit fees (audit fees regression model includes 2-digit SIC code and year indicators) analyses respectively. Two-sample *t*-tests are used to test the differences in means, and Wilcoxon two-sample tests are used to test differences in medians. *,**, and *** indicate two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Big N Treatment Effects from regressing DAC on Big N from 3,000 PSM matched samples randomized over three research design choices

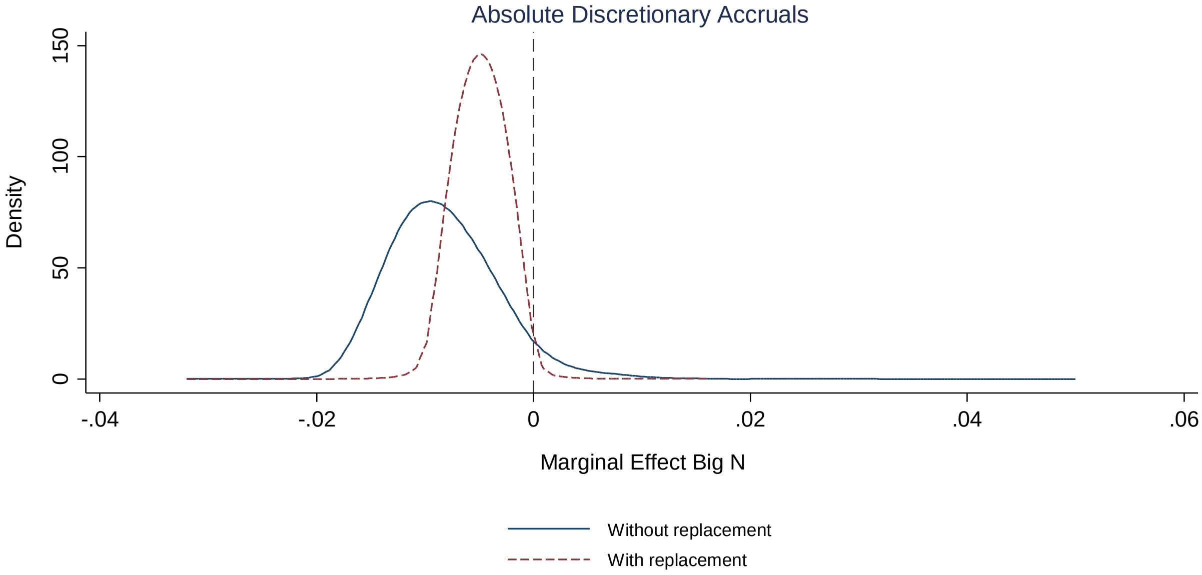


Fig. 1. This figure presents a density plot illustrating the variation in the Big N treatment effect across random combinations of design choices, for the LMZ sample period 1988-2006. The solid line represents the treatment effect from matching without replacement, and the dotted line represents the treatment effect from matching with replacement. To construct this figure we take 3,000 random draws from the full-set of non-linear terms (squares, cubes, and simple interactions) that can be constructed using the five matching variables presented in Panel B of Table 1, add the randomly selected non-linear terms to the five matching variables, and estimate the propensity score model using the selected set of matching variables. Matched samples are created by matching treatment firms to either one, two or three control firms (1,000 matched samples per matching approach) either without or with replacement. Further, for each matched sample we prune a random number of the treatment firms that have the worst overall match quality (based on the propensity score difference between the treatment and the matched control firms), where matched samples are required to be large enough to detect a moderate effect size (Cohen's *d* = 0.5) with 0.80 statistical power. Subsequently, we use the resulting matched samples to estimate the absolute discretionary accrual regression (weighted least-squares) presented in Panel D of Table 1, and use the coefficient estimates on the Big N indicator to create the density plot presented in this figure. All regressions are estimated using probability weights derived from the matching procedure. Specifically, the weight of each control increases by one over the number of allowed matches for each time that it is included in the sample. For example, for one-to-two matching the weight of each control firm increases by $^1/_2$ for each time it is included in the sample.

Big N Treatment Effects from regressing five measures of audit quality on Big N from 3,000 PSM matched samples randomized over three research design choices

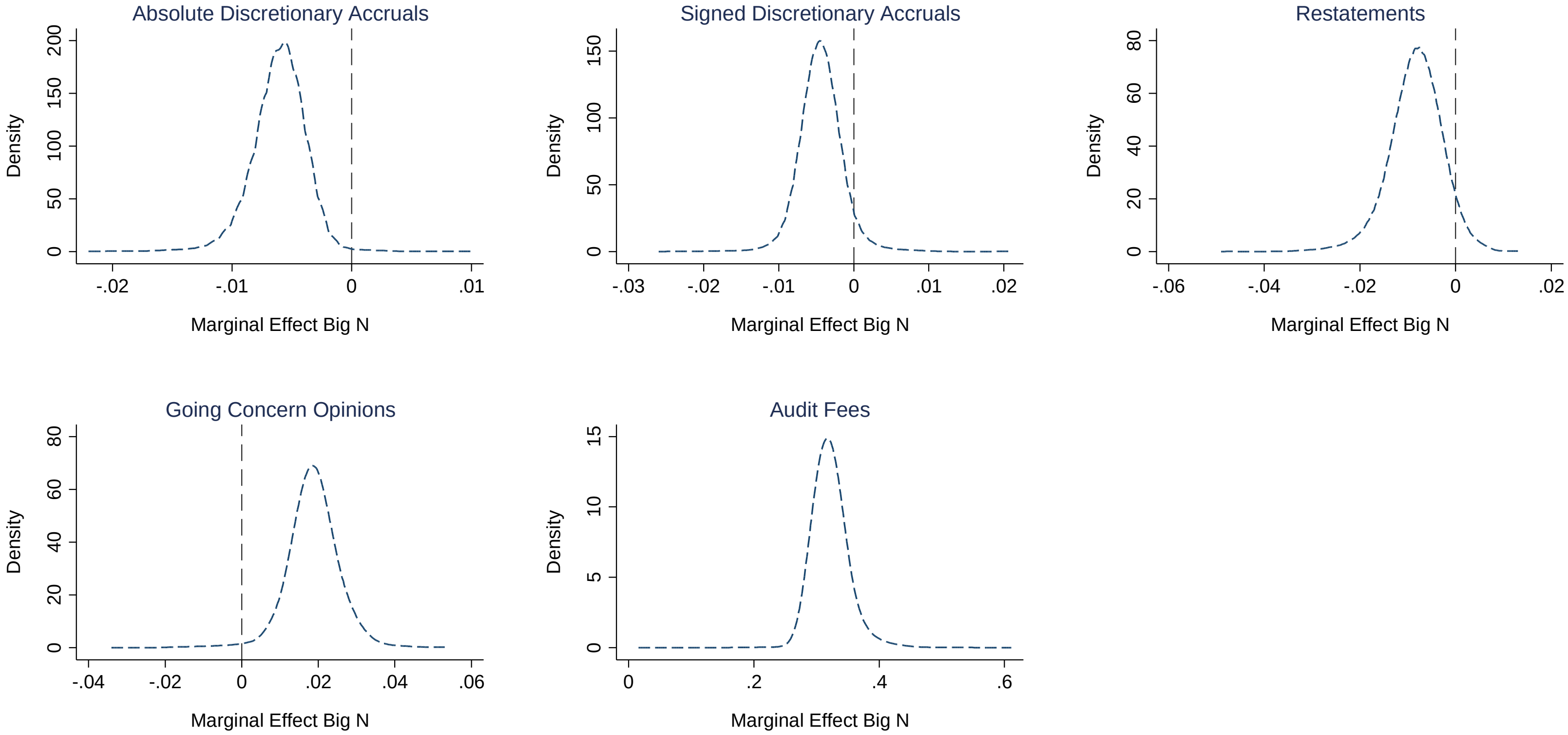


Fig. 2. This figure presents density plots illustrating the variation in the Big N treatment effect on five measures of audit quality, for our sample period 2003-2009. The approach used for creating random matched samples is as described in Figure 1. We employ matching with replacement. Using the matched samples we estimate the treatment effect of Big N on the five audit quality measures by regressing each audit quality measure on the Big N indicator and the control variables described in Appendix B. The density plots are based on Big N coefficient estimates derived from OLS regressions for absolute discretionary accruals, signed discretionary accruals and audit fees, and Big N average marginal effects derived from logistic regression models (Bartus [2005]) for going concern opinions and restatements.

The Bias-Variance Frontier

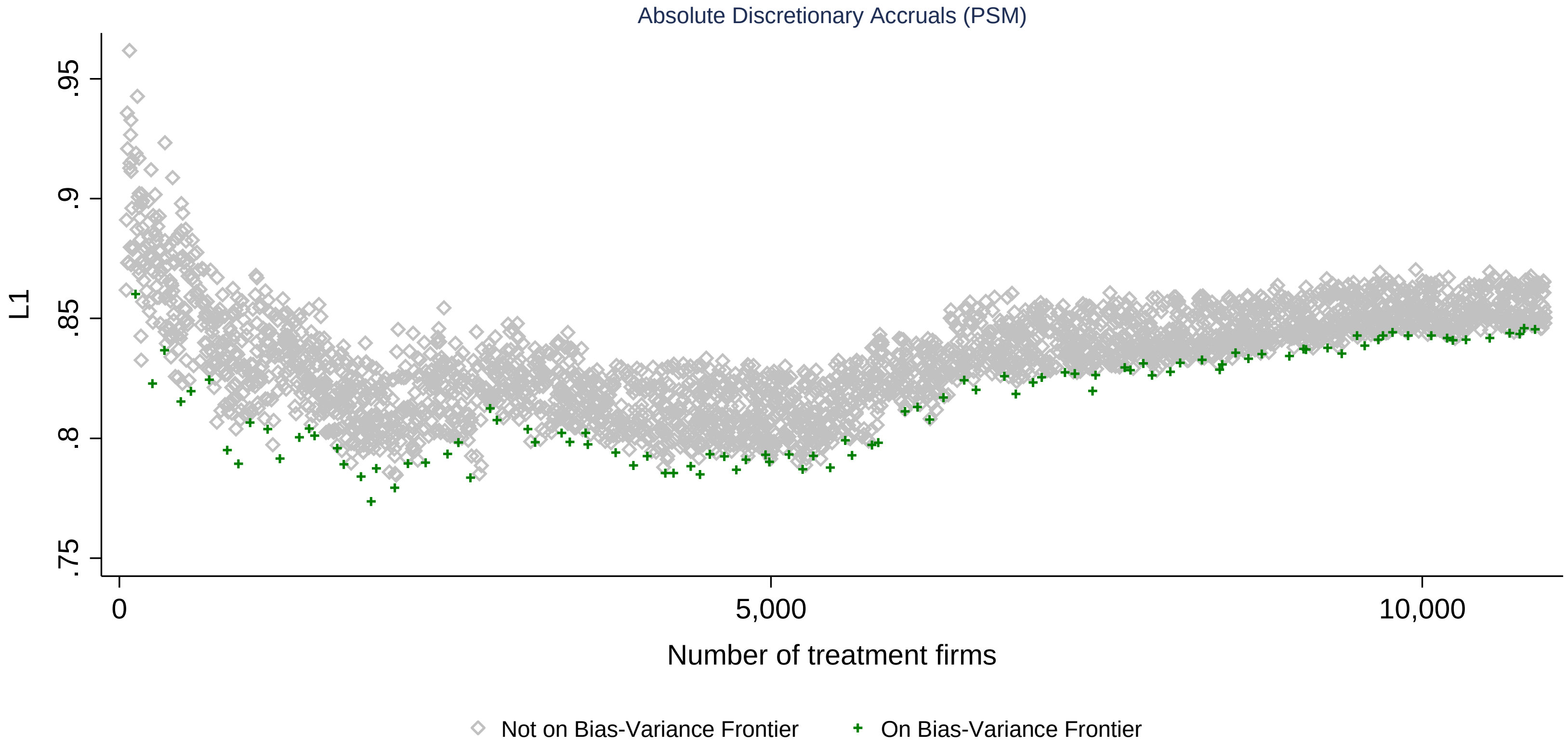
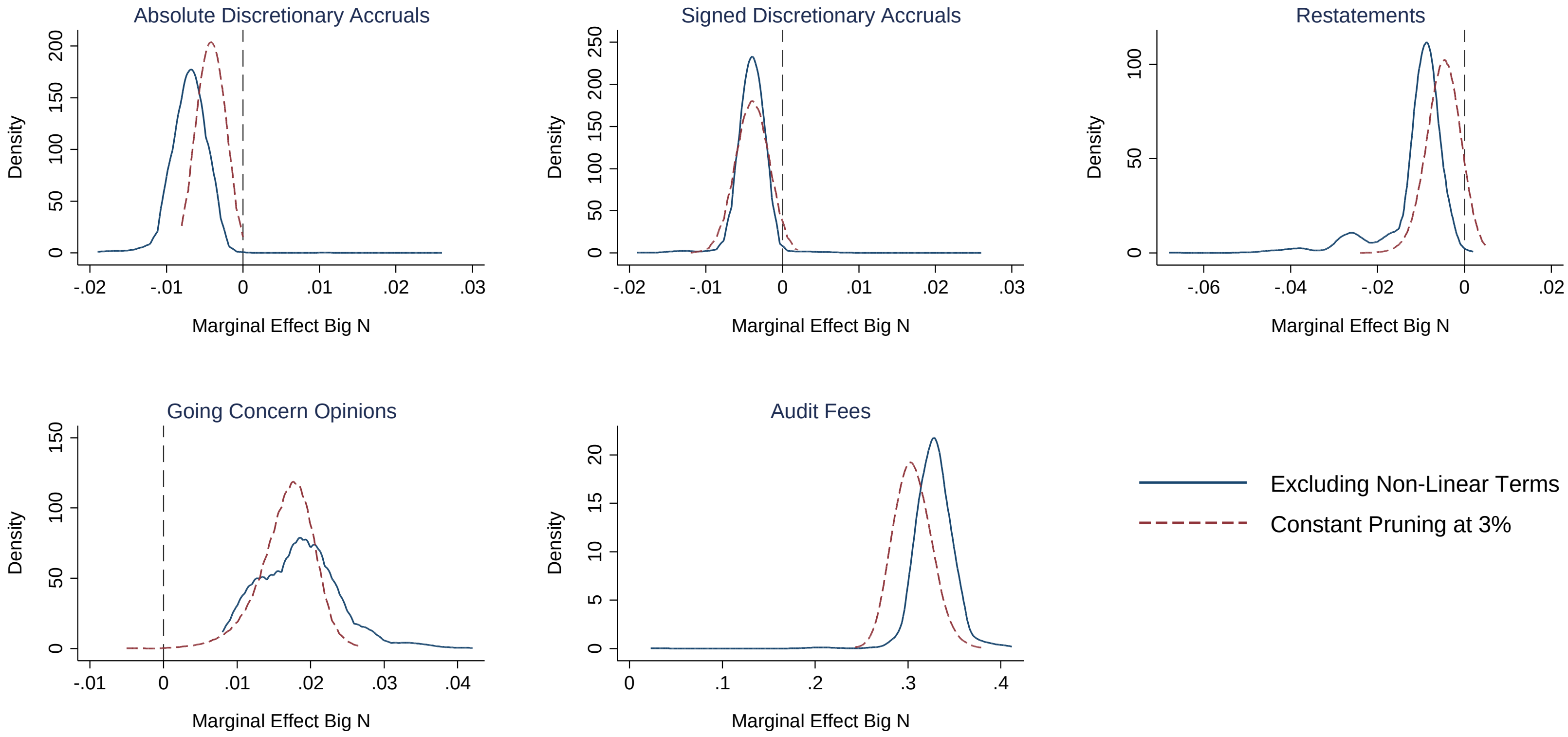


Fig. 3. This figure uses the matched samples generated for the absolute discretionary accrual analysis using PSM (Figure 2) to illustrate how we select matched samples that are on the efficient bias-variance frontier. We create 100 percentile rank groups based on the number of treatment firms and classify the matched sample with the lowest L_1 imbalance in each group as being on the bias-variance frontier.

Part I: Treatment Effects from regressing five measures of audit quality on Big N from 3,000 PSM matched samples randomized over three research design choices after holding pruning constant or excluding non-linear terms



Part II: Treatment Effects from regressing five measures of audit quality on Big N from 3,000 PSM matched samples randomized over three research design choices, separately plotted one-to-one, one-to-two, and one-to-three matches

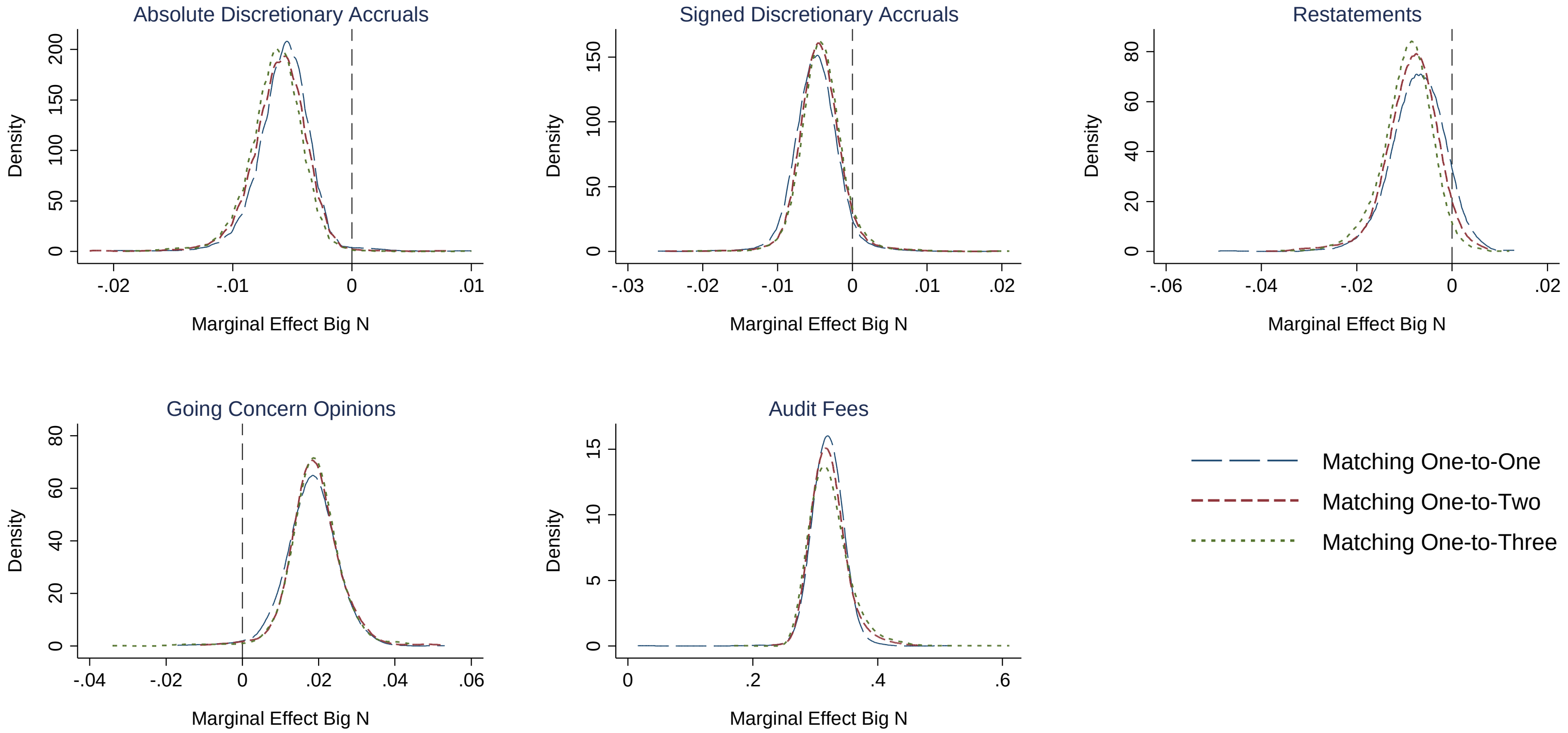


Fig. 4. This figure presents density plots that examine the individual effects of varying the pruning level, the nonlinear terms, and the treatment to control ratio on the Big N treatment effect. The figure illustrates that the variation in the Big N treatment effect is driven predominantly by varying the pruning level. The analyses presented in the top figure repeat the analyses presented in Figure 2, but either fixes the percentage of treatment firms that is pruned at 3% or excludes non-linear terms from the propensity score model. The analyses presented in the bottom figure present the results presented in Figure 2 by the number of control firms that are matched to each treatment firm.

Distribution of Big N Treatment Effects after Randomly Adding Five Additional Matching Variables

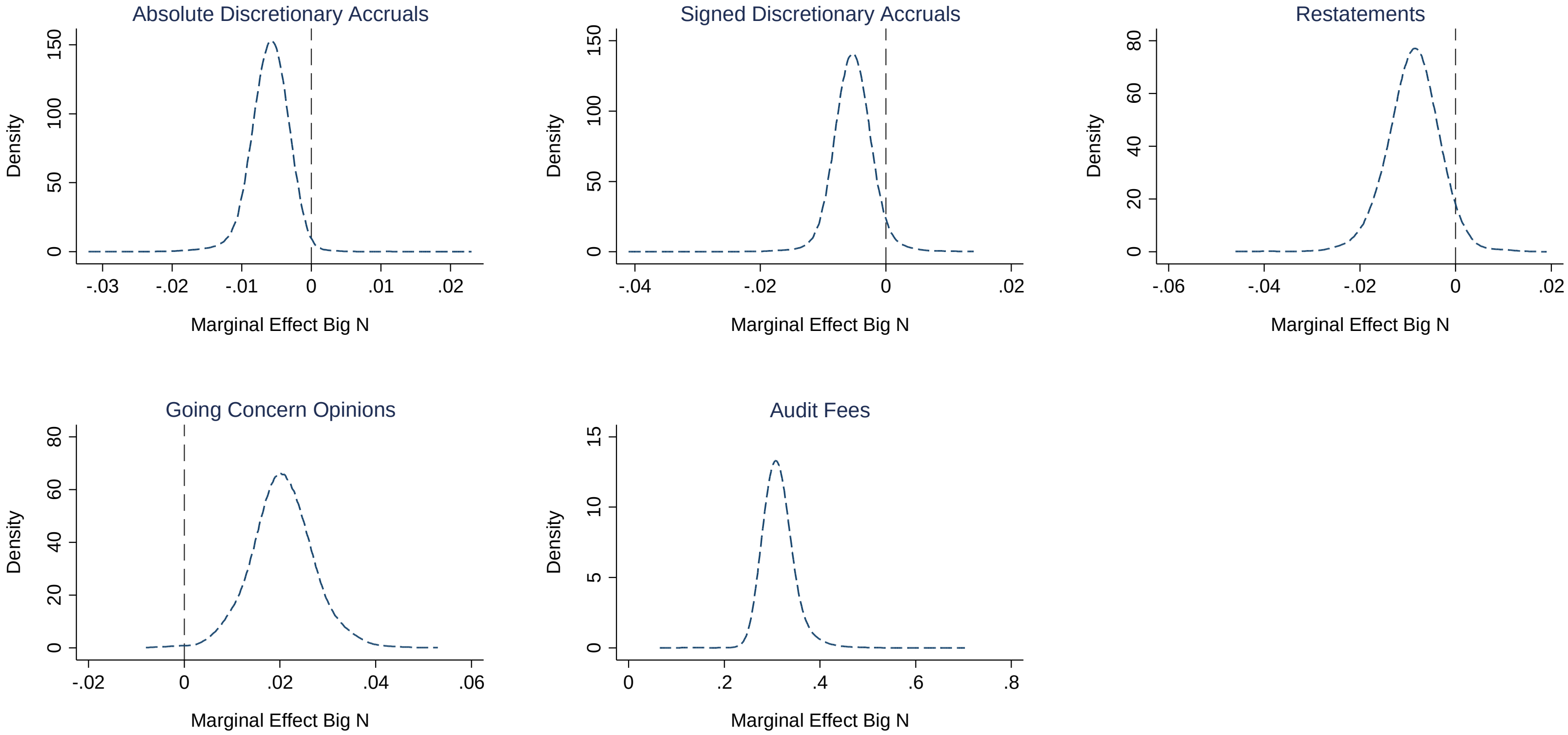


Fig. 5. This figure repeats the analysis presented in Figure 2 after randomly adding five additional variables from the treatment effect regressions (see Appendix B for the set of variables we choose from) to our five main matching variables. The density plots are based on Big N coefficient estimates derived from OLS regressions for absolute discretionary accruals, signed discretionary accruals and audit fees, and Big N average marginal effects derived from logistic regression models (Bartus [2005]) for going concern opinions and restatements.

Big N Effects from using CEM as an alternative matching procedure

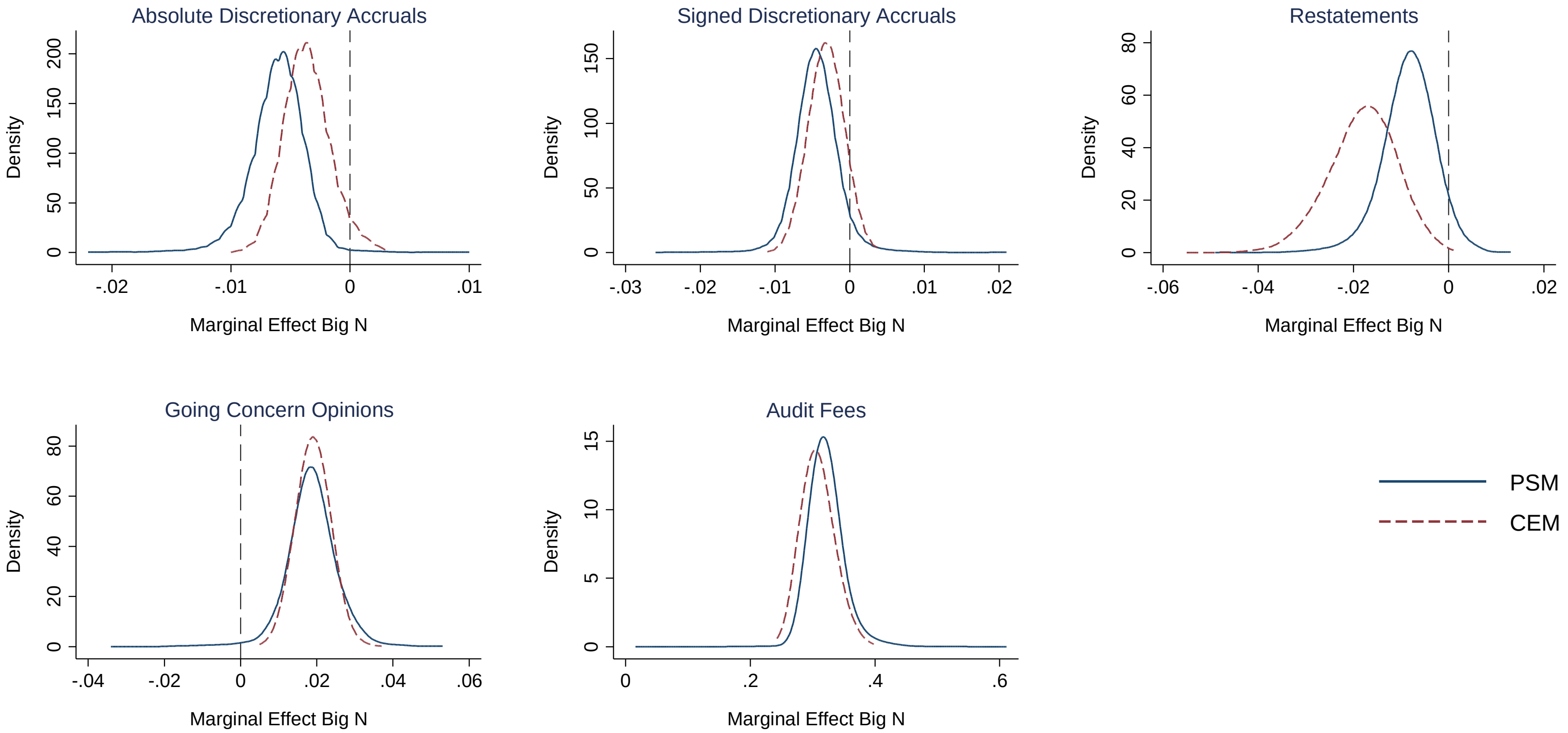


Fig. 6. This figure compares the estimated treatment effects between propensity score matching (PSM) and coarsened exact matching (CEM). For the CEM results we change the coarsenings to generate 3,000 matched samples. To search over a large number of solutions, we sample from the set of possible coarsenings equally spaced between minimum and maximum levels. To ensure that we retain a reasonable number of treatment firms we set the minimum equal to two and the maximum equal to the number of bins suggested for each covariate by Doane's formula [1976]. Consistent with our previous analyses, we require the CEM matched samples to be large enough to detect a moderate effect size (Cohen's *d* = 0.5) with 0.80 statistical power. The figure contrasts the estimated Big N treatment effects from PSM samples presented in Figure 2 with the estimated Big N treatment effects from CEM samples.

Table 1

Replicating the PSM Analysis of DAC Based on Lawrence et al. [2011]

Panel A: Big N (N=78,725) versus Other Audit Firms (N=19,544): Absolute Discretionary Accruals

	Big N			*Other audit firms*			*Difference*	
Variables	*Mean*	*Median*	*Std. Dev.*	*Mean*	*Median*	*Std. Dev.*	*Mean*	*Median*
Absolute discretionary accruals	0.080	0.049	0.095	0.134	0.087	0.135	-0.054***	-0.038***
Log(assets)	5.286	5.163	2.167	2.735	2.541	1.862	2.551***	2.622***
Log(market value)	5.161	5.104	2.284	2.711	2.583	1.970	2.450***	2.520***
Asset turnover ratio	1.145	1.001	0.822	1.201	1.031	0.982	-0.056***	-0.030
Current ratio	2.911	1.954	3.259	3.054	1.724	4.153	-0.143***	0.230***
Leverage	0.253	0.213	0.241	0.279	0.193	0.330	-0.026***	0.020***
ROA	-0.042	0.031	0.264	-0.208	-0.025	0.494	0.166***	0.057***

Panel B: Propensity Score Model

$$Big\ N = \alpha_0 + \beta_1\ Log(assets) + \beta_2\ Asset\ turnover\ ratio + \beta_3\ Current\ ratio + \beta_4\ Leverage + \beta_5\ ROA + \gamma Z + \varepsilon$$

α_0	β_1	β_2	β_3	β_4	β_5	*N*	*AUC*	*Test AUC=0.50*	
-0.773**	0.857***	0.153***	0.024***	-0.790***	-0.605***	98,269	0.858	Z-statistic	231.488
(0.376)	(0.019)	(0.028)	(0.006)	(0.069)	(0.051)			p-value	0.000

Panel C: Imbalance Metrics Big N (*N*=14,277) versus Matched Other Audit Firms (*N*=14,277)

					Imbalance	
		Mean			*Univariate*	
Variables	*Big N*	*Other audit firms*	*Difference*		*|%bias|*	L_1
Log(assets)	3.059	3.272	-0.213***		10.557	0.043
Log(market value)	3.086	3.127	-0.041*		1.928	0.029
Asset turnover ratio	1.199	1.195	0.004		0.411	0.017
Current ratio	3.192	3.195	-0.004		0.098	0.060
Leverage	0.247	0.246	0.001		0.231	0.056
ROA	-0.159	-0.142	-0.017***		4.170	0.070
				Multivariate	2.899	0.597

Panel D: Influence of Big N on Absolute Discretionary Accruals

$$Absolute\ discretionary\ accruals = \alpha_0 + \beta_1\ Big\ N + \beta_2\ Log(Market\ value) + \beta_3\ ROA + \beta_4\ Leverage + \beta_5\ Current\ ratio + \gamma Z + \varepsilon$$

α_0	β_1	β_2	β_3	β_4	β_5	*N*	R^2
0.102***	-0.002	-0.009***	-0.082***	0.002	-0.002***	28,554	0.145
(0.012)	(0.002)	(0.000)	(0.003)	(0.004)	(0.000)		

This table presents results from replicating Lawrence et al. [2011] with respect to absolute discretionary accruals for their sample period 1988-2006 using their variable definitions, where *Absolute discretionary accruals* is absolute discretionary accruals; *Log(assets)* is a logarithmic transformation of total assets at the end of the fiscal year, *Log(market value)* is a logarithmic transformation of market value of equity at the end of the fiscal year; *Asset turnover ratio* equals sales scaled by one-year lagged total assets; *Current ratio* equals current assets scaled by current liabilities; *Leverage* equals the sum of long term debt and debt in current liabilities scaled by one-year lagged average total assets; *ROA* equals net income scaled by one-year lagged average total assets; and *Z* is a vector of two-digit SIC code and year indicators. Panel A presents descriptive statistics on the measures used for examining the effect of Big N on absolute discretionary accruals. Two-sample *t*-tests are used to test the differences in means, and Wilcoxon two-sample tests are used to test differences in medians. Panel B presents the results of a logistic regression from which we derive propensity scores. Standard errors are computed using firm-level cluster robust standard errors and are presented in parentheses. We use the area under the ROC (Receiver Operating Characteristic) curve (AUC) as a measure of how accurate our logistic regression model is at discriminating between treatment and control firms (Hosmer and Lemeshow [2000], p.160-164). A greater AUC indicates better predictive ability of the model—Hosmer and Lemeshow classify anything above 0.7 as an "acceptable" level of discrimination. We test whether the AUC is statistically different from chance (0.50) using the *Z*-statistic which is equal to (AUC-0.50)/((standard error (AUC)) (see Zhou et al. [2002]). Panel C presents univariate and multivariate imbalance metrics on the matched sample. Following, Lawrence et al. [2011] we match each Big N firm to one non-Big N firm, without replacement, without non-linear terms of the covariates, and require matches to have a maximum caliper difference of 0.03. Two-sample *t*-tests are used to test the differences in means. *|%bias|* is the absolute standardized percentage bias, calculated as the difference of the sample means between the matched treatment (i.e. Big N) and control (i.e., non-Big N) samples as a percentage of the square root of the average of the sample variances in the Big N and non-Big N samples (Rosenbaum and Rubin [1985]). The multivariate version of this measure is the average of the absolute standardized percentage bias across all covariates. L_1 is the difference between the multidimensional histogram of covariates in the Big N and the non-Big N subsamples (Iacus et al. [2011]). For the univariate version of this measure we include exclusively the respective covariate, and for the multivariate version we include all covariates. Specifically, we use Sturges rule [1926] to coarsen covariates. Subsequently, we cross-tabulate the discretized variables as $X_1 \times \cdots \times X_k$ for the Big N and non-Big N samples, and record the *k*-dimensional weighted relative frequencies for the treated $f_{l_1 \cdot \cdot}$ $f_{l_k \cdot}$ and control $g_{l_1 \cdot \cdot}$ g_{l_k} units. Finally, our imbalance measure is the absolute difference over all the cell values: $L_1(f,g) = \frac{1}{2}\sum_{l_1 \cdots l_k}^{n} \left| f_{l_1 \cdots l_{1k}} - g_{l_1 \cdots l_{1k}} \right|$ and where the summation is over all cells of the multivariate histogram. The L_1 measure varies in [0,1], where perfect balance results in $L_1 = 0$, and complete separation results in $L_1=1$. Any value in interval [0,1] indicates the amount of difference between *k*-dimensional frequencies of the two groups (see Iacus et al. [2011] for a more detailed explanation). Panel D presents the results of an OLS regression that regresses absolute discretionary accruals on Big N, four control variables, and industry and year fixed effects (denoted by *Z* in the regression specification) using the PSM matched sample. Standard errors are computed using firm-level cluster robust standard errors and are presented in parentheses. *,**, and *** indicate two-tailed statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 2

Big N Treatment Effects from Regressing Absolute DAC On Big N from 3,000 PSM Matched Samples Randomized Over Three Research Design Choices.

Approach	*Mean*	*Median*	*PCT Sign*	*PCT Significant*	*Std. Dev.*	*Minimum*	*Maximum*
Without replacement	-0.008	-0.009	0.945	0.772	0.005	-0.032	0.050
With replacement	-0.005	-0.005	0.993	0.950	0.001	-0.032	0.016
Difference	-0.003	-0.004	-0.048	-0.178	0.004	0.000	0.034

This table presents summary statistics for the regression results presented in Figure 1, where *PCT Sign* is the fraction of treatment effects that are in the predicted direction (i.e., negative for absolute DAC), and *PCT Significant* is the fraction of treatment effects that are in the predicted direction and significant at p < 5% (one-tailed).

Table 3

Big N Treatment Effects from Regressing Five Measures of Audit Quality on Big N from 3,000 PSM Matched Samples Randomized Over Three Research Design Choices.

Audit Quality Measure	*Mean*	*Median*	*PCT Sign*	*PCT Significant*	*Std. Dev.*	*Minimum*	*Maximum*
Absolute discretionary accruals	-0.006	-0.006	0.991	0.910	0.002	-0.022	0.010
Signed discretionary accruals	-0.004	-0.004	0.943	0.483	0.003	-0.026	0.021
Restatements	-0.008	-0.008	0.944	0.276	0.006	-0.049	0.013
Going concern opinions	0.019	0.019	0.988	0.713	0.007	-0.034	0.053
Audit fees	0.322	0.319	1.000	1.000	0.027	0.016	0.611

This table presents summary statistics for the estimated Big N treatment effects plotted in Figure 2, where *PCT Sign* is the fraction of treatment effects that are in the predicted direction, and *PCT Significant* is the fraction of treatment effects that are in the predicted direction and significant at $p < 5\%$ (one-tailed). The predicted sign is negative for absolute DAC, signed DAC, and restatements, positive for going concern opinions and audit fees.

Table 4

Selected Matched Samples from the Bias-Variance Frontier

Panel A: Propensity Score Matching

Audit Quality Measure	*Mean*	*Median*	*PCT Sign*	*PCT Significant*	*Std. Dev.*	*Minimum*	*Maximum*
Absolute discretionary accruals	-0.006	-0.006	0.990	0.890	0.003	-0.019	0.001
Signed discretionary accruals	-0.004	-0.004	0.960	0.410	0.002	-0.014	0.001
Restatements	-0.009	-0.009	0.910	0.290	0.006	-0.025	0.004
Going concern opinions	0.018	0.018	0.990	0.740	0.006	-0.012	0.033
Audit Fees	0.315	0.311	1.000	1.000	0.024	0.193	0.391

Panel B: Propensity Score Matching (Size Balanced)

Audit Quality Measure	*Mean*	*Median*	*PCT Sign*	*PCT Significant*	*Std. Dev.*	*Minimum*	*Maximum*
Absolute discretionary accruals	-0.007	-0.006	0.990	0.900	0.003	-0.014	0.005
Signed discretionary accruals	-0.004	-0.004	0.950	0.440	0.003	-0.015	0.002
Restatements	-0.010	-0.009	0.980	0.370	0.005	-0.026	0.003
Going concern opinions	0.020	0.020	1.000	0.780	0.006	0.004	0.041
Audit Fees	0.322	0.319	1.000	1.000	0.025	0.249	0.453

Panel C: Propensity Score Matching (All Selection Variables Balanced)

Audit Quality Measure	*Mean*	*Median*	*PCT Sign*	*PCT Significant*	*Std. Dev.*	*Minimum*	*Maximum*
Absolute discretionary accruals	-0.007	-0.007	0.970	0.740	0.003	-0.015	0.003
Signed discretionary accruals	-0.004	-0.004	0.810	0.320	0.004	-0.018	0.007
Restatements	-0.010	-0.009	0.880	0.350	0.009	-0.041	0.009
Going concern opinions	0.018	0.019	0.970	0.670	0.007	-0.010	0.033
Audit Fees	0.322	0.326	1.000	0.980	0.048	0.011	0.418

This table presents the proportion of matched samples drawn from the bias-variance frontier (see Figure 3) that result in a Big N treatment effect that is in the predicted direction and significant at p < 0.05 (one-tailed), where *PCT Sign* is the fraction of treatment effects that are in the predicted direction, and *PCT Significant* is the fraction of treatment effects that are in the predicted direction and significant at p < 5% (one-tailed). We create 100 percentile rank groups based on the number of treatment firms and classify the matched samples with the lowest L_1 imbalance in each group as being on the bias-variance frontier. Thus, using this approach we obtain 100 matched samples that have the lowest imbalance for a given number of treatment firms. Using these matched samples we compute the summary statistics presented in this table. Panel A presents the results when we use the matched sample from the PSM analysis presented in Figure 2. Panels B and C present the results after we repeat the analyses presented in Figure 2 after requiring matched samples to be balanced with respect to respectively firm size or all selection variables. We use two-sample t-tests to assess whether matched samples are balanced with respect to firm size. In order to assess whether matched samples are balanced with respect to all selection variables we regress the matching variables on the Big *N* indicator and classify matched samples as being balanced when an *F*-test cannot reject the null hypothesis that the matching variables do not explain the use of a Big *N* auditor at p < 10.

Table 5

Big N Treatment Effects from Regressing Five Measures of Audit Quality on Big N from 3,000 PSM Matched Samples Randomized Over Three Research Design Choices after Randomly Adding Five Additional Matching Variables.

Audit Quality Measure	*Mean*	*Median*	*PCT Sign*	*PCT Significant*	*Std. Dev.*	*Minimum*	*Maximum*
Absolute discretionary accruals	-0.006	-0.006	0.992	0.887	0.003	-0.032	0.023
Signed discretionary accruals	-0.005	-0.005	0.961	0.623	0.003	-0.041	0.014
Restatements	-0.009	-0.009	0.950	0.255	0.006	-0.046	0.019
Going concern opinions	0.021	0.020	0.996	0.744	0.006	-0.008	0.053
Audit fees	0.313	0.309	1.000	0.998	0.030	0.065	0.703

This table presents summary statistics for the estimated Big N treatment effects plotted in Figure 5, where *PCT Sign* is the fraction of treatment effects that are in the predicted direction, and *PCT Significant* is the fraction of treatment effects that are in the predicted direction and significant at p < 5% (one-tailed). The predicted sign is negative for absolute DAC, signed DAC, and restatements, positive for going concern opinions and audit fees.

Table 6

Big N Treatment EffectsfFrom 3,000 CEM Matched Models Randomized Over The Level Of Coarsening

Audit Quality Measure	*Mean*	*Median*	*PCT Sign*	*PCT Significant*	*Std. Dev.*	*Minimum*	*Maximum*
Absolute discretionary accruals	-0.004	-0.004	0.945	0.563	0.002	-0.010	0.003
Signed discretionary accruals	-0.003	-0.003	0.907	0.328	0.002	-0.011	0.004
Restatements	-0.018	-0.018	0.998	0.756	0.007	-0.055	0.001
Going concern opinions	0.019	0.019	1.000	0.878	0.004	0.005	0.037
Audit Fees	0.308	0.306	1.000	1.000	0.023	0.243	0.398

This table presents summary statistics for the regression results presented in Figure 6, where *PCT Sign* is the fraction of treatment effects that are in the predicted direction, and *PCT Significant* is the fraction of treatment effects that are in the predicted direction and significant at p < 5% (one-tailed).