Stakeholder Perceptions of Data and Analytics based Auditing Techniques

Brian Ballou
EY Professor
Department of Accountancy
Miami University

Jonathan H. Grenier
Associate Professor
Department of Accountancy
Miami University

Andrew Reffett
Professor
Department of Accountancy
Miami University

August 2019

We appreciate the generous financial support of KPMG and Miami University and constructive feedback from Christine Earley, Kathryn Holmstrom, Bright Hong, Jennifer Joe, Kathryn Kadous, Curtis Mullis, Jon Pyzoha, and workshop participants at KPMG LLP’s US Headquarters, Georgia State University and the University of Kentucky.
Stakeholder Perceptions of Data and Analytics based Auditing Techniques

ABSTRACT

Public accounting firms have invested significant resources to develop reliable substantive tests using large data sets and sophisticated algorithms (“data and analytics based procedures”). Developing reliable procedures, however, is just one hurdle firms must clear before leveraging such data sets and algorithms. In particular, firms will also need to convince audit stakeholders that relying on data and analytics based procedures will not only enhance audit efficiency, but also improve, or at least maintain, audit effectiveness. This study provides exploratory, experimental evidence to indicate how three key audit stakeholder groups—non-professional investors, peer reviewers, and jurors—perceive two prominent data and analytics based audit procedures (population testing and predictive modeling) relative to a more traditional substantive procedure (sample testing). Results suggest that key audit stakeholders are generally open to, and, in some cases, favorably disposed to the use of data and analytics based audit procedures. Nevertheless, participants also expressed some concerns about the appropriateness of relying on data and analytics based procedures, particularly predictive modeling, as primary sources of substantive evidence. This paper reports these and other key findings to develop an agenda for future research to help firms better understand and ultimately address stakeholder concerns.

KEYWORDS: Data and Analytics; Substantive audit procedures; Analytical Procedures; Audit Sampling; Audit Quality; Auditor Liability
I. INTRODUCTION

Public accounting firms have invested, and almost certainly will continue to invest, significant resources to develop reliable technologies to harness recent advances in data availability and sophisticated analytical algorithms (see Richins, Stapleton, Stratopoulus, and Wong (2017) for a discussion of by-firm investments). A primary objective in developing such technologies is to create and implement new data and analytics based substantive audit procedures to supplement and, in some cases, replace more traditional audit procedures. Developing reliable substantive data and analytics based procedures, however, is not the only hurdle firms must clear. Specifically, firms will also need to ensure that audit stakeholders are confident that auditor reliance on substantive data and analytics based procedures will enhance (or at least maintain) audit effectiveness, regardless of any positive effects on audit efficiency.

Currently, however, little is known about how key audit stakeholder groups perceive data and analytics based audit procedures. Thus, the purpose of this study is to provide initial, exploratory experimental evidence to indicate how three important audit stakeholder groups perceive two prominent data and analytics based audit procedures relative to a more traditional audit procedure.\(^1\) Specifically, we examine investor, juror, and peer reviewer perceptions of the following three procedures: (1) traditional statistical sample based procedures (statistical sampling); (2) data and analytics based testing of complete populations for anomalies (population

---

\(^1\) While there are studies examining the effectiveness of non-substantive data and analytic auditing procedures (see Section II for a review), little is known (by the academic literature) about the relative effectiveness of data and analytic based auditing procedures (compared to more traditional procedures) used as substantive tests of financial statement balances, with the exceptions of the literatures on continuous auditing (e.g., Vasarhelyi, Alles, and Williams 2010; Zhang, Yang, and Applebaum 2015) and fraud detection (cf. Ngai et al. 2011). Since auditing firms are devoting substantial resources to the development of substantive data and analytic based auditing procedures, we assume that the firms believe that there are not only efficiency, but also effectiveness benefits. Rather than put this belief to empirical scrutiny, we examine if stakeholders share this belief. Of course, the relative effectiveness of data and analytic based auditing procedures is worthy of considerable research inquiry (see calls for research by Appelbaum, Kogan, and Vasarhelyi 2017, Brown-Liburd, Isa, and Lombardi 2015; Earley 2015; Krahel and Titera 2015; Schneider, Dai, Janvrin, Ajayi, and Raschke 2015; Vasarhelyi, Kogan, and Tuttle 2015; Yoon, Hoogduin, and Zhang 2015), but is beyond the scope of this study.
data and analytics based predictive modeling substantive analytical procedures (predictive modeling). Based on our discussion with multiple auditing firms and our review of data and analytic thought pieces (Littley 2012; AICPA 2014; Cao, Chychyla, and Stewart 2015; Appelbaum et al. 2017), population testing and predictive modeling are two primary ways that auditors can, or at least potentially can, use data and analytics to substantively test financial statement balances. We employ experiments, rather than qualitative methods (e.g., interviews) that are more typically used in exploratory research due to expectations of considerable variance in opinions and knowledge of data and analytics based auditing procedures within the participating stakeholder groups. Thus, experiments allow us to control the description of the procedures (vetted by the professional practice unit of a Big 4 firm) and document any perceptions of differential audit effectiveness across large samples of heterogeneous stakeholders.

This study is exploratory in nature, which Ledgewood et al. (2017, 45) describe as an appropriate research method that can be “enormously generative” in terms of new research questions when venturing into new scientific territory, such as understanding the intended and unintended consequences of embedding more data and analytics into the auditing profession. The auditing community describes audit firms’ shift to data-driven audit methodologies as a major change from a compliance based to a systems-based model (Crosley and Anderson 2018), which suggests that exploratory studies are of increased value. Accordingly, this study seeks to “learn

---

2 Of course, there are many other important audit stakeholders such as PCAOB inspectors and standard setters, lenders and other capital providers, educators, and board members, most notably audit committee members. Based on discussions with audit professionals and standard setters, audit standard setters are moving slow due in part about concerns about how to best ensure the reliability of data used in data and analytic auditing procedures. See Alles (2015), Earley (2015), Krahel and Titera (2015), Schneider et al. (20150, and Warren Jr., Moffitt, and Byrnes (2015) for further discussion of the effect of data and analytics on standard setting. Future research should examine the views of all of these constituencies.

3 Note that since the inception of this study and data collection, firms have continued to invest further in data and analytics oriented tools, including adding artificial intelligence to enable even more thorough population testing than described in this study (c.f., Crosley and Anderson 2018 for a practitioner description of such investments).
from suggestive patterns in the data rather than use inferential statistics for the purpose of testing

\textit{a priori} hypotheses and drawing strong conclusions” (Ledgewood et al. 2017, 45). That said, we
consider several factors that lead to intuition-based expectations, presented as research questions
as opposed to hypotheses. First, based on prior research, we tentatively expect non-professional
investors to view population testing, often referred to as \textit{complete} population testing (meaning that
the entire population has been screened to some level), as providing more assurance than either
predictive modeling or statistical sampling procedures.\footnote{Investors do not currently have access to audit procedures, but will have access to some of the auditors’ procedures in the near future, as the PCAOB (2017) will require auditors of U.S. public companies to report critical audit matters (CAM) and the associated audit procedures in their audit opinions starting for the audits of large accelerated filers in 2019 and the audits of the remaining U.S. public companies in 2020. We leverage the CAM setting in our experiment to communicate audit procedures to investor participants.} Such perceptions could increase investors’
perceptions of audit quality and ultimately their willingness to invest in companies when auditors
perform population testing procedures relative to the other two procedures.\footnote{Future investments in data and analytics and artificial intelligence likely will increase the level of assurance expected by stakeholders beyond the population-based procedures used in this study, which heighten the importance of studying its impact on external stakeholders. Whether or not these potential heightened expectations will exacerbate the audit expectations gap, as suggested in prior research, (cf. Earley 2015, Krahel and Titera 2015), is beyond the scope of this study.}

Second, similar to investors, jurors could also perceive that population testing provides
assurance beyond the levels attained by predictive modeling or sampling. However, such
perceptions might have the opposite effect on jurors’ versus investors’ judgments because
investors typically consider audit quality without knowledge of material misstatements, whereas
jurors have such knowledge. Specifically, when an audit fails to detect a material misstatement,
the perception that population testing provides a level approaching absolute assurance could make
jurors less inclined to attribute undetected misstatements to inherent limitations of audit
procedures, and more inclined to attribute undetected misstatements to either auditor incompetence
and/or malfeasance. Conversely, population testing could reduce perceptions of negligence

signaling that auditors exceeded the minimum level of audit testing (cf. Barr-Pulliam, Brown-Liburd, and Sanderson 2017). Thus, in cases of alleged audit failure, jurors’ auditor negligence judgments could be more, less or equally severe when the auditors utilized population testing versus predictive modeling or statistical sampling procedures.

Third, although we expect peer reviewers to be more attuned to the actual assurance level provided, we tentatively expect peer reviewers to prefer statistical sampling over population testing procedures, and population testing over predictive modeling procedures. We base this expectation on the intuitive possibility that peer reviewers are more familiar with statistical sampling than population testing procedures, and more familiar with population testing than predictive modeling procedures, and peer reviewers being more favorably disposed to more familiar procedures.

To examine these questions, we conducted three experiments, one designed for each stakeholder group. Each experiment provides participants with information about a fictional insurance company and then discusses the audit procedures utilized to test the company’s claims expense and associated reserve. The experiment manipulates the nature of the auditors’ procedures at three levels—statistical sampling, population testing, and predictive modeling. For investors and jurors (peer reviewers), the experiment manipulates the auditors’ procedures between-participants (within-participants). For peer reviewers and investors, the experiment also manipulates the risk of material misstatement (RMM) at two levels—high versus low to explore whether perceptions differ based on RMM. For investors, the primary dependent variables are perceptions of auditor competence, perceptions of the justifiability of the audit procedures, and willingness to invest in the audited company. For peer reviewers, the dependent variables are perceptions of the quality of, and familiarity with the audit procedures. For jurors, the dependent variables are negligence verdicts, perceptions of auditor competence, and of the justifiability of
the audit procedures. Last, it is important to note that we tailored the experimental settings to each group’s natural setting. Most notably, the experiment does not provide outcome information to investor and peer reviewer participants, but informs juror participants that the audit in question did not detect a material misstatement. This approach was taken because we are not comparing responses to data and analytics based audit procedures across stakeholders but instead investigating research questions on the impact for each stakeholder group individually so that future research can pursue the perceptions of specific stakeholder groups more effectively.

Results overall suggest that each stakeholder group is not averse to data and analytics based audit procedures. Specifically, results indicate that investor willingness to invest is unaffected by whether the auditors’ procedures are statistical sampling based, population testing or based on predictive modeling. However, for several measures, investors view predictive modeling more favorably than statistical sampling procedures. In particular, when the risk of material misstatement is relatively high, investors appear to view auditors as more competent, and their procedures as more justifiable, when they utilize predictive modeling versus statistical sampling procedures. A reasonable post hoc explanation is that investors value additional auditor effort to incorporate informative and potentially less biased external information into their expectations, especially when the risk of material misstatement is high.

Jurors view population testing as favorably as statistical sample testing, and, inconsistent with expectations, more favorably than predictive modeling procedures. In particular, jurors are less likely to hold the auditors liable for the plaintiff’s alleged damages, and view the auditors’ procedures as more justifiable, when the auditors utilized population testing versus predictive modeling procedures. This result is contrary to the notion that jurors will believe that population testing provides higher assurance (i.e., close to absolute assurance), and will attribute undetected
misstatements to auditor incompetence or malfeasance. However, jurors need to be convinced of
the sufficiency of predictive modeling as a primary substantive test.

Peer reviewer perceptions of the quality of audit procedures and their ability to detect
material misstatements did not vary across the three procedures or by the risk of material
misstatement. Further, when asked to select the most effective of the three types of procedures,
peer reviewers were approximately evenly split across the three. It is worth noting, however, that
peer reviewers are significantly more familiar with statistical sampling procedures than with either
population testing or predictive modeling, and are more familiar with population testing than
predictive modeling. However, peer reviewers’ familiarity with data and analytics based
procedures is unrelated to their perceptions of the quality of those procedures. Interestingly, we
observe no significant differences in the perceptions of the three procedures across peer reviewers
from larger versus smaller public accounting firms. Open-ended responses did reveal some peer
reviewer apprehension about the reliability of data used in predictive modeling, similar to the
views of standard setters (cf. Krahel and Titera 2015), and its sufficiency as a primary substantive
test.

This study’s findings contribute to the literature by providing exploratory experimental
evidence regarding several key stakeholder group perceptions of two prominent data and analytics
based procedures, population testing and predictive modeling. With respect to investors, our study
is a first step in answering Krahel and Titera’s (2015) call for research on how data and analytic
auditing techniques affect investor confidence by demonstrating that investors are generally open
to the techniques and even prefer them under some circumstances. Likewise, with respect to jurors,
our study is a first step in answering Earley’s (2015) and Richin et al.’s (2017) calls for research
on how data and analytic auditing techniques affects auditor legal liability. In addition, we provide
preliminary evidence that population testing may not exacerbate the expectations gap when data and analytic auditing techniques fail to detect misstatements, contrary to the expressed concerns of Cao et al. (2015), Earley (2015), Krahel and Titera (2015), Schneider et al. (2015), and Richins et al. (2017). However, we find some evidence that jurors need to be convinced of the benefits of predictive modeling perhaps due to concerns with data reliability or evidence sufficiency. Further, we do not observe that population testing reduces negligence perceptions relative to statistical sampling. This inconsistency with concurrent work (Barr-Pulliam et al. 2017) indicates that future research is needed before reaching more definitive conclusions about the effects of data and analytic techniques on auditor liability. Finally, peer reviewers also appear to be generally open to data and analytic procedures, but share the concerns of jurors regarding predictive modeling.

This study also provides an important base for future research to develop and test theoretically-based hypotheses. While there have been numerous calls for future research on the effectiveness of data analytic auditing techniques (see footnote 1 for a list of calls), our study proposes a future research agenda focused on stakeholder perceptions. By doing so, we significantly expand the specificity of previous broad calls for research on the effects of data and analytic auditing procedures on investor confidence and auditor legal liability (Earley 2015; Krahel and Titera 2015; Schneider et al. 2015; Richins et al. 2017). Additionally, our study’s findings provide helpful insights for practitioners to consider as they develop and implement data and analytic procedures. Overall, results suggest that key audit stakeholders are generally open to, and in some cases, favorably disposed to data and analytics based procedures. However, results also suggest that firms likely need to persuade certain stakeholders, most notably jurors and peer reviewers, of the appropriateness and sufficiency of predictive modeling as a substantive audit.
procedure. Thus, future research is needed to build on this study to better understand the source of stakeholder concerns about predictive modeling, and how best to alleviate such concerns.

II. MOTIVATION AND RESEARCH QUESTIONS

This section discusses relevant background information, and various factors that could affect investor, peer reviewer, and juror perceptions of population testing and predictive modeling audit procedures (both relative to each other and compared to statistical sampling procedures). After discussing such factors, we present research questions.

Data and analytic based auditing procedures

Previous studies on the effectiveness of data and analytic auditing techniques primarily fall into three categories: financial distress modeling, financial fraud modeling, and stock market prediction / quantitative modeling (see Gepp, Linnenluecke, O’Neill, and Smith (2018) for a review of studies in each of these categories as well as a review of non-accounting studies of data and analytic techniques with applicability to accounting and auditing). Ngai, Hu, Wong, Chen, and Sun (2011) provide a more comprehensive review of the financial fraud modeling literature while West and Bhattacharya (2016) review computational intelligence-based approaches to fraud detection. Although many of these previous studies have implications for the substantive testing of financial statement balances (described below), we are unaware of any studies that examine the relative effectiveness of data and analytic auditing procedures as substantive tests of the material accuracy of financial statement balances. Instead of directly addressing this gap in the literature,
we examine whether key audit stakeholders believe that two substantive procedures, population testing and predictive modeling, have effectiveness benefits.

Population testing is using high-powered analytics to analyze complete populations of client data to identify anomalies for further investigation. An anomaly is defined as a transaction with certain attributes that suggest a higher risk of material misstatement. Many papers laud the effectiveness benefits of population testing over other sampling approaches as it increases the chances of identifying material misstatements as long as the attributes used to identify anomalies are truly associated with elevated misstatement risk (Cao et al. 2015; Earley 2015; Schneider et al. 2015; Yoon et al. 2015; Appelbaum et al. 2017). As for empirical evidence of the effectiveness of population testing, we are not aware of any studies that compare its effectiveness relative to traditional auditing approaches. However, the studies showing the power of population testing approaches to detect control weaknesses and/or occupational fraud provide indirect empirical support (e.g., Dutta and Tavawala 2013; Jans, Alles, and Vasarhelyi 2014).7

Predictive modeling can be used as a substantive test by developing a formal expectation of the client’s financial statement balance. Auditors have long used substantive analytical procedures, but the rise of big data and processing capabilities is allowing them to build more sophisticated models based on larger and larger data sets, including those that do not contain data from potentially biased client internal systems (Cao et al. 2015; Earley 2015; Appelbaum et al. 2017). Such predictive models may be especially beneficial for the auditing of balances with significant uncertainty such as estimates (Krahel and Titera 2015). The studies on stock market

---

7 Another stream of studies build data analytic techniques to successfully predict instances of fraud at the company level (Huang, Tsaih, and Lin 2014; Ravisankar et al. 2011; Perols, Bowen, Zimmermann, and Samba 2017). See Gepp et al. (2018) for a review.
prediction and quantitative modeling illustrate numerous variables that could be incorporated into predictive models (see Gepp et al. (2018) for a review).

**Non-professional investors**

While there is little to no theory or evidence to suggest how non-professional investors perceive data and analytics based audit procedures relative to more traditional procedures, existing research finds various factors can affect investor perceptions of audit quality and that such perceptions affect investment decisions. Prior research finds that earnings response coefficients (ERCs), a measure of the association between unexpected earnings and abnormal changes in stock price (and thus a measure of the degree to which stock prices reflect new earnings information), are positively related to the credibility of companies’ financial statements (Teoh and Wong 1993). Thus, as perceived audit quality increases, investors view companies’ earnings as more credible, which increases ERCs. Leveraging the hypothesized positive association between audit quality and ERCs, prior research identifies several factors that affect investor perceptions of audit quality.8

As noted above, none of the discussed studies suggest how investors likely perceive new data and analytics based procedures. Thus, examining investor perceptions of the quality of emerging data and analytics based procedures has important implications for capital markets and thus is an important area of inquiry for auditors, regulators, and academics. When considering the nature of the three examined audit procedures, one possibility is that non-auditor stakeholders, such as non-professional investors and jurors, will perceive that population testing provides higher than reasonable assurance, perhaps near absolute assurance, especially when it is referred with

---

8 Specifically, Teoh and Wong (1993) report statistically greater ERCs (and thus higher audit quality/more credible financial statements) among companies that utilize Big Eight audit firms versus companies that utilize non-big Eight audit firms. Within Big N firms, Balsam, Krishnan, and Yang (2003) report greater ERCs among companies that utilize industry specialist as opposed to non-specialist auditors. Last, Krishnan, Sami, and Yang (2005) find that ERCs are negatively associated with non-audit fees, suggesting that investors believe non-audit fees impair auditor independence (in appearance); and that such factors influence their investment decisions.
terms such as “complete population testing.” As Krahel and Titera (2015, 418) state, “This difference will need to be articulated, lest the public come to expect a ‘perfect’ audit.” Such misperceptions could lead to overly favorable views of the effectiveness of population testing relative to other procedures, and ultimately a greater willingness to invest.⁹ Finally, we explore whether investor perceptions of assurance quality and willingness to invest depend on the risk of material misstatement (RMM) as they may view certain procedures as more or less valuable when RMM is elevated. These beliefs lead to the following research questions:

RQ1a: Do non-professional investors believe that population testing procedures provide greater assurance than statistical sampling and/or predictive modeling, and do their beliefs about these procedures depend on the risk of material misstatement?

RQ1b: Is there an association between investor perceptions of assurance quality and their willingness to invest?

RQ1c: Is investor willingness to invest in a company affected by whether the company’s auditors used statistical sampling, population testing, or predictive modeling procedures, and if so, is it contingent on the risk of material misstatement?

Jurors

As noted above, we expect jurors to view population testing as also providing higher than reasonable assurance, perhaps near absolute assurance, and that in the presence of outcome information, this inflated expectation of population testing could lead to less favorable judgments of auditors. In other words, jurors may hold auditors who utilize population based testing to high standard of care (Earley 2015). Gray and Debreceny (2014, 378) illustrate, “Data mining can be considered the equivalent to taking a 100% sample. If the smoking gun is in that sample, but the auditors missed it, then the auditors no longer have their traditional industry-practice defense (of

⁹ Note that such perceptions could be exacerbated should firms implement even more sophisticated population testing, such as using artificial intelligence to more thoroughly screen populations.
sampling risk).” Thus, we believe that the misperception that population testing provides close to absolute assurance could lead to more negative judgments in situations in which evaluators know that an audit failed to detect a misstatement. Specifically, overly optimistic expectations about population testing could cause jurors to attribute undetected misstatements to either auditor incompetence or auditor complicity/malfeasance, instead of limitations of the audit procedure. On the other hand, concurrent work suggests that use of population testing could actually reduce perceptions of negligence by signaling that the auditors exceeded the minimum level of auditing procedures (Barr-Pulliam et al. 2017). This leads to the following research questions:

RQ2a: Do juror perceptions of auditor competence vary when the auditors performed population testing, statistical sampling, or predictive modeling procedures?

RQ2b: Is there an association between juror perceptions (under any of the three procedures) of auditor competence and their negligence verdicts?

RQ2c: Are jurors’ auditor negligence verdicts affected by whether the auditors used statistical sampling, population testing, or predictive modeling procedures?

Peer reviewers

As noted, there is little to no theory or evidence to indicate how peer reviewers perceive data and analytics based procedures. Unlike investors and jurors, we do not expect peer reviewers to view population testing as providing more than reasonable assurance. Rather we base our expectations for peer reviewers on potential differences in their familiarity with the different procedures. Specifically, we expect peer reviewers to be more familiar with statistical sampling than population testing procedures, and more familiar with population testing than predictive modeling procedures. Further, we expect to observe a positive association between peer reviewers’ familiarity with data and analytics based audit procedures and their perceptions of the quality of those procedures. In other words, peer reviewers might not be as comfortable with procedures that
are less familiar to them, or perhaps hesitant to effectively certify them of high quality due to unfamiliarity. Finally, we explore whether peer reviewer perceptions of quality depend on the risk of material misstatement (RMM) as they may consider certain procedures as more or less appropriate when RMM is elevated. Collectively, this reasoning leads to the following research questions:

RQ3a: Are peer reviewers more familiar with statistical sampling, population testing, or predictive modeling audit procedures?

RQ3b: Is there an association between peer reviewer familiarity with data and analytics based procedures (population testing or predictive modeling) and their perceptions of the quality of those procedures?

RQ3c: Do peer reviewers view data and analytics based procedures (population testing or predictive modeling) more or less favorably than traditional statistical sampling, and do their perceptions depend on the risk of material misstatement?

III. METHOD

This section describes the experimental method for all three reported experiments. We begin with the experiment used to examine non-professional investors’ perceptions, and then proceed to the experiments used to examine peer reviewers’ and jurors’ perceptions respectively. When describing the experiments used to examine peer reviewers’ and jurors’ perceptions, we only discuss how those experiments differ from the experiment utilized for investors.

Non-professional investors

Participants

We recruited investor participants from Amazon Mechanical Turk, an online market place that offers individuals (i.e. workers) an opportunity to complete human intelligence tasks (HITs) for compensation. To ensure participants are appropriate proxies for non-professional investors, the experiment first asks prospective workers if they have bought or sold financial instruments,
and if they have ever read a company’s financial statements. Participants who answer yes to both questions are allowed to continue, and are paid $2 for completing the task. Mechanical Turk data are considered highly reliable (Paolacci et al. 2010; Buhrmester et al. 2011; Horton et al. 2011; Farrell et al. 2017). Accounting studies have also used Mechanical Turk to proxy for non-professional investors (e.g., Rennekamp 2012; Koonce, Miller, and Winchel 2015; Chen, Han, and Tan 2016; Asay, Elliott, and Rennekamp 2017).

Our sample includes 232 participants—61 percent are male and, on average, 35 years of age. Participants on average indicated that they have a moderate level of investing experience—participants’ mean self-reported investing experience level is 44 on a scale between 0 and 100, with 0 indicating no experience, and 100 indicating extensive experience.

Task

After reading and agreeing to an informed consent form, participants read background information about financial statements and the audit process. Then, participants assume the role of an investor who is considering investing in an insurance company (Big Country Insurance (BCI)), and are provided background information about the company, including a comparative balance sheet and income statement for BCI. Participants are then provided a summary of the audit procedures performed by BCI’s audit firm (INTL) during the year-end audit relating to the company’s claims expense. As audit reports typically do not disclose auditors’ procedures to investors, the experiment discusses how the auditors disclosed the procedures as a critical audit matter in the audit opinion. Finally, participants answer several questions regarding their

10 Human subjects approval was obtained from the authors’ university for all three experiments.  
11 On June 1st, 2017, the PCAOB (2017) passed a new auditing standard requiring U.S. public companies to disclose critical audit matters including discussion of the associated audit procedures. The effective date for large accelerated filers will be fiscal years ending on or after June 30th, 2019, and fiscal years ending on or after December 15th, 2020 for all other companies.
perceptions of the quality of the auditors’ procedures and their willingness to invest in the insurance company.

**Independent variables**

The experiment manipulates two variables on a between-participants basis, the auditors’ procedures and RMM. The experiment manipulates the auditors’ procedures at three levels—statistical sampling, population testing, and predictive modeling (see Exhibit 1 Panel A for the wording of the manipulations). The wording of these manipulations were vetted by the professional practice unit of a Big 4 firm for external validity. The experiment manipulates RMM at two levels, high and low. In the high RMM condition, the instrument indicates that INTL considers claims expense to be a high risk area because historical predictions of BCI’s wildfire claims (a major component of its claims expense) have been relatively imprecise. The instrument further indicates that INTL disclosed a critical audit matter (CAM) relating to claims expense. In the low RMM condition, the instrument indicates that INTL considered claims expense to be a low risk area as historical predictions have been relatively precise, but that given its significance to the financial statements, chose to disclose a CAM relating to claims expense. As noted above, we included CAMs in both conditions because there is no other realistic mechanism to inform investors of the nature of the auditors’ procedures relating to specific accounts.

**Dependent variables**

The primary dependent variables are participants’ perceptions of auditor competence, perceptions of the justifiability of the auditors’ procedures, RMM assessments, and their

---

12 In all three experiments, we refer to population testing as complete population testing as, based on anecdotal discussions with practitioners and our conversations with the professional practice unit, this was how the audit procedure was commonly referred to at the time of data collection. However, more recent anecdotal evidence indicates that the word complete is no longer commonly used, perhaps to lessen the chances of exacerbating the expectations gap. Importantly, our inclusion of the word complete biases us towards finding such exacerbation, making our failure to find supporting evidence of such exacerbation even more convincing.
willingness to invest in the audited company. To measure participants’ perceptions of auditor competence, the experiment asks participants to indicate the auditor’s competence on a scale between 0 and 100, with 0 indicating “not at all competent” and 100 indicating “completely competent”. To measure participants’ perceptions of the justifiability of the auditors’ procedures, the experiment asks participants to indicate the justifiability/defensibility of the auditors’ procedures on a scale between 0 and 100, with 0 indicating “very low”, and 100 indicating “very high”. To measure participants’ willingness to invest, the experiment asks participants how likely they would be to invest in BCI on a scale between 0 and 100, with 0 indicating “extremely unlikely”, and 100 “extremely likely”.

**Jurors**

**Participants**

We recruited juror participants from Amazon Mechanical Turk. To ensure participants were appropriate proxies for jurors, prospective workers indicated whether they are U.S. citizens and at least 18 years of age. Only participants who answer yes to both questions continue, and are paid $2 for completing the task. Our sample includes 201 observations of which 48 percent of the participants are male and, on average, are 36 years old. Mechanical Turk workers who participated in the non-professional investor study were not allowed to participate in the juror study. Several prior studies have used Mechanical Turk workers to proxy for jurors (e.g., Grenier, Pomeroy, and Stern 2015; Brasel, Doxey, Grenier, and Reffett 2016; Maksymov and Nelson 2017), as their demographics are more diverse than those of student participants.

**Task**

After reading and agreeing to an informed consent form, participants read background information about financial statements and the audit process. Then, participants are asked to
assume the role of a juror in a lawsuit involving an insurance company (BCI) and an independent auditor (INTL). Participants then read background information about BCI, and a summary of the audit procedures INTL performed relating to BCI’s claim expense during the year-end audit. Then, participants read about an undetected material misstatement involving a $5 million understatement of claims expense (and corresponding $5 million overstatement of net income). Last, participants answer questions regarding their beliefs about the quality of the auditors’ procedures, and the auditors’ culpability for investors’ losses allegedly caused by the insurance company’s misstated financial statements.

*Independent variables*

The experiment manipulates the auditors’ procedures between-participants at three levels—statistical sampling, population testing, and predictive modeling. The wording of the manipulations is the same as the wording used for investor participants (see Exhibit 1 Panel A).

*Dependent variables*

To measure participants’ verdicts, the instrument asks participants to indicate whether they would find INTL negligent, yes or no. To measure participants’ perceptions of the justifiability of the auditors’ procedures, participants indicate the justifiability/defensibility of the auditors’ procedures on a scale between 0 and 100, with 0 indicating “very low”, and 100 indicating “very high” quality. To measure participants’ perceptions of auditor competence, participants indicate the auditor’s competence on a scale between 0 and 100, with 0 indicating “not at all competent” and 100 indicating “completely competent”.

*Peer reviewers*

*Participants*
We recruited peer reviewer participants by sending e-mails to 1,857 peer reviewers (after excluding 169 undeliverable emails), using a list of peer reviewers published on the AICPA website. Our final sample includes 88 peer reviewers, resulting in a response rate of 5 percent.\(^1\) On average, peer reviewer participants have approximately 27 years of work experience in public accounting and 13 years of experience serving as peer reviewers. Peer reviewer participants received a $20 Amazon gift card for their participation.

**Task**

After reading and agreeing to an informed consent form, participants are asked to assume the role of a peer reviewer who currently is reviewing the audit of Big Country Insurance (BCI) as part of the peer review for INTL public accounting firm. The instrument describes the audit procedures relating to BCI’s claims expense performed by INTL during the year-end audit. The instrument also describes an audit adjustment made based on those procedures. After reading this information, participants answer several questions regarding their perceptions of the quality of the auditors’ procedures, and indicate their level of familiarity with those procedures.

**Independent variables**

The experiment manipulates the auditors’ procedures within-participants at three levels—statistical sampling, population testing, and predictive modeling (see Exhibit 1, Panel B). These descriptions are shorter than the ones used for investors and jurors as (1) it was not necessary to explain the procedures in as much detail to peer reviewers due to their auditing knowledge, and (2) we were extremely cognizant of adding unnecessary length to the experiment for professional participants. The experiment randomizes the presentation order of the procedures to avoid order

---

\(^1\) This response rate is similar to other accounting studies administered via email to professionals (e.g., Dichev, Graham, Harvey and Rajgopal (2013) and Brown, Call, Clement and Sharp (2014) whose response rates were 5.4 percent and 10.9 percent, respectively).
effects. The experiment manipulates RMM between-participants at two levels, high and low. In the high RMM condition, the instrument indicates that INTL considers claims expense to be a high risk area because historical predictions of BCI’s wildfire claims (a major component of their claims expense) have been relatively imprecise. In the low RMM condition, the instrument indicates that INTL considered claims expense to be a low risk area as historical predictions of BCI’s wildfire claims have been relatively precise.

**Dependent variables**

The experiment measures participants’ perceptions of the quality of the auditors’ procedures by asking participants to indicate the quality of INTL’s procedures on a scale between 0 and 100, with 0 indicating “very low” and 100 indicating “very high” quality. In addition, participants indicate their level of familiarity with each of the three procedures, and which of the three they believe is most appropriate in the given situation. They also justify their choice of most appropriate procedure in an open-ended response.

**IV. RESULTS**

**Non-professional investors**

**Manipulation check**

For the audit procedure manipulation, 90.8 percent of investor participants correctly identified the audit procedures performed by the auditors in their particular experimental case. For RMM, 91.4 percent of investor participants in the low (high) RMM condition correctly indicated that the auditors in the case assessed claims expense to be a low (high) risk audit area. These high percentages indicate that participants generally attended to, and understood the experimental manipulations. Excluding participants who failed at least one manipulation check does not change the inferences drawn from the results.
Research questions

RQ1a examines whether investors’ perceptions of the risk of material misstatement (after audit procedures are performed) are affected by whether the auditors’ primary source of audit evidence came from statistical sampling, population testing or predictive modeling procedures. To examine this question, we performed an analysis of variance (ANOVA) with the auditors’ procedures and the risk level as independent variables, and participants’ assessments of the probability that the audited financial statements are materially misstated as the dependent variable. Results indicate that there is not a significant effect of the auditors’ procedures on investors’ perceptions of RMM (F = 1.16; p = 0.316 two-tailed). Further, consistent with an effective risk manipulation, there is a significant effect of the study’s risk manipulation on perceptions of RMM (F = 5.29; p = 0.022 two-tailed). Last, we do not observe significant interactive effects of the auditors’ procedures and risk level on perceptions of RMM (F = 0.63; p = 0.533 two-tailed). See TABLE 1 for descriptive statistics by condition and ANOVA results.

RQ1b examines the association between non-professional investors’ perceptions of RMM and their willingness to invest in the given company. Consistent with intuition, we observe a significant negative association between investors’ perceptions of RMM and their willingness to invest in the given company (r = -0.296, n = 232, p < 0.001, two-tailed). This finding indicates that investor participants understood the experimental materials and responded thoughtfully. In addition, paired with the results of RQ1a, it illustrates how auditors communicate misstatement risk in CAMs can affect investors’ willingness to invest in a company.

RQ1c examines whether the nature of the auditors’ procedures affects non-professional investors’ willingness to invest in a company. Consistent with the results for RQ1a, ANOVA results indicate that the auditors’ procedures generally did not affect investors’ willingness to
invest (F = 0.61; p = 0.542 two-tailed). Further, consistent with an effective risk manipulation, there is a negative effect of increased risk on investors’ willingness to invest (F = 23.97; p < 0.001 two-tailed). Last, we do not observe significant interactive effects of the auditors’ procedures and the risk level on investors’ willingness to invest (F = 1.75; p = 0.176 two-tailed).

Supplemental analysis

It is important to note that, while the nature of the auditors’ procedures did not affect investor participants’ willingness to invest, our investor participants viewed predictive modeling more favorably than statistical sampling procedures in several other regards. Specifically, when RMM is relatively high, investors view auditors as more competent when they utilize predictive modeling versus statistical sampling procedures (t_{78} = 1.96, p = 0.054, two-tailed). Similarly, when RMM is relatively high, investors view the auditors’ procedures as more justifiable when they utilize predictive modeling versus statistical sampling audit procedures (t_{78} = 1.90, p = 0.062, two-tailed). A reasonable post hoc explanation is that investors value auditors making the effort to incorporate informative and potentially less biased external information into their expectations, especially when the risk of material misstatement is high. Overall, results suggest that non-professional investors’ perceptions of new data and analytics based audit procedures generally are as positive, and in some regards, more positive than their perceptions of more traditional statistical sampling procedures.

Jurors

Manipulation check

For the audit procedure manipulation, 95.0 percent of juror participants correctly identified the audit procedures utilized to test BCI’s claims expense. This high percentage indicates that participants generally attended to, and understood the experimental manipulation. Excluding
participants who failed the manipulation check does not change the inferences drawn from the results.

Research questions

RQ2a examines whether the nature of the auditors’ procedures affect jurors’ assessments of auditor competence. To examine this question, we performed an ANOVA with the auditors’ procedures as the independent variable, and participants’ assessments of auditor competence as the dependent variable. Results indicate that there is not a significant effect of auditors’ procedures on jurors’ assessments of auditor competence ($F = 0.73; p = 0.482$ two-tailed). Further, all possible pairwise comparisons across audit procedures indicate statistically insignificant differences in jurors’ assessments of auditor competence. See TABLE 2 for descriptive statistics by condition and ANOVA results.

RQ2b examines whether there is a negative relationship between jurors’ assessments of auditor competence and their negligence verdicts. Consistent with prior research (e.g., Reffett 2010), we observe a negative correlation between jurors’ assessments of auditor competence and their auditor negligence verdicts ($r_s = -0.594$, $p < 0.001$, two-tailed). This finding provides further evidence that juror participants understood the materials and responded thoughtfully.

RQ2c examines whether the nature of the auditors’ procedures affect jurors’ auditor negligence verdicts. Approximately 43 percent (31 percent) [46 percent] of participants in the statistical sampling (population testing) [predictive modeling] condition indicated that the auditors were negligent. We utilized Fisher’s exact test to perform pairwise comparisons. Results indicate that there is not a significant difference between the proportion of negligence verdicts in the statistical sampling versus the population testing condition ($p = 0.153$ two-tailed). Similarly, there is not a significant difference between the proportion of negligence verdicts in the statistical
sampling versus the predictive modeling condition (p = 0.733 two-tailed). There is, however, a marginally significant difference between the proportion of negligence verdicts in the population testing versus the predictive modeling condition (p = 0.077 two-tailed).

**Supplemental analysis**

To further understand the observed difference in the proportion of participants who indicated that the auditors were negligent in the population testing versus the predictive modeling condition, we compared participants’ perceptions of several other factors across those two conditions (not tabulated). We observe significant differences in participants’ perceptions of the defensibility of the auditors’ procedures, and in their emotional reactions to the auditors. In particular, participants in the population testing condition viewed the auditors’ procedures as more defensible (M = 70.7) than participants in the predictive modeling condition (M = 60.1) (t_{132} = 2.25, p = 0.026, two-tailed). Further, participants in the population testing condition had less negative feelings to the auditors (M = 50.6) than participants in the predictive modeling condition (M = 43.2) (t_{131} = 2.20, p = 0.029, two-tailed). Further, there is a significant negative correlation between jurors’ assessments of the defensibility of the auditors’ procedures and their auditor negligence verdicts (r_s = -0.672, p < 0.001). Similarly, there is a significant negative correlation between jurors’ emotional reactions to the auditors and their auditor negligence verdicts (r_s = -0.440, p < 0.001 two-tailed).

**Peer reviewers**

**Manipulation check**

For RMM, 89.8 percent of peer reviewer participants correctly indicated that the auditors in the case assessed claims expense to be a low (high) risk audit area. This high percentage indicates that participants generally attended to, and understood the experimental risk
manipulation. Excluding participants who failed the manipulation check does not change the inferences drawn from the results. As peer reviewers evaluated all three procedures, there is no manipulation check for audit procedures.

Research questions

RQ3a examines whether peer reviewers are more or less familiar with statistical sampling, population testing or predictive modeling procedures. To compare peer reviewers’ relative familiarity with the different procedures, we performed a two-way ANOVA using a block design to control for rater effects.\textsuperscript{14} Consistent with expectations, peer reviewers’ familiarity varied across procedures ($F = 107.8$, $p < 0.001$). Post-hoc comparisons with a Bonferroni adjustment indicate that peer reviewers are more familiar with statistical sampling ($M = 82.7$) than population testing procedures ($M = 58.8$) ($p < 0.001$), and are more familiar with population testing ($M = 58.8$) than predictive modeling procedures ($M = 35.9$) ($p < 0.001$).

RQ3b examines whether peer reviewers’ familiarity with audit procedures affects their perceptions of the quality of those procedures. Surprisingly, the only significant positive correlation we observe between familiarity and perceptions of quality is for statistical sampling ($p = 0.01$). The correlations between familiarity with procedures and perceptions of quality are insignificant for both population testing ($p = 0.331$), and predictive modeling ($p = 0.561$).

RQ3c examines whether peer reviewers view data and analytics based procedures (population testing or predictive modeling) more or less favorably than traditional statistical sampling. Results of a two-way ANOVA using a block design to control for rater effects reveals no significant effect of audit procedures on peer reviewers’ quality ratings ($F = 1.10$, $p = 0.336$).

\textsuperscript{14} With respect to our RMM manipulation, we do not observe a significant main effect nor significant interaction with procedures. Although we are hesitant to conclude that peer reviewer perceptions of the quality of audit procedures do not depend on RMM, we do not find any evidence to that effect and leave its investigation to future research. For expositional simplicity, RMM is not further discussed with respect to peer reviewer perceptions.
Post-hoc comparisons with a Bonferroni adjustment indicate no significant differences in peer reviewers’ quality ratings across statistical sampling (M = 70.74), population testing (M = 70.91), and predictive modeling procedures (M = 67.47) (all p-values > 0.50 two-tailed).

In addition, to examining peer reviewer participants’ assessments of the quality of the auditors’ procedures, we also examined participants’ responses regarding which of the procedures is the most appropriate and why. Consistent with the above results, peer reviewers were split approximately evenly across the three procedures—slightly over 1/3rd selected statistical sampling, slightly over 1/3rd selected population testing, and slightly under 1/3rd selected predictive modeling. Participants who selected statistical sampling commonly indicated that it is cost effective and reliable. Participants who selected predictive modeling commonly indicated that it utilizes third party data, which is more independent and provides a better test for completeness than utilizing client data. Participants who selected population testing commonly indicated that it provides a more thorough test, which is important in a high-risk area such as claims expense for insurers (see Exhibit 2 for a list of responses).

Supplemental analysis

We also examined whether peer reviewers from larger public accounting firms are more familiar with data and analytics based procedures than peer reviewers from smaller firms. Consistent with expectations, we observe a significant positive correlation between firm size and peer reviewers’ familiarity with population testing procedures (r = 0.257, n = 69, p = 0.017, two-tailed). However, inconsistent with expectations, we do not observe a significant correlation between firm size and peer reviewers’ familiarity with predictive modeling (r = -0.092, p = 0.407, two-tailed). Further, we do not observe significant correlations between firm size and peer reviewers’ perceptions of the quality of auditors’ population testing procedures (r = 0.102, p =
0.348, two-tailed) or predictive modeling procedures ($r = 0.098$, $p = 0.374$, two-tailed) (supplemental analyses are not tabulated).

Peer reviewer discussion of the relative strengths and weaknesses of the three procedures (Exhibit 2 Panel B) reveals some apprehension with predictive modeling. In fact, approximately one third of peer reviewers who provided an open-ended response (11 out of 32; see underlined passages in Exhibit 2 Panel B) discussed concerns with the reliability of data used in predictive modeling and/or its sufficiency as a primary substantive test. These peer reviewers appear to share the beliefs of some standard setters with respect to data reliability (cf. Krahel and Titera 2015) and view predictive modeling as complementary, as opposed to primary, audit evidence (cf. Yoon et al. 2015). In sum, peer reviewers appear generally open to data and analytic auditing techniques, but express some apprehensions about predictive modeling as a primary substantive test.

V. IMPLICATIONS FOR FUTURE RESEARCH

Public accounting firms are investing significant resources to develop reliable technologies to harness recent advances in data availability and data analysis methods in an effort to create and implement new efficient and reliable data and analytics based audit procedures. However, firms need to ensure that key audit stakeholders are confident that relying on such procedures not only will improve audit efficiency but will also enhance (or at least not impair) audit effectiveness. Prior to this study there was little to no evidence to indicate how key audit stakeholders perceive data and analytic based audit procedures. To addresses this gap, this study provides exploratory experimental evidence to indicate how three key audit stakeholder groups (non-professional investors, peer reviewers, and jurors) perceive two prominent data and analytics based audit procedures (predictive modeling and population testing) relative to a more traditional audit procedure (sample testing).
In general, results suggest that key audit stakeholders are open to, and in some regards favorably disposed to new data and analytics based audit procedures. However, while the examined stakeholders generally are comfortable with population testing, they appear less comfortable with predictive modeling. As such, firms will need to convince stakeholder groups, notably peer reviewers and jurors, of the effectiveness of predictive modeling.

We designed and employed exploratory study, using a bottom-up approach to identify patterns in data sets that can help in informing future studies by identifying research questions (Ledgewood et al. 2017). Accordingly, one of this study’s primary contributions is identifying phenomena that help future research to identify theory-based explanations researchers can test using behavioral experiments, experimental markets, or archival studies designed to isolate variables of interest and control for alternate explanations. To help researchers in this regard, we present a set of general research questions categorized by each data and analytics based procedure (predictive modeling and population testing) across three primary issues: stakeholder education, audit quality, and perceived procedure effectiveness. Given the results of our study, these questions largely (but not entirely) focus on better understanding why many jurors, and some peer reviewers are unconvinced of the effectiveness of predictive modeling, and how best to alleviate such concerns—see Exhibit 3 for the complete list of questions.

---------Insert Exhibit 3 about Here---------

This study is subject to several important limitations. First, it is exploratory in nature, so the results should be used more to help identify theory-based hypotheses for future studies. Second, the participants in this study have limitations in their ability to proxy for other stakeholders of interest. For example, while Amazon Mechanical Turk participants are effective proxies for jurors and non-professional investors, the perceptions of sophisticated investors are important
considerations and our study does not capture such perceptions. Further, peer reviewers have the limitation of being practicing auditors, who are subject to biases that independent regulators (such as PCAOB inspectors) might not possess.

In closing, there is little doubt that using big data sets, particularly those from third-parties, along with analytical algorithms can be used to improve audit quality. In fact, accounting firms likely have, or will develop effective data and analytics based audit procedures. However, auditors also need to convince stakeholders, most notably users of audited information and auditing regulators, of their reliability. Our study provides exploratory experimental evidence to help initiate the process of achieving this objective.
REFERENCES


Dutta, V., and A. Tavawala. 2013. *Applying data analytics to address fraud risk*. Proceedings of the 28th World Continuous Auditing & Reporting Symposium, Newark, NJ.


**EXHIBIT 1**

Experimental Manipulation of Audit Procedures

*Panel A: Non-professional investors and jurors*

**Statistical sampling condition**

“INTL’s primary audit procedures to test BCI’s 2015 claims expense relied on statistical sampling techniques using a stratified population in which representative samples of expenditures from the fiscal year under audit and the subsequent year are selected for testing (i.e., examining source documentation of the claim and payment) from each strata based on key attributes associated with misstatement risk. Statistical sampling, i.e., testing less than 100% of the population, is allowed under current auditing standards. INTL also had their own actuaries review the reserve for reasonableness. Results of these audit procedures indicate that BCI’s claim expense and its associated loss and loss adjustment expense reserve (i.e., claims payable liability) are materially accurate in compliance with Generally Accepted Accounting Principles. As noted above, based on the results of these, and other, audit procedures, INTL issued an unqualified opinion for BCI’s 2015 financial statements.”

**Population testing condition**

“INTL’s primary audit procedures to test BCI’s 2015 claims expense relied on complete population testing with data and analytics based procedures using client data. The firm has proprietary software that can analyze BCI’s entire population of claims and identify claim attributes that are diagnostic of potential misstatement—i.e., it identifies specific claims where attributes, e.g., claim size or location, fall outside expected ranges. Identified claims are then tested (i.e., examining source documentation of the claim and payment). Complete population testing, i.e., screening the entire population transaction for risk attributes and further testing the identified transactions, is allowed under current auditing standards. INTL also had their own actuaries review the reserve for reasonableness. Results of these audit procedures indicate that BCI’s claim expense and its associated loss and loss adjustment expense reserve (i.e., claims payable liability) are materially accurate in compliance with Generally Accepted Accounting Principles. As noted above, based on the results of these, and other, audit procedures, INTL issued an unqualified opinion for BCI’s 2015 financial statements.”

**Predictive modeling condition**

“INTL’s primary audit procedures to test BCI’s 2015 claims expense relied on proprietary software that implements predictive modeling using third party data to predict claims expenses. Based on technological advancements, historical claims information and correlations with specific claim inducing events are available for all insurance companies from 1980 to 2015, including all of BCI’s major competitors. INTL has proprietary software that uses this industry data to predict BCI’s claims expense. INTL then compares the predicted and recorded amounts for potential material differences. Predictive modeling, i.e., testing an account balance by developing an overall expectation for the account instead of testing individual transactions, is allowed under current auditing standards. INTL also had their own actuaries review the reserve for reasonableness. Results of these audit procedures indicate that BCI’s claim expense and its associated loss and loss adjustment expense reserve (i.e., claims payable liability) are materially accurate in compliance with Generally Accepted Accounting Principles. As noted above, based on the results of these, and other, audit procedures, INTL issued an unqualified opinion for BCI’s 2015 financial statements.”
Panel B: Peer reviewers

Statistical sampling condition

“INTL’s primary audit procedures to test BCI’s 2015 claims expense relied on statistical sampling techniques using a stratified population in which representative samples of claims expenditures from the fiscal year under audit and the subsequent year are selected for testing (i.e., examining source documentation of the claim and payment) from each strata based on key attributes associated with misstatement risk. INTL also had their own actuaries review the reserve for reasonableness.”

Population testing condition

“INTL’s primary audit procedures to test BCI’s 2015 claims expense relied on complete population testing with data and analytics based procedures using client data. The firm has proprietary software that can analyze BCI’s entire population of claims and identify claim attributes that are diagnostic of potential misstatement—i.e., it identifies specific claims where attributes, e.g., claim size or location, fall outside expected ranges. Identified claims are then tested (i.e., examining source documentation of the claim and payment). INTL also had their own actuaries review the reserve for reasonableness.”

Predictive modeling condition

“INTL’s primary audit procedures to test BCI’s 2015 claims expense relied on proprietary software that implements predictive modeling using third party data to predict claims expenses. Based on technological advancements, historical claims information and correlations with specific claim inducing events are available for all insurance companies from 1980 to 2015, including all of BCI’s major competitors. INTL has proprietary software that uses this industry data to predict BCI’s claims expense. INTL then compares the predicted and recorded amounts for potential material differences. INTL also had their own actuaries review the reserve for reasonableness.”
EXHIBIT 2

Peer Reviewer Comments

Panel A: Justification for choice of best procedure

1. Data analytics. Entire population is evaluated.
2. Risk is understatement; the best source of data is subsequent payments which is only selected with the statistical sampling model. The data analytics is cool but it is testing existing claims, not the risk - unbooked claims. Don't know much about the predictive modeling, not sure how effective that would be.
3. In my view, the best approach would be to use all three.
4. Any one of them is an appropriate technique. I would expect, however, that the predictive modeling with 3rd party data would not be used as the only technique, but rather as corroborative evidence to buttress either the stat sample or the complete population testing.
5. As a peer reviewer, I expect to see thought behind the approach. With IBNR there is no certainty - so I expect to see multiple approaches to assess ranges / seek comfort that no one consideration "sells" a risky outcome. We have a humbling profession - thinking there is one right answer here is exceedingly dangerous.
6. Due to climate change predictive models may understatement the numbers of losses due to fires.
7. Complete population testing - most thorough
8. Statistical sampling using stratified population based on key attributes. I will also use analytical techniques to estimate the liability for IBNR claims.
9. All can be appropriate depending on facts and circumstances
10. Third party data. Use of external data as well as for prior periods (retrospective review procedures)
11. Complete pop test - uses their rate and reimbursement structure to determine what is outside of the norm. very reasonable considering the 2015 wildfires. Also, some of it depends on the result of the testing of the actuary info, of which the results were not made known in the examples.
12. Because of their prior record of precision, I would have to say the complete population testing.
13. Complete population testing appears to give the best coverage in what should have been considered high risk.
15. Complete population testing with data analytics - Provides complete testing of population along with predictability of estimate with the specific and unique attributes of the Company and their industry insurance risks.
16. Complete population testing is most appropriate, given assumptions about its proven ability to identify the appropriate claim attributes. But it would be reasonable, depending on the risk assessment, to supplement this with some predictive modeling.
17. Predictive modeling. They are searching for something not recorded. Other companies might have had the losses not recorded on BCI books.
18. Statistical Sampling. Generally you're going to be looking for unrecorded or under-recorded claims. The statistical sampling method looked at claims made in both the current year and the subsequent year to determine the reserve. Also it looked at actual claims as opposed to relying on diagnostics.
19. All three with the most telling the predictive modeling and industry data.
20. Complete population testing. It includes the entire population of an almost abnormal set of circumstances
21. Statistical sampling as long as the sample is a valid and comprehensive sample. Although not 100% certain, a high degree of reliance can be achieved through sampling.
22. Complete population testing - I am a big proponent of data extraction software.
23. Statistical sampling - covers current and subsequent year
24. Complete population testing with data analytics and actuarial review
25. Predictive modeling as the CA wildfires was an outlier event and the modeling would use other outlier events over its data.
26. Predictive modeling - it sounds akin to a recalculation of an actuarial estimate
27. Complete population testing - I don't have experience with this but it seems like if you can test the entire population, then it would be best
28. Statistical sampling and inspection of subsequent activity because it was based on what actually occurred.
29. Either the Complete Population testing or Predictive Modeling could be appropriate. However in the population testing, they would need to do a better job of addressing IBNR.
30. Statistical sampling. An audit is to see if items are materially misstated and statistical sampling can achieve that without testing the entire population.
31. Predictive modeling as it best captures the total exposure including IBNR.
32. Predictive modeling would decipher material misstatements.
33. Predictive modeling - since such tests tend to provide a higher level of precision for analytical procedures.
34. Predictive modeling - auditors knew that past experience was low and internal measurements were imprecise
35. Statistical Sampling. If done right, it should be reliable.
36. I believe complete population testing with data analytics would be the best approach to identify material misstatements, since you're are looking at the whole populations and testing those that fall outside set limits.
37. The statistical sample. Proprietary software would need additional testing to ensure it is accurate.
38. Assuming 100% testing was not feasible, statistical sampling. Don't like to rely on client data for predictive tests.
39. Predictive modeling - risk is understatement
40. Data analytics of entire population client data to help assess level of risk against identified characteristics
41. Complete population testing since it looks like a the entire population for outliers.
42. Predictive modeling if you add in a look-back test to determine the reliability of the predictive model.
43. I think you need a combination of all three based on the high risk and less than precise ability to predict from the client.
44. Statistical sampling probably b/c I have never used data analytics or predictive modeling as an audit procedure. Small firm environment doesn't usually have access to those tools.
45. Statistical sampling is area most familiar with, so that would be my choice
46. Predictive modeling - most comprehensive test.
47. All three are appropriate depending on how much more information is available in the workpapers. Statistical Sampling is the older method and used most often in the past. However, if the firm could show that their data analytics system or the predictive modeling software and data had a good track record for evaluating client data and successfully determining amounts that were misstated or wrong, then those approaches could actually provide a better answer that the sampling approach. There would have to be some evidence of reliability of the two new systems using a look-back approach in the workpapers before I would say they are preferable to statistical sampling. However, with that evidence, the results of using those approaches would be stronger to me that sampling. It really hinges on proving that the systems are actually capable of doing what they purport to be able to do.
48. Complete population testing would be most appropriate given the level of risk and the size of misstatements identified.
49. Statistical sampling because many if not most firms would not be familiar with the testing needed to validate the models being used in the second and third alternatives.
50. Predictive modeling using third party data as you are able to rely, to a greater extent, on data gathered from independent parties.
51. Predictive Modeling Using Third Party Data - It relies on independent information that is industry based.
52. Predictive modeling with actuaries as well.
53. I still like the old school statistical sample, perhaps supplemented by predictive modeling using 3rd party data.
54. All are appropriate as long as properly utilized.
55. Complete population has the most coverage.
56. Complete testing in this situation since losses could be high and impact the financial condition of the Company.
57. Statistical sampling of claims expenditures. Substantive testing is more comforting to me than predictive modeling.
58. Predictive modeling - Provides more depth in the analysis and an improved likelihood of a satisfactory outcome.
59. Statistical sampling, complete population too expensive, time consuming, not familiar with other approach.

Panel B: Discussion of relative strengths and weaknesses of the three procedures

1. Statistical sampling fails to consider the big picture; the other two fail to consider the underlying transactions. Each used with no other substantive procedures is not adequate. And none works effectively without reliance on internal controls.
2. Cost benefit is a consideration.
3. I think that the predictive modeling with 3rd party data is probably not as strong as a standalone procedure as the other two. It would be very useful in conjunction with either the stat sample or the complete population testing.
4. The methodology used to perform data analytics should be reviewed by an outside party.
5. Statistical sampling with stratified data can be useful.
6. Looking at the entire population using data analytics should produce the most sound result.
7. Statistical sampling is becoming a thing of the past. Complete testing w/ analytics is the way I see the future going. I do not know much about predictive modeling.
8. Statistical sampling is the most acknowledged method, but the other two still provide adequate documentation for determining accuracy of unknown liabilities.
9. It would appear, in the stated scenario, that statistical sampling and predictive modeling would leave too much room for the client to manipulate the data to their advantage.
10. Data analytics and software driven approaches can provide more effective audit procedures due to the power of software in comparison to human error/techniques.
11. Statistical sampling has inherent weaknesses in that the population sampled must be relatively uniform. Insurance losses can be a result of many different factors. That is true with other industries as well. Statistical sampling also will not identify things not recorded. It only speaks to what is there.
12. You should use two methods if the amounts of errors could be material. I feel predictive modeling is based upon averages not regional data.
13. I don't like the predictive modeling technique with third party data because that data could be inaccurate.
14. Testing actual client data would appear to be more useful than using a predictive third party model.
15. The modeling method is very dependent upon the quality and suitability of the modeling software, and the practitioner’s ability to identify and properly adjust for known outlier situations. In all three cases appropriate attention will need to be given to subsequent claims activities.
16. Inefficient to test the complete population. Predictive modeling has too many variables and unknowns.
17. Predictive modeling may not match specific client circumstances.
18. Complete population testing would seem not to touch unreported claims. Statistical testing may be valid if the audit was done well beyond the end of the audit year, but this is doubtful.
19. Auditors’ judgment and their evaluation of control risk would play a large part in making the evaluations expected in this situation.
20. Statistical sampling is still subject to auditor's judgment, which could lead to incorrect selections or conclusions. Predictive modeling is the hardest to apply since you don't know how each third party accounts for certain items.
21. Testing items outside expectations seems to be the most efficient, but may not be appropriate for a high risk area.
22. With the second two, I will be relying on the developer of the analytics or predictive modeling. I would be good with all three methods if I could satisfy myself regarding the development of the analytics and the predictive modeling.
23. Prefer complete population testing when feasible due to high risk area.
24. We would use judgmental sampling, we rarely use statistical sampling unless it is required by contract or regulation, normally for a government agency. GAO has published guidance and judgmental sampling with a 90% level of confidence and 10% precision is the standard. Small firms like ours would not have access to either the third party software or data. The population testing using analytics on client data would be time consuming and probably not cost effective for firms smaller than a very large regional firm with specialist that could design the analytics or engage a consultant to perform due diligence of the model's level of effectiveness. Normally ninety percent would be the minimum.
25. I believe all three have merits but should be coupled with additional procedures and not solely relied upon to determine the potential misstatement.
26. Anything third party is best assuming the third party data is comparable to the client you are auditing. Outside source data is the best support especially in this scenario.
27. I believe it is time for us to use the modeling and data analytics approaches but the new software offerings will need to use real-world examples to prove that they are capable of performing as expected with look-back analysis to see how well they predict/analyze. Field testing on real client data is the key to getting the buy-in from practitioners. If they perform well, it should be easy to prove that they can be superior to a statistical sampling approach and should save time on the budget, get a better answer, and reduce risk.
28. The second and third methods may have validity, but would require significant testing of the data, assumptions, and model validity. Most firms are not familiar with this type of testing and thus would be inclined to utilize sampling of actual transactions.
29. I have concerns that predictive modeling does not provide enough substantive evidence to mitigate a high level of inherent risk.
30. I like the idea of predictive modeling, but not sure a smaller firm could afford the costs.
31. I would use a combination or all.
32. The predictive modeling would be distorted if the current year had a number of fires near the last quarter of the year, which claims had not be fully reported and developed by field work. The auditor needs to review development of loss data for the prior year during field work to get a handle on loss development.

Note: Responses are verbatim besides editing for typos and elimination of full or partial responses that were not relevant to the question. Underlining emphasis has been added for responses expressing apprehension with predictive modeling.
Exhibit 3


Panel A: Predictive Modeling Procedures

Stakeholder Education:
- How can firms, and the profession in general, most effectively educate stakeholders on the use of predictive modeling using big data sets?
- What is the composition of a third-party data set—what characteristics of a data set suggest analysis from that data set can provide relevant and reliable substantive evidence to test assertions?
- How does the use of third-party data account for client-specific circumstances?
- What other assumptions are included in the predictive model and are they robust to changes—i.e., does the model withstand sensitivity analyses?

Audit Quality:
- Given a sufficient understanding of the procedures, what are the specific reasons certain stakeholders (e.g., the jurors and peer reviewers in this study) believe predictive modeling provides lower quality audit evidence than traditional sampling-based tests of details of balances or population testing?
- Do reduced hours involved in predictive modeling relative to traditional procedures suggest to stakeholders an emphasis on audit efficiency over effectiveness?
- To what extent are there concerns about the security or quality of the underlying data?
- To what extent is there a preference or not with using an indirect approach to testing (predictive modeling) relative to using a direct approach to testing (traditional audit procedures and/or population based testing)?
- Are stakeholders concerned about a failure of predictive modeling to account for underlying aspects of the specific client transactions?

Perceived Procedure Effectiveness:
- Given that firms believe predictive modeling is at least as effective (and more efficient) than traditional sampling-based substantive procedures, how can firms most effectively demonstrate such effectiveness to stakeholders such as regulators, jurors, clients, and peer reviewers?
- Can outcomes from beta testing predictive modeling using internal audit clients or other test groups persuade key regulators?
- Do firms need to persuade specific stakeholder groups of effectiveness or should firms focus solely on persuading standard setters (i.e., PCAOB in U.S., IAASB internationally, etc.), who then persuade other stakeholders?
- Should predictive modeling be shown to be as or more effective than traditional sampling based substantive testing, would standard setters explicitly incorporate predictive modeling into auditing standards, and if so, would this requirement convince other stakeholders of its effectiveness?
- Can expert witnesses who extoll the benefits of predictive modeling persuade jurors?
Panel B: Population-based Substantive Testing Procedures

Stakeholder Education:
- How can firms, and the profession in general, most effectively educate stakeholders on the limitations associated with population-based substantive testing to combat the unwarranted expectations associated with a perception of “testing everything”?
- Can stakeholders understand that population testing involves identifying outlier transactions to be subject to other types of evidence to perform tests of details of balances (e.g., send confirmations, inspect inventory, inspect supporting documentation, etc.)?
- Which elements of the software do auditors use to filter account populations for identifying outlier transactions for additional substantive testing (i.e., statistical sampling techniques are based on statistical theory and judgmental sampling based on persuasive arguments)? What is basis for software filtering and will stakeholders understand it?
- Do stakeholders understand or support the cost versus benefit rationale for choosing whether to use or not to use population-based substantive testing relative to traditional sampling-based procedures?

Audit Quality:
- What specific concerns do investors, regulators, and jurors have regarding the use of population-based substantive testing procedures?
- What type of evidence will persuade stakeholders that firm software used to identify outliers within populations of accounts are sufficient, meaning that such evidence is at least as effective as traditional sampling-based substantive procedures?
- What evidence is needed to justify why firms opt to use or not to use population-based substantive testing relative to traditional sampling-based procedures?

Perceived Procedure Effectiveness:
- How does population-based substantive testing affect financial statement users’ perceptions of the level of assurance provided by an audit relative to traditional sampling-based substantive testing?
- Will stakeholders mistakenly believe that an audit opinion provides a guarantee (i.e., absolute assurance) since the auditors “tested populations” of accounts (i.e., will the lack of “sampling risk” based on statistical theory as a defense harm auditors)?
- Will stakeholders be more likely to assert auditor negligence (e.g., file shareholder lawsuits, perform SEC inquiries, etc.) when a restatement occurs for engagements in which auditors have used population based substantive testing?
## TABLE 1

Non-Professional Investors

### Panel A: Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>All</th>
<th>Low Risk</th>
<th>High Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RMM</strong></td>
<td>Mean</td>
<td>41.70</td>
<td>35.26</td>
</tr>
<tr>
<td><strong>Willingness to Invest</strong></td>
<td>Mean</td>
<td>56.62</td>
<td>66.49</td>
</tr>
<tr>
<td></td>
<td>Std Dev</td>
<td>22.69</td>
<td>19.19</td>
</tr>
</tbody>
</table>

### Panel B: Analysis of Variance – RMM

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure</td>
<td>1,532.83</td>
<td>2</td>
<td>766.42</td>
<td>1.16</td>
<td>0.316</td>
</tr>
<tr>
<td>Risk</td>
<td>3,500.70</td>
<td>1</td>
<td>3,500.70</td>
<td>5.29</td>
<td>0.022</td>
</tr>
<tr>
<td>Procedure x Risk</td>
<td>834.77</td>
<td>2</td>
<td>417.38</td>
<td>0.63</td>
<td>0.533</td>
</tr>
<tr>
<td>Error</td>
<td>149,619.30</td>
<td>226</td>
<td>662.03</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Panel C: Analysis of Variance – Willingness to Invest

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure</td>
<td>480.83</td>
<td>2</td>
<td>240.42</td>
<td>0.61</td>
<td>0.542</td>
</tr>
<tr>
<td>Risk</td>
<td>9,393.63</td>
<td>1</td>
<td>9,393.63</td>
<td>23.97</td>
<td>0.000</td>
</tr>
<tr>
<td>Procedure x Risk</td>
<td>1,370.72</td>
<td>2</td>
<td>685.36</td>
<td>1.75</td>
<td>0.176</td>
</tr>
<tr>
<td>Error</td>
<td>88,552.53</td>
<td>226</td>
<td>391.83</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:** RMM: participants’ assessment of the risk of material misstatement on a 100 pt scale; Willingness to Invest: participants’ assessment of how likely they would be to invest in BCI; Procedure: between participants manipulation of the primary audit procedure used to test claims expense and the associated reserve as either Statistical Sampling, Population Testing, or Predictive Modeling; Risk: between participants manipulation as High Risk or Low Risk based on the historical precision of claims expense and the associated reserve.
TABLE 2
Jurors

Panel A: Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Statistical Sampling</th>
<th>Population Testing</th>
<th>Predictive Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auditor Competence</td>
<td>Mean</td>
<td>59.93</td>
<td>64.22</td>
</tr>
<tr>
<td></td>
<td>Std Dev</td>
<td>23.89</td>
<td>24.70</td>
</tr>
<tr>
<td>Verdict</td>
<td>Percent Negligent</td>
<td>43.3%</td>
<td>30.8%</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>67</td>
<td>65</td>
</tr>
</tbody>
</table>

Panel B: Analysis of Variance – Auditor Competence

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure</td>
<td>907.45</td>
<td>2</td>
<td>453.72</td>
<td>0.73</td>
<td>0.482</td>
</tr>
<tr>
<td>Error</td>
<td>122,490.77</td>
<td>198</td>
<td>618.64</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel C: General Linear Model with a Logit Link - Verdict

<table>
<thead>
<tr>
<th>Source</th>
<th>χ2</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure</td>
<td>3.72</td>
<td>2</td>
<td>0.156</td>
</tr>
</tbody>
</table>

Notes: Auditor Competence: participants’ assessment of the auditors’ competence on a 100 pt scale; Verdict: binary measure of whether or not the participant indicated that the auditors were negligent; Procedure: between participants manipulation of the primary audit procedure used to test claims expense and the associated reserve as either Statistical Sampling, Population Testing, or Predictive Modeling.
### TABLE 3
**Peer Reviewers**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Statistical Sampling</th>
<th>Population Testing</th>
<th>Predictive Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Familiarity</strong></td>
<td>Mean</td>
<td>82.66</td>
<td>58.82</td>
</tr>
<tr>
<td></td>
<td>Std Dev</td>
<td>17.84</td>
<td>25.41</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>88</td>
<td>87</td>
</tr>
<tr>
<td><strong>Quality</strong></td>
<td>Mean</td>
<td>70.74</td>
<td>70.91</td>
</tr>
<tr>
<td></td>
<td>Std Dev</td>
<td>16.84</td>
<td>20.91</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>88</td>
<td>87</td>
</tr>
</tbody>
</table>

**Panel B: Analysis of Variance – Familiarity**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>procedure</td>
<td>92,688.16</td>
<td>2</td>
<td>46,344.08</td>
<td>107.78</td>
<td>0.001</td>
</tr>
<tr>
<td>rater</td>
<td>66,453.04</td>
<td>87</td>
<td>763.83</td>
<td>1.776</td>
<td>0.001</td>
</tr>
<tr>
<td>Error</td>
<td>73,095.84</td>
<td>170</td>
<td>429.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Panel C: Analysis of Variance – Quality**

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>procedure</td>
<td>624.90</td>
<td>2</td>
<td>312.45</td>
<td>1.10</td>
<td>0.336</td>
</tr>
<tr>
<td>rater</td>
<td>49,781.33</td>
<td>87</td>
<td>572.20</td>
<td>2.01</td>
<td>0.001</td>
</tr>
<tr>
<td>Error</td>
<td>48,352.10</td>
<td>170</td>
<td>284.42</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Notes:**
- **Familiarity**: participants’ assessment of their familiarity with each audit procedure on a 100 pt scale;
- **Quality**: participants’ assessment of the quality of each audit procedure on a 100 pt scale;
- **Procedure**: between participants manipulation of the primary audit procedure used to test claims expense and the associated reserve as either Statistical Sampling, Population Testing, or Predictive Modeling;