

Analysis of Non-Random Samples In Auditing Research

Summary

The use of choice-based, matched, and stratified sample designs is common in auditing research. Earlier research analyzed these samples as if the data were randomly sampled. However, data gathered from choice-based, matched and stratified samples are not random samples. Data analysis for choice-based, matched and stratified samples may be improved by the use of analysis assuming non-random sampling of data. Using non-random data analysis might improve the power of the tests used in the analysis and may give us better information about the variables of interest in the models. We review choice-based, matched and stratified sample auditing papers in three areas of audit research from 1980 to 2003 and use these papers to illustrate how nonrandom data analysis can be used to improve the external and internal validity of choice-based, matched and stratified samples.

1. INTRODUCTION

Statistical methods used in analysis of auditing data often assume that the data is gathered by random sampling from a population. Data gathered in non-random sampling, including stratified sampling, matched samples or choice-based research designs, requires different statistical methods for analysis. This paper examines prior research in three contemporary audit research fields—auditor practices, audit litigation, and management fraud—and considers errors that might occur in data analysis when the statistical assumption of data gathered from a random sample is violated and describes what the researcher can do to avoid these errors in sample selection and analysis. Archival auditing research in these three areas has moved beyond its infancy and is gaining critical mass. It is time for more sophisticated and appropriate analyses on nonrandom research designs for studies in these areas to have an impact on auditing practice or proposed regulation. The purpose of this paper is to provide some direction in terms of the more sophisticated analyses that might be appropriate for these research fields.

Prior research (e.g. Cram, Karan, Stuart, 2008) has identified research design categories using choice-based and matching techniques in accounting research. They show that choice based and matched samples have often been analyzed assuming data is randomly sampled, when in fact the assumption is not justified. This paper continues the discussion of appropriate analyses of nonrandom research data using three areas of auditing research to illustrate the analysis. These three areas have been chosen because they are important areas for future auditing research. The areas are: audit practices, determinants of audit litigation, and detecting management fraud. For each area, we discuss prior nonrandom research studies and provide specific suggestions for more sophisticated research methods that might be used to advance knowledge in each area. We also discuss several *tactical errors* that might limit explanatory power when used in research studies.

The paper is organized as follows. The next section describes common motivations for use of non-random samples and explains errors that might occur if nonrandom samples are analyzed as random samples. The main part of this paper provides discussion of the three

audit fields and discussion of research methods helpful in advancing knowledge in these areas. Finally, we summarize and conclude.

2. NONRANDOM RESEARCH DESIGNS IN AUDIT RESEARCH

Choice-based, matched, and stratified samples are nonrandom research designs used to economize on data collection. If some variables require hand-collection or other costly investment to gather, but some other variables are freely or more cheaply available, then one can be strategic in choosing the observations on which to collect complete data. Choice-based sampling is especially useful when particular types of outcomes are rare and few of these rare outcomes would be obtained under random selection. For example, Mutchler (1985) identifies all 119 manufacturing firms that received a going concern qualification during an 11 month period and collects the data for all of those. She then compares those to a randomly selected sample taken from all manufacturing firms having financial distress but not receiving a going concern qualification.

A refinement of the choice-based sampling approach is to use pair-matching in the selection of the control sample. For example, all firms experiencing auditor litigation during a period may be identified, and for comparison, rather than randomly selecting from all non-litigation firms, a control sample containing matched firms, matching one or a few firm-year observations to each litigation firm by industry, year, and firm size, may be collected. Stice (1991), Lys and Watts (1994), Shu (2000), and Heninger (2001) all follow this strategy. This form of analysis is appropriate if matching factors such as industry, year, and firm size are likely to have a large effect on the likelihood of the explained outcome (audit litigation) but are not themselves of primary research interest. In these circumstances, the use of a matched sample design allows the researcher to focus power on estimating parameters for variables of interest while applying control for those matching variables. Use of a matched sample is especially convenient when a matching variable may have a nonlinear effect, as it suffices just to match on it without modeling and estimating its nonlinear effect explicitly.

Another type of nonrandom design is one that is merely stratified, meaning that samples from different subpopulations are selected at varying sampling rates. “A **stratified**

random sample is one obtained by separating the population elements into non-overlapping groups, called strata, and then selecting a simple random sample from each stratum.”

(Scheaffer, Mendenhall, Ott, 1979, p. 59). A distinguishing characteristic of stratified samples is that, although the selection is not evenly random, there is positive probability of selection for all observations in the population. For appropriate analysis, it will be essential that sampling rates for each stratum be noted.

Previous researchers (e.g. Cram, Karan, and Stuart, 2008) have identified three errors that can apply when nonrandom samples are analyzed as random samples. We discuss these errors briefly below.

Error 1: The use of unconditional analysis, when analysis conditional upon effects of matching variables is needed. This is the most serious error that has been identified. A correct, conditional analysis of data with matched pairs could involve regressing pair-wise differences in outcome upon pair-wise differences in explanatory variables. A pooled regression that does not take into account pairings is an **unconditional, incorrect** analysis, for a matched pairs sample. When researchers analyze nonrandom designs without considering the conditional nature of the data, the analyses lack internal validity. There is no plausible set of statistical assumptions which would justify the standard errors and t-statistics reported in analysis suffering Error 1.¹ **Solution:** Researchers avoid Error 1 by analyzing pairwise differences of all variables, or by including a dummy variable for each matched pair or set in the analysis.²

¹ A rare exception would be an OLS regression analysis where the matching variables were both independent of the dependent variable and independent of the included independent variables. Then, omission of the matching variables could not bias the coefficients of included variables. However, we observe no case where matching variables are unlikely to meet such standard; on the contrary, we find researchers' arguments to be convincing that the industry, size, and other variables that they use in matching would relate to the outcomes they study and that these variables are usually correlated with the financial ratios and other variables that are included in their models. For logit regression and other nonlinear analysis, even if the matching variables are uncorrelated, the estimated coefficients would be biased.

² In very general terms, appropriate reweighting can be used to avoid Error 1, but is hard to apply. For example, a pair-matched sample can be analysed by a pooled regression analysis without taking pair-wise differences, if each observation is reweighted appropriately in the analysis. That is, each observation would be reweighted proportionally to the size of the subpopulation from which it is drawn (e.g., the subpopulation of non-bankrupt firms of a certain size category in a certain industry). But to set the weights correctly, it would be necessary to know a great deal about the population from which the sample is drawn, specifically the prior probability that each observation taken would be selected. Each observation could have a different appropriate weight. And, the reweighting would effectively undo the desired effect of selecting the most informative observations for analysis.

Error 2: Failure to control for effect of imperfectly matched variables. This applies to analysis of fully-matched samples where matching is between firms that are matched by *closest* size. Residual differences in size (or another closest-matched variable) can still influence the result. The effect of these differences might be controlled for by including a linear term. But as size or another variable's contribution might be non-linear, in general, there is no fully satisfactory resolution. The researcher should make some effort to examine the possibility that the results are driven by the omitted residual effect. Research that fails to consider the effect of differences remaining when matching by closest size may lack internal validity. The main statistical assumption which would justify excluding the residual effect (if matching by size) is that size has no influence upon the outcome (in which case it is not helpful to match upon it). In an OLS regression, the omission of residual size would also be justifiable if size was not correlated with any included variables (it is only omitted correlated variables that cause bias in all other coefficients); however in logit, probit and other nonlinear analyses, the omission of even an uncorrelated variable that should be in the model will bias estimates of all other coefficients. **Solution:** Perform sensitivity analyses using linear and quadratic terms for the "closest" match variable. If there is no difference in the data results between the original model and the model with linear and quadratic terms, then the remaining difference in the "closest" matched variable appears not to affect the results and you can submit that as evidence on the internal validity of your research study.

Error 3: Failure to reweight observations according to differing sampling rates. In auditing research, we note just a few studies employing weighting to correct for such problems. **Solution:** One way to re-weight observations is to use a weighted exogenous sampling likelihood method like WESML in SAS. A limitation of this method however is that exogenous population information is needed. In many research settings the required information—the proportions of all individuals from the population that fall in each of the matching set categories—might be readily available, such as when sampling from the finite population of publicly traded firm-year observations appearing in CRSP and Compustat databases, but often it is not.

A variant to Error 3 has to do with what we term a “choice-based logit exemption” to the need for reweighting. This exception applies to logit regression models explaining a choice variable that itself is used in guiding sample selection, as long as fully saturated models—ones which account for matching and hence avoid Error 1—are used. Those who used the guidance provided by prior researchers (e.g. Palepu, 1986; Maddala, 1991; and Zmijewski, 1984) often did not recognize that if matching is used within choice-based samples, it is then necessary for an intercept to be estimated for each matched set, i.e. for the model to be fully saturated, hence avoiding Error 1. In estimation, each of those intercepts on matching variables is likewise biased, even though their presence is necessary to enable consistent estimation of coefficients on variables of research interest. However, logit regressions that explain a dependent variable other than the groupings of the data into case and control samples, i.e. that are not choice-based, do not benefit from the exemption to reweighting. The choice-based logit exemption to the need for reweighting applies only rarely.

Research Design or Analysis Errors

In addition to the three errors identified in previous research (e.g. Cram, Karan, and Stuart, 2008), we expand the list of errors that may occur in non-random research designs to describe what we term “tactical errors.” Tactical errors are errors that reduce the *power of the analysis* performed by the researchers. They may not lead to biased conclusions, but they will reduce the power of the test and make it less likely for the researcher to find results. These tactical errors include: (1) over-reliance upon univariate analyses, (2) the belief that equal sized samples are needed in studies using choice-based or matched sample research designs, (3) using matched samples for comparison when the statistic of interest relates to a sample-level quantity, and (4) inconsistency between the form of the matched variable and the form of the variable in the model. Each tactical error is discussed briefly below.

Univariate analysis of each variable that might explain a discrete outcome, e.g. as might be presented in a t-test comparing means across two subsamples, does not allow a researcher to control for other variables and provides inconclusive results in situations where many variables impact the result. Some researchers may present univariate analysis because they

appreciate that multiple variable methods commonly employed do not control for matching, while they are unaware of the appropriate multiple variable methods to use. The only statistical methods that avoid Error 1 and are taught in introductory statistics textbooks are univariate tests such as matched pair *t*-tests and Wilcoxon signed rank tests. Other researchers use univariate comparisons to argue that their matching worked well, prior to presenting multiple variable analyses. Previous research (e.g. Cram, Karan, and Stuart 2008) has suggested that such comparisons do not accomplish what the researchers intend and that little is provided by univariate analysis preliminary to presentation of multivariate analysis in matched sample studies. In this situation, the researcher should simply present the multivariate analysis without the preliminary step of the univariate analysis.

A second tactical error made by researchers is the throwing away of data to get equal size samples. In matched pairs, by construction the case and control samples are of equal sizes. But it is not necessary to have equal sized case and control samples for choice-based samples. For example, Mutchler (1985) collects only as many control observations out of 684 available distressed firms as the 119 case observations of going concern qualification that are available to her. In general, there are declining returns to collecting additional data of either type, and in general it is more useful to collect data of the less numerous type. But there are still benefits to collecting more data, even of the more numerous type, and the decision to stop collecting should be based on the marginal cost of collecting more data and the marginal benefit. Equally sized subsamples in matched analysis are not necessary either. Shu (2000) collects 10 control observations for each case observation, and that is more informative than collecting just 1 control for each case. Varying numbers of cases and controls can be accommodated. It is not necessary to collect only 10 controls in any subpopulation where more control observations are freely available. Dropping data of one type to make two samples seem “more comparable” simply reduces the power of the analyses that can be conducted, and introduces the necessity of accounting for the non-random nature of the reduced data. (Dropping data to make the two samples seem more similar doesn’t make the analysis valid.)

A third tactical error is the selection of matched samples for comparison, when a sample-level quantity such as R-squared (that explains the coherence of data or the explanatory power of models in one regime vs. in another regime) is of interest to the researcher. While a Vuong test is one way to statistically test the difference of explanatory power of two models in the same data, testing in which of two data regimes one model performs better is not a traditional statistical test. The performance check testing explanatory power of a model in two data areas might possibly be examined by a bootstrapping approach. However, we know of no way to account for the matching in that analysis of difference in explanatory power across two data regimes, and all analyses omitting matching are invalid. So, matched samples should not be used in this situation because they cannot be properly analyzed.

A fourth tactical error is inconsistency in the form of the variable between the match selection phase and the model estimation phase of analysis. For example, a number of researchers choose to use “closest matching” on assets when selecting control firms to fully match case firms, but then include a transformed version of the variable, i.e. $\log(\text{assets})$, in the model estimation. Their insight that the transformed version is more appropriate in the model is probably correct, but then the match selection should also have been conducted using the transformed version. It would be more appropriate to select firms that are closest in $\log(\text{assets})$.

3. DISCUSSION OF AUDIT RESEARCH STUDIES BY RESEARCH STREAM

Prior research from 1980-2003 using nonrandom research designs in three audit research areas will be examined to illustrate how data analyses designed for nonrandom samples might be used to increase explanatory power and assure the continued contribution of choice-based and matched samples in these research streams.

Research Stream 1: Auditor Practices

In the auditor practice literature, researchers address issues related to measuring auditor practices in ways useful for exploring various questions. Auditor practices have been

characterized in various ways. Kinney (1983) measures audit structure and examines the impact of structured or nonstructured audit approaches on the judgments made by auditors. Kreutzfeld and Wallace (1986, 1990, and 1995) examine the content of audit work papers.

Papers in this category use actual audit work papers, publicly available information, or surveys to study issues in audit practice. Seven nonrandom design research papers appear in this stream in the time period identified. Table 1 provides additional detail on the specific research questions in each study, the sample selection, the principal analyses, and the errors present in the study.

Kinney and McDaniel (1993) study audit delays for matched pairs of firms with opposite-matching on whether they report a correction of previous earnings or not, with similarity-matching by industry, sign of earnings change, and closest size. Audit delay is defined as the difference between the year-end close of the accounting period and the date of release of the audit report, i.e. the audited financial statements. This is a between-subjects fully-matched sample that is not choice-based (as the outcome variable studied—audit delays—is not used to guide sample selection). This study accounts for pair-matching and hence it successfully avoids Error 1. The analysis utilizing the pair-matching is essentially an OLS regression of pairwise differences in audit delay upon pairwise differences in explanatory variables.³ The imperfection in size-matching is not controlled for in the analysis. Nonrandom research methodology suggests that in this situation, the imperfection in size-matching could be considered in the model by including pairwise difference in size as a control variable. Without this modification in data analysis, we are left with some doubt as to whether the omitted variable of residual pairwise differences in firm size could partially explain pairwise audit delays and be correlated to included variables. Nonrandom research

³ Actually, one explanatory variable is not a pairwise difference, but rather is a difference between the correcting firms' returns and the market return. This exception undermines the interpretability of the work; we believe that it would have been preferable for the returns variable to be computed in a comparable way.

methodology further suggests that reweighting for differences in population sizes is needed to improve the generalizability of the studies results. Reweighting can be used to improve the external validity of the research study.

Lawrence and Glover (1998) is a study of changes in audit delays, i.e. the delay between fiscal year end and release of audited financial statements. This is a within-subjects fully-matched sample study that is not choice-based (as the outcome variable studied—audit delays—is not used to guide sample selection). Audit delays are measured for 204 audit clients of Big8 firms in 1986 and for the same clients again in 1991. Overall, average audit delays are reduced from 34 to 31 days. The research focus is on whether merged vs. non-merged auditors make more progress in reducing delays, which is explored by univariate t-tests of changes in audit delays within subsamples. Our reading of the researchers' presentation is that they employed unmatched, two-sample t-tests that did not take into account the natural matching of each client firm in 1986 with itself in 1991. Perhaps only for lack of appropriate analysis, they are left to conclude that their main theory is not supported. The tactical error of relying too heavily on unmatched univariate tests could be avoided in analysis of this type by applying matched tests. The researchers fail to find a statistically significant change in audit delay for the merged firms, but that could merely be the effect of using the unmatched test rather than the matched test that will increase the power of the test. The main research question of the study might have been addressed by a single test of differences-in-changes (i.e., whether audit changes were different across merged vs. non-merged auditors) but the researchers did not construct such a test.

A similar paper is Elder and Allen (2003) that examines trends in auditor risk assessments and audit sample size decisions. Audit tests in 1994 are matched by audit client, by type of test (accounts receivable confirmation, inventory test count, inventory price test) to audit tests in 1995. Primary analyses are unmatched OLS regression analyses of audit sample

sizes within each type of test (pooling across three audit firms) and within each of the three audit firms (pooling across type of audit tests). Results include unexpected findings that measures of inherent risk and control risk mostly appear unrelated to sample size. Although these regressions included a year dummy, they did not include the 35 client dummies that are needed to control for the client matching. Nonrandom data analysis methods suggest that the researcher should include 35 client dummies to answer the question of whether inherent and control risk is related to audit sample sizes. The failure to include client dummies may have reduced the power of the test to establish a relationship between audit sample size and inherent or control risk.

Sweeney and Summers (2002) examine employee “burnout” during auditors’ busy season. Their analysis of before and after survey data collected from auditors does not take into account the natural matching of each auditor with herself or himself (identifying information was collected on both surveys to enable matching up each auditor’s responses). They employed a sophisticated method, structural equation modeling using AMOS software. Their model has “before” and “after” constructs, each based on two or three component measures which are compared, but the comparison does not account for matching. To allow for comparisons that account for matching and consider the nonrandom nature of the data, each major construct could have been composed as a difference, itself, based on components that are already converted to pair-wise differences. The analysis in the paper suffers from the omission of 142 correlated indicator variables.

Kreutzfeldt and Wallace (1986), Kreutzfeldt and Wallace (1990), and Wallace and Kreutzfeldt (1995) use a sample from Arthur Anderson of 260 audit clients constructed by stratified sampling in strata of size, industry, and publicly vs. privately held status. They examine types and causes of errors in client firms’ accounting processes that necessitate auditors making correcting journal entries. The authors stated that they wanted a more

representative sample than one chosen by pure random selection. Because they did not explicitly sample at the same rate in each grouping, it is implicit that they over-sampled in some strata and under-sampled in others. Although this study is not choice-based and the manner of selection does not involve matching, we include the study as an example of a stratified sample which is a nonrandom sample requiring the researchers to use reweighting of data to generalize the results of the study to the population. For conclusions in these three papers to be generalizable, technically, reweighting to the proportions observed in the general population is needed but not performed. If this data is not available, researchers can only note that their data is not randomly selected and that generalizations are subject to question for that reason.

In summary, analysis appropriate to nonrandom research designs would be easy to implement in all cases, with the possible exception of re-weighting if the population totals are not known. To avoid error 1, the researcher should include a dummy variable for each pair, or use differences between the pairs as the variable in the model. This is a simple matter of model construction. To avoid error 2, the researcher should either include the “closest” size variable in the model, or use sensitivity analysis to consider whether the difference in the matching variable has an impact on the data results. This again is an easy step in the data analysis process. To avoid error 3, the researcher should reweight the samples from the two populations to the totals in the populations. This is a fairly simple process if the population totals are known. Incorporating matching and reweighting into the analyses would lessen the issues of internal and external validity for these studies.

Research Stream 2: Determinants of Auditor Litigation

Four papers using nonrandom samples in the time period of our study investigate determinants of auditor litigation using choice-based samples. Table 2 provides additional

detail on the specific research questions in each study, the sample selection, the principal analyses, and the errors present in the study.

Stice (1991) creates two semi-matched samples for comparison to 49 firms having auditor litigation: one is matched on year only, the other is matched by year and by industry, and then one is selected randomly, achieving fewer than 49 year-industry sets, so we categorize this as semi-matched. Stice throws away the information contained in his sample that would be available for a choice-based analysis by replacing valid identification of which firms were sued by randomly generated 1's and 0's. His probit analysis explaining these random numbers, indicates, surprisingly, that hypothesized relationships between variables and the audit litigation outcome are reflected by relationships found between the variables and the random numbers.⁴ We attribute Errors 1 and 3 to the analysis as, even if auditor litigation rather than random values had been employed as the dependent variable, the probit analysis failed to account for the matching and would not have benefited from the choice-based logit exemption to the need for reweighting. Stice might have included dummy variables for each pair or used pair-wise differences in his model. Both changes would exploit the fact that the data selected for the sample was nonrandom and would have been easy changes to make in the models. Reweighting might have been possible if the population totals where the samples were drawn were known.

Lys and Watts' (1994) sample consists of 163 firms whose auditors were sued and 163 firm-year observations pair-matched by year, industry, Compustat delisting code (if any), and then by closest size (assets). They report OLS regressions explaining auditor litigation or not using analysis methods based on random selection of data. Statistical analysis appropriate for

⁴ We note that many of the variables entering into Stice's model seem plausible, and their sign of entry may accord with researchers' intuition. However, the results he obtained have no more validity than those that would be obtained if the opposite random number sequence substituting 1 for each 0 and vice versa had been drawn, in which case every variable would have received exactly the same magnitude and apparent statistical significance but the opposite sign.

nonrandom sampling would have been easy to implement in this study. The authors could have added a dummy variable for each pair to their model or have used pair-wise differences as a variable in the model. Lys and Watts partially control for imperfection in their matching on size by including $\log(\text{size})$ as a variable in their analysis, mitigating concern for Error 2 in their work, although matching on $\log(\text{size})$ upfront would be more internally consistent with this step and would increase the power of their analysis (avoiding the tactical error of matching on a different variable than is included in the model.)

Shu (2000) uses a list of audit litigation firms from Ross Watts, augmented by additional searching, yielding 282 firm-year observations, and constructs a semi-matched sample by selecting 10 controls matched by year, for each litigation outcome. Her analysis suffers Error 1 for failing to include year dummies. It would be quite easy to analyze the choice based sample as a nonrandom sample by adding a dummy variable for each year to the model. This addition would correctly account for the conditional analysis appropriate when data is selected in a nonrandom fashion, by selecting a sample conditional on the litigation outcome. It is noted that 80.2% classification performance is achieved, however, we note that this predictive performance is less accurate than the 91% accuracy of a naïve “predict zero” model, given that 91% of her observations are non-litigations by construction.

Heninger (2001) identifies 67 auditee firms with lawsuits against auditors, and semi-matches to 67 firms by year and industry, achieving somewhat fewer than 67 year-industry sets. The authors argue convincingly that industry and/or year are important, but the data analysis has been done as if the sample were a random sample. The researcher could consider the nonrandom nature of the data by including a dummy variable in the model for each matched pair of firms.

All four of these papers use *random sampling methods* of data analysis when more contemporary research methods employing *nonrandom data analysis* might be useful in

advancing the questions of interest in this field. The industry and year matching is important. As shown by prior researchers (e.g. Cram, Karan, and Stuart 2008), the effect of omitting these correlated variables from analysis is severe. Further, since saturated models are not employed in the logit analyses, they do not enjoy the logit exemption for reweighting.

The impact of erroneous analysis in these papers carries through many other papers that rely upon them for their measurement of audit litigation risk. Shu's misanalysis of litigation determinants undermines her analysis in the same paper of causes of auditor resignations. Krishnan and Krishnan (1997), a paper that suffers Error 1, relies upon Stice's reported coefficients for a Z-score-type index of auditor litigation likelihood. This application is inappropriate for several reasons. If Stice's sample selection had not involved matching and he had performed a nonmatched analysis explaining audit litigation (which he did not), then Stice's coefficients other than the intercept would have been unbiased and higher index values would have fairly represented higher auditor litigation likelihood. Use of an unconditional index based on the conditional estimates from a matched sample (assuming Stice had explained audit litigation as his dependent variable) is, however, invalid. For one, the index omits year-industry-specific intercepts that would likely be important (as Lys and Watts, Stice, Shu, and Heninger do argue that industry and year do matter). And even if Stice had explained audit litigation as his dependent variable and accounted for the matching by including year-industry dummies in the logit regression, the intercepts estimated would not be valid, for the same reason that the overall intercept in logit regression of choice-based samples is not valid. Estimation of those intercepts for use in an index requires a random sample, or reweighting of each observation according to the sampling rate in each outcome-year-industry vis-a-vis the general population.⁵

⁵ A conditional index omitting intercepts could indeed be constructed that would be valid in narrow circumstances: in applications where fixed effects for the industry are included, and where the industry

Nonetheless, would such an index be informative about litigation likelihood, in the same way that Altman's Z-score turns out to be informative about bankruptcy, although it is equivalently miss-specified? The answer to that is unclear: Altman's Z-score has been proven to be informative by classification performance tests by Altman and many others. But an auditor litigation score has not been proven informative in such ways. Further, it is arguable that industry-year effects could be relatively stronger for auditor litigation and omitting them would undermine the scores' validity in two ways: first the direct omission of the important effect of industry and year; second the coefficients included in the model would be more severely miss-estimated.

In a recent accounting study, Krishnan and Zhang (2005) similarly use a Z-score-like measure of audit litigation risk, but one based on Shu (2000)'s coefficients instead of Stice's. Again it is incorrect to use the *conditional* estimates from Shu to construct an *unconditional* index of audit litigation risk, one which necessarily omits the year effect that Shu's work deemed important, besides the fact of Error 1 in Shu's work.

We are aware of no study of audit litigation risk that does not both use a choice-based and matched sample and suffer Error 1, the most severe error. Hence it seems that the impact of misanalysis of choice based and matched samples significantly affects the estimation of audit litigation risk, which continues to influence research in corporate governance and other areas.

Research Stream 3: Detecting Management Fraud

Research in the area of fraud detection is designed to identify factors present when fraud occurs. This information would be useful for managers interested in preventing fraud, for auditors responsible for designing the audit so material fraud is found, and for the regulators making institutional changes likely to lessen the occurrence of fraud within a

categorization is the same as in Stice's estimation setting. But we note that Krishnan and Krishnan's setting does

company. We identify four nonrandom design research papers during the time period of the study. Table 3 provides additional detail on the specific research questions in each study, the sample selection, the principal analyses, and the errors present in the study.

Green and Calderon (1995) identify 86 instances of 10-K statements subsequently found to be fraudulent, and match to 86 non-fraud firm year observations, with matching by industry, year, and “comparable” size. The work tests whether simple analytical procedures in the early stage of an audit are effective in signaling management fraud. Their analysis is by univariate ranked sums of differences and takes matching into account, but suffers Error 2 for not addressing residual difference in size, and suffers Error 3 because reweighting is not done that would allow the results to generalize to any larger population.

Beasley (1996) is an oft-cited paper from the management fraud research stream. The data in the paper consists of 75 matched pairs of fraud and non-fraud firms matched by year, stock exchange, size (market value within 30%), and industry to consider whether including a larger proportion of outside members on the board of directors reduces the likelihood of fraud. The results indicate that firms without fraud have a significantly higher percentage of outside board members than firms with fraud. Audit committee presence was found *not* to be significant, seemingly contrary to Beasley’s prior beliefs, perhaps due to the presence of Error 1 in the analysis. Section 301(3)(A) of the Sarbanes-Oxley Act of 2002 (SOX) requires that there be an audit committee (and that its members be independent), rather than making larger requirements on the independence of its board of directors, hence implementing somewhat the reverse of what was implied by this paper. It seems possible that, in reanalysis, Beasley’s main result, that percentage of outside directors affects the likelihood of financial fraud, could be reversed, and that the presence of audit committee may be found to be a significant factor

if improved data analysis models were used employing a nonrandom analysis of choice based data.

Abbott, Park, Parker (2000), in a similar matched sample study employing unmatched analysis, find contrasting results to those of Beasley (1996) in that they “do not find a significant relation between fraud and the proportion of outside directors” (p. 61). They attribute the difference in results to sample differences, while we note their analysis suffers all three errors and hence we view their results as unreliable.

Green and Choi (1997) use the sample of firms from Green and Calderon (1995) to apply a neural network. They assert that their neural network model achieves higher reliability levels than previous models developed to detect the presence of fraud in financial statements. However, their approach appears to pool the data and not account for the pairings; hence it suffers Error 1 in addition to Errors 2 and 3.

Bell and Carcello (2000) is a study that develops a model to estimate the likelihood of fraudulent financial reporting for an audit client based on the presence or absence of fraud-risk factors. The researchers identified 7 factors from the list of 46 fraud risk factors that were associated with the likelihood of fraudulent financial statements. Because this paper suffers from Error 1, the variables identified as significant or insignificant may change when the data is analyzed using nonrandom models consistent with the nature of the data.

The five papers in this stream try to determine fraud factors associated with management fraud. We question in each paper whether fraud factors have been correctly identified due to the use of data analysis that assumes random sampling of data when the data has been collected based on nonrandom sampling.

4. SUMMARY AND CONCLUDING REMARKS

This article examined the use of choice based, matched, and stratified sample designs in three research streams and suggested new research methods that may be appropriate in advancing knowledge in each of these research fields. Prior research notes that when non-random samples are collected, they must be analyzed taking their non-randomness into account. If this is not done correctly, coefficient estimates are inconsistent and may be attenuated, exaggerated, or even reversed from their true values. When less sophisticated analysis is used, the estimated coefficients' signs, magnitudes, and significance are unreliable. Thus both the internal and external validity of the research may be questioned and the need for better research methods increasing the internal and external validity of the research is apparent.

Revisiting earlier papers in research streams may be important. When less accurate research methods are applied to research questions, research streams may not reach a critical mass of knowledge as quickly as possible with more accurate research methods. More sophisticated research methods may be called for to move research to the next step.

REFERENCES

- Abbott, L. J., Y. Park, and S. Parker. 2000. The Effects of Audit Committee Activity and Independence on Corporate Fraud. *Managerial Finance* 26 (11): 55-67.
- Beasley, M. S. 1996. An Empirical Analysis of the Relation Between the Board of Director Composition and Financial Statement Fraud. *The Accounting Review* 71 (4): 443-465.
- Bell, T. B., and J. V. Carcello. 2000. A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. *Auditing: A Journal of Practice & Theory* 19 (1): 169-184.
- Cram, D., V. Karan, and I. Stuart. 2008. Three Threats to Validity of Choice-Based and Matched Sample Studies in Accounting Research. *Contemporary Accounting Research* forthcoming.
- Elder, R. J. and R. D. Allen. 2003. A Longitudinal Field Investigation of Auditor Risk Assessments and Sample Size Decisions. *The Accounting Review* 78 (4): 983-1002.
- Green, B. P., and T. G. Calderon. 1995. Analytical Procedures and Auditors' Capacity to Detect Management Fraud. *Accounting Enquiries* 5 (1): 1-48.
- Green, B. P., and J. H. Choi. 1997. Assessing the Risk of Management Fraud Through Neural Network Technology. *Auditing: A Journal of Practice & Theory* 16 (1):14-28.
- Heninger, W. G. 2001. The Association between Auditor Litigation and Abnormal Accruals. *The Accounting Review* 76 (1): 111-126.
- Kinney, W. R., and L. S. McDaniel. 1993. Audit Delay for Firms Correcting Quarterly Earnings. *Auditing: A Journal of Practice & Theory* 12 (2): 135-142.
- Kreutzfeldt, R.W., and Wallace, W. A. 1986. Error Characteristics in Audit Populations: Their Profile and Relationship to Environmental Factors. *Auditing: A Journal of Practice & Theory* 6 (1):20-43.
- Kreutzfeldt, R.W., and Wallace, W. A. 1990. Control Risk Assessments: Do They Relate to Errors? *Auditing: A Journal of Practice & Theory*.
- Krishnan, J. (Jagan), and J. (Jayanthi) Krishnan. 1997. Litigation Risk and Auditor Resignations. *The Accounting Review* 72 (4): 539-560.
- Krishnan, J. (Jagan). and Y. Zhang. 2005. Auditor Litigation Risk and Corporate Disclosure of Quarterly Review Reports. *Auditing: A Journal of Practice & Theory* 24(Supplement): 115-138.
- Lawrence, J. E. and H. D. Glover. 1998. The Effect of Audit Firm Mergers on Audit Delay. *Journal of Managerial Issues* 10 (2): 151-164.
- Lys, T. and R. Watts. 1994. Lawsuits Against Auditors. *Journal of Accounting Research* 32 (Supplement): 65-93.
- Maddala, G. S. 1991. A Perspective on the Use of Limited-Dependent and Qualitative Variables Models in Accounting Research. *The Accounting Review* 66: 788-807.
- Mutchler, J. 1985. A Multivariate Analysis of the Auditor's Going-concern Opinion. *Journal of Accounting Research* 23 (2): 668-682.
- Palepu, K.G. 1986. Predicting Takeover Targets: a Methodological and Empirical Analysis. *Journal of Accounting and Economics* 8: 3-35.
- Scheaffer, R. L., W. Mendenhall, and L. Ott, 1979. *Elementary Survey Sampling*, 2th edition. Boston: Duxbury Press.

- Shu, S. Z. 2000. Auditor Resignations: Clientele Effects and Legal Liability. *Journal of Accounting and Economics* 29: 173-205.
- Stice, J. 1991. Using Financial and Market Information to Identify Pre-Engagement Factors Associated with Lawsuits Against Auditors. *The Accounting Review* 65: 516-533.
- Sweeney, J. T. and S. L. Summers. 2002. The Effect of the Busy Season Workload on Public Accountants' Job Burnout. *Behavioral Research in Accounting* 14: 223-245.
- Wallace, W. A. and R. W. Kreutzfeldt. 1995. The Relation of Inherent and Control Risks to Audit Adjustments. *Journal of Accounting, Auditing and Finance*: 459-481.
- Zmijewski, M. 1984. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research* 22: 59-82.

Table 1
Non-random Research Design Papers in Audit Practices Stream
1980-2003

Paper Author(s), Journal, Title	SSCI citation count in 2007	Choice-Based and/or Matched Sample Selection		Principal Analyses Explaining Choice and/or Using Matching	Analyses Incorp. Matching/Stratification?		Additional control for "closest" matching		Reweight-ing for generalizing	
		Identification of Cases, or Stratified Sample	Selection of Controls or Comparison Sample, If Any		Needed ?	Applied ?	Needed ?	Applied ?	Needed ?	Applied ?
Kreutzfeldt and Wallace (1986) <i>Auditing: A Journal of Practice & Theory</i> , "Error Characteristics in Audit Populations: Their Profile and Relationship to Environmental Factors"	62	Identify 260 Arthur Anderson clients, stratified in sampling by client size, industry, and public vs. privately held	No comparison sample created.	Mean and standard deviation of number of audit adjustments as percentage of account balances; ANCOVA of same comparing strata	No	No	No	No	Yes	No
Kreutzfeldt and Wallace (1990) <i>Auditing: A Journal of Practice & Theory</i> , "Control Risk Assessments: Do They Relate to Errors?"	6	Use sample of Kreutzfeldt and Wallace (1986) Identify 260 Arthur Anderson clients, stratified in sampling by client size, industry, and public vs. privately held	No comparison sample created.	Correlations of control structure attributes with error rates and with severity of errors.	No	No	No	No	Yes	No
Kinney and McDaniel (1993). <i>Auditing: A Journal of Practice & Theory</i> . "Audit Delay for Firms Correcting Quarterly Earnings"	1	Identify 85 firms announcing, at year-end, corrections of previously reported interim earnings	From NAARS, fully match 85 firms by 4 digit SIC, sign of earnings change, having no extraordinary items, and closest in size (revenues).	OLS of paired differences in audit delay upon pairwise difference in explanatory variables.	Yes	Yes	Yes	No	Yes	No
Wallace and Kreutzfeldt (1995). <i>Journal of Accounting, Auditing and Finance</i> , "The Relation of Inherent and Control Risks to Audit Adjustments"	Not in SSCI	Use sample of Kreutzfeldt and Wallace (1990): select 260 Arthur Andersen audit clients by stratified sampling in groups by size, industry, and public versus private	No comparison sample created.	OLS regression models explaining number of audit adjustments, or in other words, error rate upon control risk factors	No	No	No	No	Yes	No
Lawrence & Glover (1998) <i>Journal of Managerial Issues</i> , "The Effect of Audit Firm Mergers on Audit Delay"	Not in SSCI	204 audit clients of Big8 auditors in 1986.	The same 204 audit clients, now of Big6 firms.	Univariate t-tests that appear to be unmatched, two-sample tests.	Yes	No	No	No	No	No
Sweeney & Summers (2002) <i>Behavioral Research in Accounting</i> , "The Effect of the Busy Season Workload on Public Accountants' Job Burnout"	Not in SSCI	Survey experienced auditors within one big firm, in 13 offices, of which 142 before and after responses received.	"After" responses form natural comparison sample to "Before" responses.	Structural equations model that does not incorporate matching of before and after pairings.	Yes	No	No	No	No	No
Elder & Allen (2003) <i>The Accounting Review</i> , "A Longitudinal Field Investigation of Auditor Risk Assessments and Sample Size Decisions"	1	Identify 35 clients of three audit firms where audit sample size data is available in both 1994 and 1999.	Same client firms in 1999	Unmatched regression of audit sample size on measures of inherent risk, control risk, and other explanatory variables.	Yes	No	No	No	No	No

Table 2
Non-random Research Design Papers in Auditor Litigation Stream
1980-2003

Paper Author(s), Journal, Title	SSCI citation count in 2007	Research Design Category	Choice-Based and/or Matched Sample Selection		Principal Analyses Explaining Choice and/or Using Matching	Analyses Incorp. Matching/Stratification?		Additional control for "closest" matching		Reweight-ing for generalizing	
			Identification of Cases, or Stratified Sample	Selection of Controls or Comparison Sample, If Any		Needed ?	Applied ?	Needed ?	Applied ?	Needed ?	Applied ?
Stice (1991). <i>The Accounting Review</i> , "Using Financial and Market Information to Identify Pre-Engagement Factors Associated with Lawsuits Against Auditors".	62	CB-SM	Identify 49 cases of auditor litigation, excluding financial and service firms.	Create two semi-matched samples from Compustat firms: one is matched on year only and then random selection, the other is matched on year and industry (SIC3).	Unmatched probit regression explaining random numbers (although explaining audit litigation or not was intended).	Yes	No	No	No	Yes	No
Lys and Watts (1994). <i>Journal of Accounting Research</i> , "Lawsuits Against Auditors". This paper was discussed by Jennifer Francis' "Discussion of Lawsuits against Auditors".	32	CB-FM	Identify 163 auditee firms whose auditors were sued during 1955-1994, and for which Compustat data was available	Fully-match 163 firm-year observations: match by year, industry (3 digit SIC), and Compustat delisting code, if any, and then select firm of closest size (total assets).	Unmatched OLS (and unmatched logit not reported) regressions of litigation or not	Yes	No	Yes	Yes	Yes	No
Krishnan, J. and J. Krishnan (1997). <i>The Accounting Review</i> , "Litigation Risk and Auditor Resignations"	21	CB-SM	Identify 141 firms whose auditors resigned during 1989-1995 and all required data available.	Semi-match firm-year observations of firms who dismissed auditors (auditors did not resign), following Stice (1991): one sample of 141 matched on year only and then random selection, a second set of 141 is matched on year and industry.	Unmatched logit regressions of resignation versus dismissal	Yes	No	No	No	No	No
Shu (2000). <i>Journal of Accounting and Economics</i> . "Auditor Resignations: Clientele Effects and Legal Liability"	15	CB-SM	282 audit litigations	For each of 282 auditor litigation firms randomly select ten firms matching by year. Paper uses other control groups corresponding to auditor resignations and client-initiated auditor changes, as well.	Unmatched logit regression of audit litigation or not, and unmatched logit regression of auditor resignations explained by litigation risk and clientele effects.	Yes	No	No	No	No	No
Heninger (2001). <i>The Accounting Review</i> , "The Association Between Auditor Litigation and Abnormal Accruals".	12	CB-SM	Identify 67 auditee firms with lawsuits against auditors having 8 years of Compustat data.	Semi-match 67 firm-year observations: match on year, industry (SIC4: 62; SIC3: 5), and then randomly select one.	Unmatched logit regression of litigation or not	Yes	No	No	No	No	No

Table 3
Non-random Research Design Papers in Audit Fraud Stream
1980-2003

Paper Author(s), Journal, Title	SSCI citation count in 2007	Research Design Category	Choice-Based and/or Matched Sample Selection		Principal Analyses Explaining Choice and/or Using Matching	Analyses Incorp. Matching/Stratification?		Additional control for "closest" matching		Reweight-ing for generalizing	
			Identification of Cases, or Stratified Sample	Selection of Controls or Comparison Sample, If Any		Needed ?	Applied ?	Needed ?	Applied ?	Needed ?	Applied ?
Green and Calderon (1995). <i>Accounting Enquiries</i> , "Analytical Procedures and Auditors' Capacity to Detect Management Fraud"	Not in SSCI	NCB-FM-B	86 firms having 10-K statements subsequently found to be fraudulent, and data available.	Fully-match 86 nonfraud firm-year observations from Compustat retrospectively: match by year, size ("comparable" assets), and industry.	Univariate Wilcoxon signed rank tests of differences	Yes	Yes	Yes	No	Yes	No
Beasley (1996). <i>The Accounting Review</i> , "An Empirical Analysis of the Relation Between the Board of Director Composition and Financial Statement Fraud"	63	CB-SM	Identify 75 public firms having an occurrence of financial statement fraud publicly reported during 1980-1991.	Semi-match 75 firm-year observations having complete data: Match on year, stock exchange, firm size within +/- 30% in market value, industry (SIC4 or 3 or 2)	Unmatched logit regression explaining fraud	Yes	No	No	No	No	No
Green and Choi (1997). <i>Auditing: A Journal of Practice & Theory</i> , "Assessing the Risk of Management Fraud Through Neural Network Technology"	9	CB-FM	86 firms having 10-K statements subsequently found to be fraudulent, and data available.	Fully-match 86 nonfraud firm-year observations from Compustat: match by year, size, and industry (4 digit SIC).	Neural network explaining Fraud or non-fraud, without use of matching	Yes	No	Yes	No	Yes	No
Abbott, Park, Parker (2000). <i>Managerial Finance</i> . "The Effects of Audit Committee Activity and Independence on Corporate Fraud"	Not in SSCI	CB-FM	Identify 78 firms subject to SEC Accounting and Auditing Enforcement Releases.	Fully-match to 78 firm-year observations in non-sanctioned firms matched by size, industry, trading exchange, and time period.	Appears to be unmatched linear probability model (OLS regression) explaining 1-0 sanction or not. May instead be an unmatched logit regression.	Yes	No	Yes	No	Yes	No
Bell and Carcello (2000). <i>Auditing: A Journal of Practice & Theory</i> , "A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting"	6	CB-SM	Employ sample of 77 audit client engagements of an audit firm where material financial statement fraud was identified during 1960-1989, as identified in a prior study.	For comparison, solicit survey responses from a stratified sample of non-fraud audit engagements of the same firm during 1990, with stratification ensuring industry representation was proportional to the audit firm's client base (but we presume in differe	Unmatched logit regressions explaining fraud or not.	Yes	No	No	No	No	No