Psychophysiological Responses to Data Visualization and Visualization Effects on Auditors’ Judgments and Audit Quality

ABSTRACT

We conduct experiments with practicing Big 4 auditors and business students in order to investigate the psychophysiological responses to Big Data visualizations and the effects of different visualization techniques on auditor judgment and ultimately audit quality. More specifically, the first experiment with students examines whether visualizations can be designed to increase the level of a users’ arousal. Such increases in arousal have the capacity to yield significant benefits to the audit profession by drawing auditors’ attention to important patterns in data and promoting the evaluation of these patterns during evidence evaluation. Results of the first experiment using cognitive pupillometry and eye gaze measurement indicate that different visualization techniques produce significant differences in the level of arousal without interfering with information evaluation efficiency. The second experiment then investigates whether visualizations that were shown to promote higher and lower levels of arousal have differential effects on auditor judgments and audit quality. In addition, the second experiment investigates whether the reliability of the data sources underlying visualizations affect auditors’ judgments. Results from the second experiment indicate that visualizations that increase arousal enhance auditors’ ability to recognize disconfirming evidence and incorporate this evidence into their decisions. That is, auditors who view visualizations of disconfirming evidence that are designed to promote arousal recommend greater reductions to management estimates of reported revenue and increase their budgeted audit hours more than auditors who view visualizations that promote less arousal. In addition, auditors who view visualizations that increase arousal are more likely to attend to the reliability of data used to create the visualizations. Overall, the experiments reveal that understanding the root causes of different visualization techniques on arousal and auditor judgment present multiple opportunities to enhance audit quality.

Keywords: arousal, auditor judgment, Big Data, data reliability, eye tracking, psychophysiological response, pupillometry, visualization.

Psychophysiological Responses to Data Visualization and Visualization Effects on Auditors’ Judgments and Audit Quality

I. INTRODUCTION

In recent years, the largest accounting firms have devoted significant resources to developing sophisticated analytical tools that can leverage the value of emerging data sources such as Big Data. New data sources are changing the way auditors obtain and assess audit evidence and appear to have significant capacity to improve audit quality (Rose, Rose, Sanderson, and Thibodeau 2017; Brown-Liburd, Issa and Lombardi 2015; Vasarhelyi, Kogan, and Tuttle 2015). Despite the advances that have been made regarding the use of improved analytical tools and Big Data in the financial statement auditing process, deployments of innovations in the field, such as the use of visualizations of Big Data, have been slow to materialize. There appears to be reluctance among partners on public company audits to fully integrate analytics of emerging data sources into the auditing process or include visualizations of Big Data in the audit workpapers (Gepp, Linnenluecke, Terrence, and Smith 2018). Much of this reluctance stems from a lack of understanding of the potential effects of data visualizations on auditor judgment and beliefs that current PCAOB regulation does not require data visualization or the use of emerging data sources (Franzel, Rose, Thibodeau, and Williams 2018). Therefore, to inform audit firms and regulators of the potential costs and benefits of using visualizations of Big Data as audit evidence, it is critical for research to examine how visualizations of Big Data affect auditor judgment and audit quality.

Research into the effects of visualizations of Big Data on audit quality is also essential because auditors are required to employ a balanced evaluation of evidence that both supports and contradicts management’s assertions (PCAOB 2010a; 2010b). Contrary to regulatory requirements to document evidence that contradicts management assertions, prior literature finds that auditors often fail to adequately consider disconfirming evidence in their judgment and decision-making processes (e.g., Cloyd and Spilker 1999; Kadous and Magro 2001; Asare and Wright 2003; Earley, Hoffman, and Joe 2008; Thayer 2011). Because third-party data sources are potentially more likely than client-provided data to yield evidence that contradicts managements' assertions, Big Data visualizations have the potential to and are likely to become a key source of disconfirming evidence (Rose et al. 2017).

Before audit firms can employ Big Data visualizations as a customary source of disconfirming evidence, however, it is critical to understand how different types of visualizations could lead to different auditor judgments and how different types of visualizations could affect audit quality. In this study, we employ psychophysiological measures of eye gaze and cognitive pupillometry to determine how different visualizations of the same data affect the arousal levels of visualization users. By investigating the root psychophysiological responses to visualization techniques, we are able to examine core constructs of visualization design that have the capacity to affect auditor judgment and audit quality. We then investigate the effects of visualizations that increase arousal on auditor judgment and audit quality. Thus, in one study, we are able to provide researchers, audit firms, and regulators with a theory-driven examination of the potential effects of new methods of evidence evaluation on auditor judgment and audit quality.

In the first experiment, we measure psychophysiological responses to different visualization techniques of Big Data that are employed as audit evidence. The results of the first experiment with business students indicate that different visualization techniques produce significantly different levels of cognitive and emotional arousal. Specifically, we find increased pupil dilation in response to Big Data that is presented in a Word Cloud format, relative to a Bar

Graph format. These differences in pupillary response are indicative of increased cognitive and emotional arousal (e.g., Kahneman and Beatty 1966; Stanners, Coulter, Sweet, and Murphy 1979; Verney, Granholm, and Dionisio 2001; Hayhoe and Ballard 2005; Nuthmann and van der Meer 2005; Wedel and Pieters 2008; Rayner 2009; Day 2010; Reutskaja et al. 2011), which translate into increased top-down cognitive processing control, and associated with improved attention, executive functioning (i.e., higher order cognitive skills) and memory encoding (e.g., Van Steenbergen and Band 2013; Querino, dos Santos, Ginani, Nicolau, Miranda, Romano-Silva, and Malloy-Diniz 2015). Importantly, the different visualization formats do not increase the time required to analyze the visualizations, and there are no significant differences in gaze times between the formats. Similarly, we find no evidence of differences in the number of gazes (eye fixations) on the visualizations, indicating that the potential benefits of increased arousal can be achieved without causing significant reductions in the efficiency of evidence evaluation.

The second experiment involves a 2 X 2 between-participant experiment using 120 auditor participants from two Big 4 firms, which employs the same visualizations that were evaluated for their effects on arousal in the first experiment. We manipulate two constructs related to visualization of data: *Arousal* (less versus more) and data *Reliability* (less versus more). The experiment includes a manipulation of data reliability because data from external sources such as Big Data are generally more challenging to verify than client-provided data. That is, Big Data sources are often easily manipulated or hacked, and such alterations are difficult to detect (Appelbaum 2016; Nearon 2005). This is in stark contrast to the prevailing view that audit evidence obtained from sources external to the client can be more reliable than evidence obtained from the client. As a result, auditors may over-rely on unreliable data when visualizations are designed to promote high levels of arousal, and it is therefore important to examine how visualization design and data reliability work together to affect auditor judgment and audit quality.

The results of the second experiment indicate that when there is evidence to disconfirm a management assertion, visualizations that promote higher levels of arousal, relative to less arousal, cause auditors to pay more attention to disconfirming audit evidence, to perceive that a misstatement at the overall financial statement level is more likely, and to increase planned substantive testing hours. Thus, in line with existing literature, visualizations that are designed to increase arousal will increase the likelihood that auditors will recognize disconfirming audit evidence and incorporate this evidence into their decision-making processes. There is no evidence of a main effect of data reliability on auditor judgments, which suggests there are opportunities to enhance auditor training in this area. However, when visualizations promote higher levels of arousal, auditors are more likely to consider the reliability of the underlying data when they are evaluating specific accounts related to the Big Data visualization. Overall, these results indicate that designing visualizations to increase arousal offers multiple and meaningful benefits to audit firms.

Our study is important for several reasons. First, this study builds on the evolving literature on Big Data visualization and auditor consideration of disconfirming evidence (e.g., Rose et al. 2017; Hackenbrack and Nelson 1996; Kadous, Kennedy and Peecher 2003; Kadous, Magro and Spilker 2008; Earley, Hoffman, and Joe 2008) by providing evidence that the increased arousal improves auditor consideration of disconfirming evidence. Second, we find that auditors attend more to the reliability of the underlying data used to generate Big Data visualizations when the visualization promotes higher levels of arousal. Given that the issue of reliability is critical to the potential use of Big Data in the audit process (Brown-Liburd, Issa, and Lombardi 2015;

Vasarhelyi, Kogan, and Tuttle 2015), our study is timely and addresses an issue that is important to practitioners as they consider the effects of Big Data visualizations on the audit process and standard setters as they consider changes to audit standards in light of advanced analytical tools and Big Data (Franzel et al. 2018). Finally, with regards to developing a deeper understanding of the role of Big Data and visualizations in the audit process, we provide an approach to evaluate the root psychophysiological responses to different types of visualizations that can be applied before field deployment of new methods of data analysis and evidence presentation.

II. BACKGROUND AND HYPOTHESES DEVELOPMENT

Big Data and Evidence Evaluation in Financial Statement Auditing

Big Data is revolutionizing the way auditors gather and evaluate audit evidence and make auditing decisions (Rose, et al. 2017; Brown-Liburd et al. 2015; Vasarhelyi et al. 2015). Whereas auditors typically gather traditional audit evidence by examining historical accounting records from client sources, making observations or obtaining confirmations from third parties, Big Data offers auditors an alternate source of evidence that is real-time, electronic, voluminous, and from a wide variety of sources (Yoon, Hoogduin, and Zhang 2015). Using Big Data evidence as part of the financial statement audit allows auditors to access data from nontraditional sources to corroborate management assertions and employ a more holistic approach to evidence evaluation with goals of decreasing the probability of material misstatement and audit failure (Yoon et al. 2015). For example, where a client's sales forecasts are not available or are of poor quality, auditors can use Big Data from social networks, news articles or product discussion forums to obtain an understanding of the client's sales trends (Yoon et al. 2015). Similarly, analysis of email-related Big Data can assist auditors in understanding employee sentiment and motivations as auditors seek to detect evidence of fraudulent behavior (Yoon et al. 2015; Holton 2009).

Compared with traditional audit evidence, Appelbaum (2016) indicates that the provenance of Big Data from external or third-party (non-client) sources is often difficult to verify. Digital data can be easily altered, and the alterations may be undetectable without examination of the actual underlying data files and the controls surrounding the access and storage of such data (Appelbaum 2016; Nearon 2005). Consequently, Big Data from sources which are external to the client versus Big Data that are generated within the client (e.g., social media sources versus client emails, respectively), would likely be more difficult for auditors to verify as they are not able to examine the corresponding data logs and reliability controls (Appelbaum 2016). However, there are currently no studies that investigate auditors' consideration of the reliability of Big Data sources or the potential effects of such consideration on their professional judgments. This study fills this gap in the literature by investigating how auditors evaluate Big Data from differentially reliable sources and assimilate this evidence into their judgments and decisions.

Auditing standards require auditors to employ a well-balanced evaluation of evidential matter that both corroborates and contradicts management's representations and financial statement assertions (e.g., PCAOB 2010a, AS No. 1105). Despite this mandate, research finds that auditors often fail to adequately consider contradictory or disconfirming evidence (e.g., Asare and Wright 2003; Earley, Hoffman and Joe 2008). Specifically, Asare and Wright (2003) and Earley et al. (2008) examine auditors' decision processes and find similar underweighting effects when auditors evaluate disconfirming evidence. However, Earley et al. (2008) also find that auditors' underweighting of disconfirming evidence is unintentional and that restructuring the task can help to mitigate this effect. Building on these prior studies, we extend the literature by investigating another potential factor that can mitigate auditor inattention to disconfirming audit evidence. Using a context that features Big Data visualizations as evidential matter, we examine whether the level of arousal (either cognitive or emotional) caused by the design of evidential matter can improve auditors' attention to and integration of disconfirming evidence into their judgment and decision-making process.

Psychophysiological Responses to Visualizations of Big Data

In the context of this study, arousal is regarded as the psychophysiological response to Big Data visualizations that is particularly relevant for understanding the potential effects of visualization design choices on auditor judgment. Being able to objectively measure the level of arousal is important because arousal levels significantly affect individuals' responses to stimuli, intensity of attention, and effort intensity (Kahneman 1973; Howells, Stein and Russell 2010). Without arousal, effort is not directed towards understanding and interpreting external stimuli (Howells et al. 2010; Kahneman 1973; Sirois and Brisson 2014; Mathôt et al. 2015). Thus, arousal captures multiple dimensions of the effects of visual stimuli on decision-makers, and arousal is predictive of the amount of effort, intensity of effort, and intensity of attention that is associated with a stimulus. The two main forms of arousal are cognitive arousal and emotional arousal.

Cognitive arousal levels can be effectively measured with cognitive pupillometry (Hayhoe and Ballard 2005; Wedel and Pieters 2008; Rayner 2009; Day 2010; Reutskaja et al., 2011). Cognitive pupillometry involves measurement of pupil dilation in response to visual stimuli (Sirois and Brisson 2014). Pupil dilation reveals arousal and the level of cognitive resources that an individual allocates to a stimulus or task (Kahneman and Beatty 1966; Nuthmann and van der Meer 2005; Stanners, Coulter, Sweet, and Murphy 1979; van der Meer et al. 2010; Verney, Granholm, and Dionisio 2001). By comparing changes in pupil dilation in response to a stimulus compared to a baseline period, pupillometry reveals the arousal created by a visual stimulus (Sirois and Brisson 2014; van der Wel and Steenbergen 2018). Pupillary responses also reveal emotional arousal to visual stimuli, and both positive and negative emotional reactions result in more pupil dilation relative to neutral stimuli (Partala and Surakka 2002; Bradley, Miccoli Escrig and Lang 2008).[1]

Given the potential importance of arousal on decision-making processes, the purpose of our first experiment is to examine the effects of different visualization formats for the same underlying data on levels of arousal, and we then employ these findings to test the effects of different arousal levels on practicing auditors' judgments and, ultimately, audit quality.

Effects of Arousal Levels Caused by Visualizations on Auditor Decision Making

Given the volume and variety of Big Data, sophisticated analytical approaches such as the use of data visualizations are required to effectively analyze Big Data. Visualizations provide spatial summarization and comparison of data and help individuals to assimilate and interpret data more easily (Yoon, Hoogduin and Zhang 2015; Wright 1995; Benbasat and Dexter 1986). The types of visualizations used to conceptualize data have evolved from conventional line, pie and bar charts to more sophisticated word clouds and network, arc, and alluvial diagrams (Duke University Library 2018; Yoon, Hoogduin and Zhang 2015). Recent research finds that auditors can fail to detect even simple patterns in Big Data visualizations (Rose et al. 2017), suggesting that visualizations need to be designed to facilitate auditors' pattern recognition and attention to evidential matter that disconfirms a management representation.

While there are endless formats that can be used to visualize data, the change in arousal levels created by presentation formats are known to increase individuals' cognitive processing and encoding of the information. Increased pupillary response, for example, is related to increased

[1] Pupil dilation does not, however, reveal emotional valence, also known as hedonic valence (Bradley, Miccoli, Escrig, and Lang 2008).

levels of cognitive processing (e.g., van Steenbergen and Band 2013; Querino, dos Santos, Ginani, Nicolau, Miranda, Romano-Silva, and Malloy-Diniz 2015). The literature has also shown that individuals are better able to cognitively process visual stimuli that promote higher levels of arousal, thereby allowing individuals to better interpret and incorporate related information into their decision-making processes (Keller and Block 1997; McGill and Anand 1989; Shedler and Manis 1986). Finally, the literature finds that higher levels of arousal caused by visual stimuli are associated with better storage of information in memory (e.g., Shedler and Manis 1986; Goldinger and Papesh 2012; Kucewicz, Dolezal, Kremen, Berry, Miller, Magee, Fabian and Worrell 2018), which facilitates greater recall and cognitive elaboration and increases the influence of displays on subsequent decision making (McGill and Anand 1989; MacLeod and Campbell 1992; Keller and Block 1997).

We extrapolate well-rooted findings from the psychology and neuropsychology literatures and expect that auditors will similarly increase cognitive processing and encoding of information that is presented in formats that increase arousal. Prior accounting literature has established that auditors are susceptible to a wide range of cognitive effects found in the psychology literature (e.g., Shanteau 1989; Griffith, Kadous, and Young 2015). We expect that auditors who view Big Data visualizations that promote more arousal, relative to less arousal, will be more likely to recognize and attend to disconfirming evidence in the Big Data visualizations. As a result of attending to evidence that disconfirms management's assertions, auditors who view visualizations that increase arousal will be more skeptical of management's reporting and will increase subsequent audit testing more than when they view visualizations that promote less arousal.

H1a: *Auditors who view Big Data visualizations that promote more arousal and disconfirm a management representation will assess more financial misstatement than will auditors who view Big Data visualizations that promote less arousal.*

H1b: *Auditors who view Big Data visualizations that promote more arousal and disconfirm a management representation will increase audit hours more than will auditors who view Big Data visualizations that promote less arousal.*

Audit Evidence Reliability

AS No. 1105 (PCAOB 2010a) states that when evidence from "one source is inconsistent with that obtained from another," and where the auditor has concerns about the reliability of the audit evidence, the auditor should consider the effects this evidence has on other aspects of the audit (Auditing Standard No. 1105 ¶29, PCAOB 2010a). This regulation suggests that auditors should carefully evaluate the reliability of the evidence they use in forming judgments and should consider the related effects of inconsistent evidence on their overall decision process. Prior research using more traditional evidence settings finds that in situations of increased risk of misstatement, auditors appropriately consider the audit implications of disconfirming evidence and adjust their decision making accordingly (e.g., Hackenbrack and Nelson 1996; Kadous, Kennedy and Peecher 2003; Kadous, Magro and Spilker 2008). However, there is currently no research that has examined whether auditors consider the reliability of information when audit evidence is in the form of Big Data visualizations.

The reliability of audit evidence is of critical importance to the quality of the audit process. Reliability pertains to the "nature and source of the evidence and the circumstances under which it is obtained" (Auditing Standard [AS] No. 1105 ¶8, PCAOB 2010a). While AS No. 1105 indicates that auditors are not expected to be authentication experts, they should use professional judgment to evaluate the reliability of audit evidence regardless of whether the evidence is generated from internal or external sources, and without regard to the format. Inappropriate weighting of unreliable evidence can compromise audit efficiency and effectiveness.

Overweighting of unreliable evidence can lead auditors to under-audit just as under-weighting can lead to over-auditing (Hirst 1994).

Consistent with the prescriptions in auditing standards, most prior auditing research indicates that auditors consider evidence reliability in light of its source (e.g., Knechel and Messier 1990; Hirst 1994; Reimers and Fennema 1999; Tan and Jamal 2001; Kadous, Leiby and Peecher 2013). Knechel and Messier (1990) find that auditors elect to review more reliable evidence, and they revise their judgments more when they review more reliable evidence. Hirst (1994) examines auditors' consideration of evidence from a specialist-prepared inventory report and finds that auditors consider evidence from a more trustworthy source to be more reliable than evidence from a less trustworthy source. Similarly, Reimers and Fennema (1999) find that reviewers are sensitive to the perceived trustworthiness of the evidence source. In that spirit, research also demonstrates that auditors evaluate memos more favorably when they are prepared by more competent and reliable audit managers (Tan and Jamal 2001), and auditors place greater weight on advice received from advisors perceived to be more competent (Kadous, Leiby, and Peecher 2013).

There are also studies, however, which find that auditors can fail to consider evidence reliability when evaluating audit evidence (e.g., Joyce and Biddle 1981; Jenkins and Haynes 2003). Given that evaluation of Big Data visualizations involves different decision processes than more traditional evidence evaluation tasks (i.e., examination of graphics rather than review of calculations or financial data), and auditors often never see the data underlying visualizations due to data complexity and volume, auditors may fail to adequately attend to the reliability of the data sources and integrate reliability considerations into their decision making. While we expect auditors to be less accepting of evidence when that evidence is from a source of lower reliability, the effects of Big Data visualizations on auditors' attention to data reliability remains unknown.

H2a: *Auditors who view Big Data visualizations that disconfirm a management representation and that were acquired from more reliable evidence sources will assess more financial misstatement than will auditors who view visualizations of Big Data acquired from less reliable evidence sources.*

H2b: *Auditors who view visualizations of Big Data that disconfirm a management representation and that were acquired from more reliable evidence sources will increase audit hours more than will auditors who view visualizations of Big Data acquired from less reliable evidence sources.*

Level of Arousal and Auditors' Evidence Reliability Evaluations

The research in neuropsychology and psychology described in the development of H1 finds that visualizations that increase arousal have more influence on decision makers and increase individuals' processing and storage of the information presented (e.g., Shedler and Manis 1986; Goldinger and Papesh 2012; van Steenbergen and Band 2013; Querino, dos Santos, Ginani, Nicolau, Miranda, Romano-Silva, and Malloy-Diniz 2015; Kucewicz, Dolezal, Kremen, Berry, Miller, Magee, Fabian and Worrell 2018). In turn, increased processing and storage results in improved recall and greater cognitive elaboration (Hastie and Park 1986; MacLeod and Campbell 1992; Hertwig, Pachur, and Kurzenhauser 2005). MacLeod and Campbell (1992) find that information that is more accessible in memory has a greater influence on individuals' decisions and that these effects are quite significant. Hertwig, Pachur and Kurzenhauser (2005) also examine accessibility of recollections and find that increased recollections translate to more accurate estimates and risk frequencies. These findings all suggest that visualizations that increase arousal will improve auditors' encoding of evidence into memory and will allow them to better recall and process this information when making decisions. Overall, auditors should attend more to Big Data visualizations of disconfirming evidence when these visualizations promote higher levels of arousal (either emotional or cognitive), and we propose that increased arousal also has the capacity to influence their evaluations of evidence reliability.

Given the effects of arousal on attention to information, memory and subsequent cognitive processing, we propose that auditors who review Big Data visualizations that promote higher levels of arousal will be more likely to consider the reliability of evidence than will auditors who view Big Data visualizations that promote lower levels of arousal. Specifically, we expect that auditors will pay more attention to the source of the data underlying the visualization when the visualization promotes more arousal, thereby allowing them to better discriminate against less reliable data sources in their audit judgments. Thus, data reliability will have more influence on auditors' judgments when Big Data visualizations lead to higher levels of arousal, relative to when visualizations lead to lower levels of arousal.

H3: Data reliability will have more influence on auditors' assessments of the likelihood of management misrepresentations when auditors view Big Data visualizations that promote more, versus less arousal.

III. DESIGN - EXPERIMENT ONE

The first experiment involves assessing the effects of different visualization techniques on arousal and effort, and this experiment employs psychophysiological measures based upon pupillometry and eye tracking.

Participants

Seventy-four students (62.2% female, 37.8% male) voluntarily participated in this study as part of their accounting information systems course[2]. Their recruitment was guided via a Department of Accounting student subject pool at a major Australian university, and the project was approved by the university's Human Research Ethics Committee. As an incentive, students received bonus credit in their course for participating in the experiment. The same bonus is applied

[2] This course covers risk, internal control, financial modeling, and visualization of financial and non-financial data with regards to managerial decision-making.

for participating in any experiment in the department's pool or for working on an alternative research participation task. The average age of the participants was 20.72 years (*SD* = 1.29), ranging from 18 to 25 years. On average, the participants had 0.68 years (*SD* = 0.81) of professional business experience. Approximately eight percent (8.1%) of the participants had financial statement auditing experience. All participants had corrected-to-normal vision.

Design

The experiment employed a within-participant design. Specifically, two experimental treatments were presented to the participants while the order of presentation was randomly alternated. The treatment was the *Visualization Type* used to display information about employee sentiment toward their employer, which was varied across two conditions (i.e. Word Cloud vs. Bar Graph). The visualizations are presented in Appendix I.

We expect that word cloud formats will increase arousal levels relative to bar graphs due to higher levels of vividness and emotional effects. The psychology literature indicates that more vivid visual displays attract more attention because they are more emotionally interesting and imagery provoking (McGill and Anand 1989). The word cloud display reveal more about specific emotional responses (i.e., more than valence alone) and are more imagery provoking than are bar graphs. In order to intentionally bias against our expectations that word clouds promote higher levels of arousal, we designed the Bar Graph to contain a more complex visual representation with more points of interest, which would be expected to produce stronger pupillary responses and fixation counts relative to the word cloud. Stronger responses to the Bar Graph could only be offset by pupillary responses to the word cloud if the word cloud format produces higher levels of emotional and cognitive arousal. Similarly, we also provided additional information in the Bar Graph (i.e., information about the time series of sentiment that indicates persistent negative valence) in order to further bias against our expectations and increase the potential for negative emotional arousal. Again, this design choice is expected to increase pupillary response, fixation count and fixation duration for the Bar Graph. Thus, if we find that pupillary responses are stronger for the Word Cloud, we have high confidence that a greater cognitive and/or emotional arousal is produced by a Word Cloud format relative to a Bar Graph format.

Dependent Variable

The visualizations, which were presented sequentially on two separate screens, formed Areas of Interest (hereafter AOIs), which were used to calculate measures of arousal and visual attention, including pupillary dilation, fixation duration, and fixation count. The three dependent variables (DVs) used in this study represent individual psychophysiological responses observed while attending to the different visualizations of sentiment data. The first DV, *Pupillary Response*, is a continuous proportional variable, which reflects an individual’s percentage change between (1) pupil dilation or constriction in response to attending to either the world cloud or Bar Graph AOI and (2) the baseline pupil diameter, calculated as the mean pupil dilation level observed while attending both AOIs by an individual. Using an individual’s mean response while attending to both AOIs as a baseline provides a conservative and highly reliable measure of change associated with attending to each individual AOI, while also making it possible to control for individual differences in pupil size. In recent reviews of the experimental cognitive pupillometry literature, increased proportional pupillary response is associated with higher levels of cognitive and emotional arousal, as well as attentional effort (Sirois and Brisson, 2014; van der Wel and van Steenbergen 2018).

Using this measure, the visualization that results in a larger increase in proportional pupillary response is the visualization that produces higher levels of arousal.

The remaining two DVs reflect characteristics of participants' visual attention and take into consideration either the number or duration of their fixations on each of the two visualizations (i.e., AOIs). The second DV, *Fixation Count*, is measured using eye tracking technology and represents the number of times the participant fixates on an AOI. A fixation occurs when the eyes stay focused on a single location on the screen of the eye-tracking device (used for presenting all experimental materials) longer than a particular time threshold (Holmqvist et al. 2011). As discussed in more detail in the Materials and Apparatus section, a fixation is recorded when participants fixate at least 60ms on a particular spot on the screen. *Fixation Count* is among the most widely used measures in eye tracking research (Holmqvist et al. 2011) and has recently been applied in accounting research (Fehrenbacher, Schulz and Rotaru 2018).

Finally, the third DV, *Fixation Duration*, is another commonly used eye tracking measure that has been adopted in the accounting literature (Chen, Jermias, and Panggabean 2016) which measures the length of time of fixation on an AOI. *Fixation Duration* is strongly associated with the fixation count measure (Holmqvist et al. 2011). This measure made it possible to investigate whether one of the visualizations is associated with more effort duration than the other visualization. We employ the fixation measures to determine whether the two visualizations produce different levels of attention and time spent analyzing the visualization details.

Materials and Apparatus

Decision Case

The sentiment data presented to the participants formed part of the hypothetical scenario in which participants were asked to assume that they were evaluating a claim made by a company (in the scenario presented, the company was referred to as Absolute Solutions Agency Inc., or Absolute). The participants were provided with a brief outline of the company's business activity and one of its key performance indicators (namely, employee turnover). The participants read information about the performance estimates of the company's management, after which they were presented with two visualizations of independently collected data from social media web sites. The two visualizations (Word Cloud versus Bar Graph) served as the treatment conditions, and we were interested in determining the level of arousal created by each visualization type. The information communicated via the two visualizations disconfirmed the highly positive sentiment data provided by the company's management.

Psychophysiological Apparatus

Eye movements were recorded using a table-mounted eye tracking system (Tobii TX300) with a temporal resolution of 300 Hertz (Hz) and a screen resolution of 1920 x 1080 pixels. See Appendix II for a more complete overview of technical specifications of the eye tracking measurements conducted in this study. The average viewing distance was 65 cm from the screen (range: 50-80 cm), binocular accuracy was 0.4° and precision 0.14°.[3] Fixations were computed using the velocity-based I-VT algorithm (Komogortsev et al. 2010). To define fixations, a rather conservative 60ms threshold was selected, denoting that any fixation below 60ms threshold was considered as a saccade (a rapid movement of the eye between fixation points) and was not included in the analysis.

The experiment was conducted in a light-controlled, dimly lit booth. Participants sat on height-adjustable chairs with their head supported by a height-adjustable forehead and chin rest

[3] The Tobii TX300 infrared system has precision of 0.14 degrees and accuracy of 0.4 degrees. For more detailed product specifications, please refer to http://www.tobiipro.com/product-listing/tobii-pro-tx300/

(Heavy Duty Chin Rest with Clamps from Richmond Products, Inc.). At the beginning of the experiment, the eye tracker was calibrated using a nine-point fixation technique, which is the most rigorous calibration technique for the device used. This calibration adjusts for participants' individual differences in eye characteristics and their seating position.

Procedure

At the beginning of the experimental session, participants sat in front of the TobiiTX300 eye tracking system, and a 9-point calibration procedure was performed as described in the Materials and Apparatus section and in Appendix II. The participants were randomly assigned to one of the two visual order representation conditions (either Word Cloud first or Bar Graph first), upon which they were presented with the background information about the case. On the page where the participants were presented with the background information about the hypothetical case (see the Decision Case section), the participants were reminded that they could spend as much time as needed to read and comprehend the instructions and that once they proceeded to the next page, this information would no longer be available to them. Upon pressing the space bar key, the participants were presented with the first visualization, which was time-limited for 90 seconds. In the instructions, the participants were told that this page will be presented to them during a specific amount of time, upon which a new page will automatically appear. The two visualizations were administered in consecutive order for the duration of 90 seconds each. After this, the participants responded to a brief post-experimental questionnaire where we collected demographic information.

IV. RESULTS AND ANALYSES – EXPERIMENT ONE

The first analysis employs cognitive pupillometry to evaluate the effects of the Word Cloud and the Bar Graph on arousal levels. Table 1 presents the results of a repeated-measures ANCOVA where the within-participant measures are the pupil dilation ratios for the Bar Graph and Word Cloud, and a covariate is included to control for the order of presentation of the two visualizations. After controlling for order, there is a significant effect of visualization type on pupil dilation (F = 120.52, p < 0.001), and the dilation ratio is higher for the Word Cloud (1.023) than for the Bar Graph (0.979). Thus, the results indicate that the Word Cloud causes significantly higher levels of arousal than does the Bar Graph.

[Insert Table 1 about here]

We next analyze eye tracking data in order to determine whether participants spend more time examining either visualization type or analyze more specific features of either visualization type. For both analyses, we control for the order of presentation by including order as a covariate. Results for the two eye-tracking DVs are presented in Table 2. There are no significant differences in either *Fixation Count* (p = 0.315) or *Fixation Duration* (p = 0.334) between the Word Cloud and the Bar Graph. Thus, while the Word Cloud creates higher levels of arousal than does the Bar Graph, users of these visualizations spend approximately the same amount of time analyzing the visualizations and evaluate similar numbers of features. These results indicate that increases in arousal can be achieved without increasing the time demands needed to analyze the visualizations, which is favorable news for the potential to increase arousal while maintaining audit efficiency.

[Insert Table 2 about here]

V. DESIGN - EXPERIMENT TWO

Revenue misstatement is always a significant risk and concern for auditors during the financial statement audit process. In fact, Auditing Standard No. 2110 explicitly emphasizes that auditors should design audit procedures to detect “unusual or unexpected relationships involving revenue accounts that might indicate a material misstatement” (PCAOB 2010b). In addition, research consistently finds that revenue misstatement is a significant audit risk factor and one of the leading causes of sanctions and the issuance of Accounting and Auditing Enforcement Releases (AAERs) by the U.S. Securities and Exchange Commission (e.g., Dechow, Ge, Larson, and Sloan 2011; Marquardt and Wiedman 2004). We examine how arousal levels of visualizations and the source of Big Data evidence affect auditors' evaluation of audit evidence and their subsequent judgments within a contemporary revenue recognition context.

Participants

We recruited auditors from two Big 4 public accounting firms. Participants completed the instrument during firm training sessions, and multiple authors attended the training sessions to ensure appropriate administration of the experiment. One hundred and forty-three auditors completed the experiment. Of these participants, 23 provided responses that were outside the possible ranges of the questions (e.g., above total possible revenues), indicating that these participants failed to attend to the case. Consequently, we do not include these participants in our analyses resulting in a final sample of 120 auditors. Table 3 presents the descriptive statistics for these participants. Participants had an average of 2.7 years of auditing experience, 2.0 years of experience conducting analytical reviews, and 59% were male.

[Insert Table 3 about here]

Task Description and Experimental Procedures

The experimental procedures are depicted in Figure 1. Participants read a hypothetical case describing a publicly traded professional services audit client company that provides management services (e.g., human resources, accounting and administration, and sales and marketing). The case begins with background information about the client. Historically, the audit client provided services only to not-for-profit entities but extended service offerings to for-profit entities during the fourth quarter of the current year under audit. The case elaborates that the audit client earns revenue primarily from management fees charged for these professional services. The case then presents participants with summary results of third quarter and fiscal year-end financial reporting for the current year and the prior year. Reported results include revenue, cost of services, gross margin and gross margin percentage. In addition, year-end reporting (i.e., revenue, cost of sales, etc.) for the current year is also broken down for each of the two divisions – not-for-profit and for-profit. The reporting shows a significant increase in revenue over the two years, which management attributes to the recent service expansion into the for-profit sector.

Participants are then told that the manager on the audit requests that they review revenue related to the audit client's sole contract recorded in the for-profit division during Q4. In this contract, the audit client provides three management service offerings. As a provision of the contract, the audit client receives a standard quarterly fee and can earn performance incentive bonuses if certain criteria are satisfied. Of particular interest in the case, the audit client can earn a performance incentive bonus for human resources services if the customer's employee turnover is low (i.e., below an established threshold) at the end of each quarter. Although the customer provides audited employee turnover information four months after quarter end, due to financial reporting time constraints, the audit client estimated the maximum Q4 incentive bonus for human resource services at year-end.

There are five data visualizations included in the experimental instrument. We use visualization types with which our auditor participants are familiar to reduce familiarity effects. In addition, our participants have previously received training on the use of the graphical displays used in the experiment. Participants are presented with data analytic visualizations related to the customer's employees as well as the audit client's performance relative to other competitors in the industry. The case deliberately presents visualizations that are informative as well as uninformative with regards to the revenue and human resources performance bonus recognized in Q4. The first visualization is uninformative and presents weekly hashtag analysis that compares the number of social media mentions of the audit client to that of their competitors. The second and third visualizations are informative with respect to the performance bonus revenue. The second visualization reports the results of the customer's employee engagement survey that was administered by the client during Q3. This visualization suggests positive employee morale and supports the audit client's expectation of low employee turnover, which lead to the recording of the entire performance incentive bonus at year-end.

The third visualization is manipulated between conditions, and provides disconfirming evidence suggesting poor employee morale. This visualization suggests the possibility of high employee turnover and inappropriate recording of the performance incentive bonus at year-end. Using the findings from the first experiment, the manipulation of arousal involves presenting the same underlying data as either a Word Cloud or a Bar Graph. Data reliability is manipulated by altering the source of the underlying data as coming from text sentiments analyses of client emails versus social media postings made by the customer's employees. By design, employees' sentiments are more negative than positive such that they disconfirm the audit client's assertions. The fourth and fifth visualizations are informative about the client's revenue in general and present the number of new customers obtained by the audit client compared to its three main competitors, as well as the related revenue trends over the same period. After reviewing the visualizations, participants answered dependent variable questions, questions about their perceptions of employees' emotions and perceptions of evidence reliability. Finally, they completed post-experiment and demographic questions (see Figure 1).[4]

[Insert Figure 1 about here]

Independent Variables

We employ a 2 X 2 between-participant full factorial design where we manipulate the arousal level of a visualization and the reliability of the underlying data source for the visualization. The first independent variable, *Arousal*, is manipulated on two levels (Higher Arousal versus Lower Arousal). To operationalize arousal, we employ the two visualizations that were analyzed in experiment one. We employ a Word Cloud to operationalize higher levels of arousal and a Bar Graph to operationalize lower levels of arousal. The Word Cloud contains emotionally charged words (both negative and positive) that express more details about employee emotion than the positive and negative cues in the Bar Graph. In the higher arousal condition, the Word Cloud presents results of employees' text sentiment analyses in order to determine whether the overall sentiment in the postings represent a positive, negative, or neutral attitude towards the audit client. Word Clouds are generated using algorithms to present textual data analyses by manipulating the visual characteristics of words using font size, color and weight. Unlike other forms of visualizations, Word Clouds combine the data variable with the data label (i.e., the word itself) and use variations in font size, and word proximity to convey information about the frequency and provide rich context of how the words appear in the textual data.

We operationalize the lower arousal condition by using a Bar Graph to visualize the sentiments in the same textual data used to generate the Word Cloud. While the Bar Graph depicts

[4] We developed the final experimental instrument after conducting pilot testing with current and former auditing participants. We also refined the instrument based on feedback from partners and research committee members from one of the participating Big 4 firms. The partners found the case to be relevant, timely and realistic.

the extent of employees' positive, negative and neutral attitude towards the client using the height of the bars, unlike the Word Cloud, it does not provide striking imagery of the sentiments (see Appendix I). Both visualizations were derived from the same underlying data in order to represent the same data in formats with differing levels of arousal. Thus, much like audit firms must choose how to visualize data, we examine how these important design choices affect auditor decision processes.

The second independent variable, *Reliability*, is manipulated on two levels (More Reliable versus Less Reliable). To operationalize *Reliability*, we manipulate the source of the underlying data used to generate the visualizations. This is in keeping with guidance from auditing standards on audit evidence (e.g., AS No. 1105). Evidence from sources over which the audit client has controls that are verifiable by the auditors are considered to be more reliable compared with evidence from sources over which there are less verifiable controls. Therefore, consistent with auditing standards and with current Big Data literature, we operationalize the more reliable condition by using email data which is communicated by the customer to the audit client and is more verifiable by the auditors (PCAOB 2010; Vasarhelyi et al. 2015; Applebaum 2016). Participants in the More Reliable condition are told:

> *The visualization group analyzed emails sent by [Customer Name] employees to the [Customer Name] Human Resources department and upper management and conducted text sentiment analyses on these emails to determine…*

We operationalize the Less Reliable condition by using data from social media sources (e.g., Twitter, Facebook, and Google+). Evidence generated by social media sites are removed from the audit client's control and are less verifiable by the auditors. Participants in the less reliable condition are told:

> *The visualization group analyzed social media postings that are believed to have been made by [Customer Name] employees and conducted text sentiment analyses on these postings to determine…*

Dependent Variables

We focus our analyses on two dependent variables. First, we measure *Misstatement Amount* by asking participants to indicate the amount of performance incentive revenue they recommend be recorded: "Assuming that you need to make a recommendation to the engagement partner regarding any need to adjust the client's revenue from the new for-profit contract with [Customer Name], what amount of revenue would you recommend to the partner as the appropriate amount to be recorded for the Human Resources performance incentive?" Participants were instructed to indicate an amount between $0 to $4 million (the maximum amount of the incentive bonus). Second, to determine the effects of the manipulations on audit processes, we measure how auditors would adjust budgeted audit hours for substantive revenue testing. We provide auditors with a baseline of audit hours taken to audit revenue in the prior year (i.e., 100 hours). We then ask participants to indicate the hours they would budget for the current year's audit. To measure *Audit Hours*, we ask participants to "Indicate how many hours you want to budget this year for substantive testing of the revenue account."

IV. RESULTS AND ANALYSES – EXPERIMENT TWO

Manipulation Checks

We used several measures to determine whether manipulations of *Arousal* and *Reliability* effectively influenced the underlying constructs of interest. We previously analyzed the effects of the two visualization formats on arousal in Experiment One. It was not possible to measure psychophysiological responses to the visualizations in the experiment with auditors. Thus, we rely upon the findings from the first experiment to support differences in levels of arousal between the two visualizations. However, it is possible to measure auditors' recognition of the emotional content of the visualization, and we collect a measure of their emotional response. We evaluate participants' perceptions of the *Emotions* represented by the visualizations audit evidence using five questions. Participants indicated their perceptions that employees were 1) happy, 2) discouraged, 3) angry, 4) frustrated, and 5) depressed. Responses were measured on an anchored 100-point scale where 0 = "Not at All" and 100 = "Completely." We formed a net measure of *Emotions* by subtracting responses to negatively framed emotional questions (i.e., discouraged, angry, frustrated, and depressed) from the positively framed emotional question (i.e., employees were happy).[5] The possible values for this measure range from – 400 to +100. Comparing the mean *Emotions* score of participants in the Higher Arousal condition to those in the Lower Arousal condition, we find that participants in the Higher Arousal condition perceived significantly stronger negative emotions from the case materials than those in the Lower Arousal condition (-208.23 versus -133.46, respectively; $p < 0.001$).

To determine whether the *Reliability* manipulation was effective, we employed three measures. For the first measure, we ask, "In your opinion, how reliable was the data that was used to create the visualization of employee sentiment?" Responses were measured on a 100-point anchored scale with 0% indicating "Not Reliable at All" and 100% indicating "Completely Reliable." We find that participants in the More Reliable condition reported higher mean scores to the reliability question than participants in the Less Reliable condition (46.72 versus 38.76, respectively; $p = 0.063$). Using post-experimental questions, we further evaluate within-participant perceptions of the verifiability of email data sources versus social media data sources. Responses

[5] The values for the Emotions measure range from – 400 to 100, where -400 represents the lowest negative interpretation of the employees' emotions in the case materials, and 100 represents the highest positive interpretation. Factor analysis confirms that all questions load on one factor.

were measured on an anchored 100-point scale where 0 = Definitely Cannot Verify; 100 = Definitely Can Verify. Results indicate that auditors perceive email to be a more verifiable data source than social media (54.04 versus 44.38, respectively; $p = 0.005$). Similarly, we asked participants about the reliability of an email data source compared with a social media data source. Responses were measured on an anchored 100-point scale where 0 = Definitely Not Reliable; 100 = Definitely Reliable. Results indicate that auditors perceive email to be a more reliable data source than social media (40.67 versus 26.60, respectively; $p < 0.001$).

Hypothesis Testing

H1a predicts that auditors who view Big Data visualizations that provide disconfirming evidence of a management estimate and promote higher levels of arousal will assess more misstatement than will auditors who view visualizations that promote lower levels of arousal. H1a and other hypotheses related to auditor assessment of misstatement are evaluated with an ANCOVA model where the independent variables are *Arousal* and *Reliability*, the covariate is audit experience, and the dependent variable is auditors' assessment of the amount of revenue that should have been recognized for the human resources incentive bonus (see Table 4). We do not find a main effect for *Arousal* of Big Data visualizations on *Misstatement Amount* ($p = 0.265$, one-tailed). However, there is a significant and disordinal interaction of *Arousal* and *Reliability* ($p = 0.029$, one-tailed), which is discussed further in our analyses of the interaction.[6]

[Insert Table 4 about here]

[6] We also analyzed the main effect of *Arousal* using an alternative dependent variable (a scaled rating of the likelihood that Total Revenues are overstated). Using this alternate scale measurement (not tabulated), there is a significant main effect of *Arousal* on perceptions that Total Revenues are misstated ($p = 0.050$), and auditors perceive that revenues are more likely to be overstated when the disconfirming visualization produces more arousal, relative to less arousal.

In H1b, we posit that auditors who view Big Data visualizations that disconfirm management's estimates and promote more arousal will increase audit hours more than when auditors view visualizations that promote less arousal. To test the audit hours hypotheses, we employ an ANCOVA model where the dependent variable is auditors' determination of budgeted *Audit Hours*.[7] We find results consistent with this prediction. Auditors in the Higher Arousal condition report that they would budget significantly more audit hours than auditors in the Lower Arousal condition (means = 139.93 versus 129.28, respectively; p = 0.034; Table 5 Panel B), which supports H1b.

[Insert Table 5 about here]

H2a posits that auditors who view disconfirming Big Data visualizations from more reliable sources will assess more revenue misstatement relative to auditors who view visualizations from less reliable sources. Consistent with our prediction, we find evidence of a main effect (p = 0.056, Table 4 Panel B). However, similar to the main effect of *Arousal* on *Misstatement Amount*, this result must be considered in light of the significant interaction between *Arousal* and *Reliability*.

H2b proports an effect of *Reliability* on auditors' budgeted hours decisions. This hypothesis proposes that auditors who view disconfirming Big Data evidence from more reliable sources will increase their budgeted audit hours more than when evidence is from less reliable sources. We do not find support for this prediction. Table 5 Panel B shows that there is no main effect of *Reliability* on *Audit Hours* (p = 0.365). These results suggest that when evaluating audit evidence in the form of Big Data visualizations, auditors may not adequately integrate evidence reliability into their

[7] The model includes a covariate for years of audit experience because preliminary analyses revealed that audit experience is a significant determinant of budgeting decisions.

audit planning decisions, even though they are influenced by the *Arousal* level triggered by the visualizations.

H3 predicts that auditors who view visualizations that promote higher levels of arousal will be more influenced by the reliability of audit evidence than will auditors who view visualizations that promote lower levels of arousal. To test this interaction hypothesis, we examine the interactive effects of *Arousal* and *Reliability* on *Misstatement Amount*. In Table 4 Panel B, the overall ANCOVA model indicates a significant interaction (p = 0.057; Figure 2) and provides evidence to support our hypothesis. To further evaluate this interaction result, we examine planned contrast for the effects of *Reliability* on *Misstatement Amount* in each *Arousal* condition. In the Higher Arousal condition, participants suggest larger reductions to the amount of revenue recorded by management when the data underlying the disconfirming evidence are more reliable relative to when the data are less reliable (means = $2.96M versus $1.97M, respectively; p = 0.010; Table 4 Panel C). However, when in the Lower Arousal condition, there is no statistical difference in the *Misstatement Amounts* participants suggest across the *Reliability* conditions (p = 0.810).

Results of the planned contrasts indicate that the arousal level created by the Big Data visualization is an important factor that facilitates auditor integration of evidence source reliability into their decision making. The analyses of the interaction also reveal that the expected effects of arousal on revenue misstatement decisions only materialized when the data underlying the visualization were more reliable. That is, visualizations that disconfirm management assertions and promote higher arousal levels caused auditors to be more likely to determine that management had overstated revenues when the data used to create these visualizations were from more reliable, relative to less reliable sources.

[Insert Figure 2 about here]

Supplemental Analysis

Our theory predicts that higher arousal levels will cause auditors to have stronger reactions to disconfirming audit evidence. While we did not find a main effect of visualizations on assessments of the amount of revenue to record, we did find a main effect of *Arousal* on an alternative scale measure of perceptions of total revenue overstatement. To investigate the potential for mediating effects of auditors' responses to the emotional content of visualizations (*Emotions*) on the relationship between the *Arousal* of the visualization and assessment of Total Revenue overstatement, we apply the Hayes and Preacher (2014) bootstrapped mediation technique. Utilizing this approach, we use 5,000 bootstrap sampling iterations of the data to compute a bias-corrected 95% confidence interval for the indirect effect of *Emotions* that we propose. Consistent with our theoretical expectations, *Emotions* mediate the relationship between *Arousal* and perceptions of the likelihood of Total Revenue overstatement (95% CI = LL: 0.043; UL: 0.459; where the confidence interval excludes zero; not tabled).

VII. CONCLUSIONS

We examined the effects of data visualization on auditor judgment by conducting two experiments with practicing Big 4 auditors and business students. We conducted these experiments within the context of revenue recognition standards that allow for significant management judgment, and we extend the literature that has examined auditor integration of disconfirming audit evidence into their decision making. Taken together, the two experiments allow auditors, regulators and academics to better understand how new evidence sources and different types of visualizations influence auditor judgments and, ultimately, audit quality.

Our first experiment employed a novel psychophysiological approach to examine the potential for different visualization formats to create significantly different levels of user arousal.

By employing cognitive pupillometry, we demonstrate that different visualizations of the same audit evidence can yield different levels of arousal. Prior psychology research finds that arousal levels are important determinants of attention, cognitive processing, and memory encoding, and we propose that designing visualizations to enhance user arousal levels presents important opportunities to improve audit quality. In addition to measuring pupillary responses to visualizations, the first experiment also employed eye-tracking techniques to determine whether the two visualization types we investigate result in reductions in the efficiency of audit evidence examination. We found no significant differences in fixation duration or fixation counts between the two visualization types which suggests that increased arousal can be achieved without causing significant decreases in the efficiency of evidence evaluation.

The second experiment employed the visualizations from the first experiment in an audit judgment context. This experiment involved auditors from two Big 4 firms where we manipulated the format of a data visualization (Bar Graph versus Word Cloud) and the source of data (social media versus client email) used to create a visualization that disconfirmed a management estimate of revenue. The theoretical constructs of interest were the arousal level created by the visualizations and the reliability of data source used to create the visualizations. Auditors have many options when determining how to visualize Big Data and other emerging data sources, and these visualization decisions could have significant implications for audit quality. We examined the potential for different visualizations to increase auditor attention to the disconfirming nature of evidence they review. Further, the experiment investigated the effects of data reliability on auditor decisions because no prior research has investigated how auditors integrate the reliability of data sources into their decisions when they interpret visualizations of Big Data and other contemporary data sources. If auditors fail to attend to the reliability of data when evaluating visualizations, there would be significant threats to audit quality.

Results of the auditor experiment indicate that creating visualizations that increase arousal caused auditors to place more weight on the disconfirming nature of the visualizations when assessing potential revenue misstatement and when determining audit effort. Auditors who viewed visualizations of Big Data that disconfirmed management's estimate of revenue and promoted higher levels of arousal (relative to lower levels of arousal) recommended greater decreases to reported revenue and proposed greater increases to budgeted audit hours. With regards to the reliability of the data used to create visualizations, there was no main effect of reliability on auditors' assessments of audit budgeting decisions. While this was unexpected, prior research has found that auditors can fail to appropriately consider data reliability.

Trotman and Wright (2012) find that, instead of consistently considering reliable evidence, auditors only considered more reliable evidence (i.e., externally generated traditional audit evidence) in their fraud assessments when less reliable evidence (i.e., management-controlled evidence) was inconsistent. Similarly, Joyce and Biddle (1981) find evidence to suggest that auditors do not adequately distinguish between the reliability of variance explanations provided by management (less reliable) versus the same explanation provided by a third party (more reliable). These studies suggest that although auditors are charged with considering the reliability of the evidence they evaluate, they are inattentive to some evidence sources and experience difficulties in differentiating and assimilating reliable evidence. Our results suggest a need for increased auditor training around the reliability of data underlying visualizations. However, the experiment results also revealed that visualizations that promoted higher levels of arousal caused auditors to increase their attention to data reliability, suggesting multiple benefits of increased visualization arousal in audit practice.

Overall, our experiments reveal opportunities to enhance audit quality and offer a method for examining the potential effects of different visualization types on auditor judgment prior to deployment of these visualizations in the field. Presenting Big Data using visualizations that promote arousal has the capacity to enhance auditor recognition and use of disconfirming evidence and increase auditor attention to the reliability of new and emerging data sources. In addition, we identify a potentially significant training opportunity for audit firms. Auditors' judgments about misstatement and audit planning were not influenced by data reliability for the visualization that promoted less arousal. This finding indicates a current and meaningful risk to practice.

REFERENCES

Appelbaum, D. 2016. Securing Big Data provenance for auditors: The Big Data provenance black box as reliable evidence. *Journal of Emerging Technologies in Accounting* 13: 17-36.

Asare, Stephen K. and A. Wright. 2003. A Note on the Interdependence between Hypothesis Generation and Information Search in Conducting Analytical Procedures. *Contemporary Accounting Research* 20: 235-251.

Bradley, M. M., L. Miccoli, M. A. Escrig, and P. J. Lang. "The pupil as a measure of emotional arousal and autonomic activation." *Psychophysiology* 45 (2008): 602-607.

Benbasat, I., and A. Dexter. "An investigation of the effectiveness of color and graphical information presentation under varying time constraints." *MIS quarterly* (1986): 59-83.

Brown-Liburd, H., H. Issa, and D. Lombardi. “Behavioral implications of Big Data's impact on audit judgment and decision making and future research directions.” *Accounting Horizons 29* (2015): 451-468.

Chen, Y., Jermias, J., and Panggabean, T. (2016). The Role of Visual Attention in the Managerial Judgment of Balanced-Scorecard Performance Evaluation: Insights from Using an Eye-Tracking Device. *Journal of Accounting Research*, *54*(1) 113-146.

Cloyd, C., and B. Spilker. "The influence of client preferences on tax professionals' search for judicial precedents, subsequent judgments and recommendations." *The Accounting Review* 74 (1999): 299-322.

Day, R. "Examining the validity of the Needleman–Wunsch algorithm in identifying decision strategy with eye-movement data." *Decision Support Systems* 49 (2010): 396-403.

Dechow, P., W. Ge, C. Larson, and R. Sloan. “Predicting material accounting misstatements.” *Contemporary Accounting Research* 28 (2011): 17-82.

Duke University Library. 2018. Data Visualizations. Available at: https://guides.library.duke.edu/datavis/vis_types

Earley, C., V. Hoffman, and J. Joe. "Reducing management’s influence on auditors’ judgments: An experimental investigation of SOX 404 assessments." *The Accounting Review* 83 (2008): 1461-1485.

Fehrenbacher, D. D., Schulz, A. K. D., and Rotaru, K. (2018). The moderating role of decision mode in subjective performance evaluation. *Management Accounting Research, 41*, 1-10.

Franzel, J., J. Rose, J. Thibodeau, and L.T. Williams. “The Audit of the Future: A Guide for Researchers and Regulators.” Unpublished paper, Bentley University, 2018.

Gepp, A., M. Linnenluecke, J. Terrence, and T. Smith. "Big data techniques in auditing research and practice: Current trends and future opportunities", *Journal of Accounting Literature* 40 (2018): 102-115.

Goldinger, S., and M. Papesh. 2012. Pupil Dilation Reflects the Creation and Retrieval of Memories. *Current Directions in Psychological Science*. 21, 2: 90-95.

Griffith, E., K. Kadous, and D. Young. "How insights from the "new" JDM research can improve auditor judgment: Fundamental research questions and methodological advice." *Auditing: A Journal of Practice & Theory* 35 (2015): 1-22.

Hackenbrack, K. and M. Nelson. "Auditors' incentives and their application of financial accounting standards." *The Accounting Review* 71 (1996): 43-59.

Hastie, R. and B. Park. "The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line." *Psychological Review* 93 (1986): 258-268.

Hayes, A. F., and K. J. Preacher. "Statistical mediation analysis with a multicategorical independent variable." *British Journal of Mathematical and Statistical Psychology* 67 (2014): 451-470.

Hayhoe, M., and D. Ballard. "Eye movements in natural behavior." *Trends in cognitive sciences* 9 (2005): 188-194.

Hertwig, R., T. Pachur, and S. Kurzenhäuser. "Judgments of risk frequencies: tests of possible cognitive mechanisms." *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 31 (2005): 621-642.

Hirst, D. E. "Auditors' sensitivity to source reliability." *Journal of Accounting Research* 32 (1994): 113–126.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H. and Van de Weijer, J. (2011). *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford, UK: Oxford University Press.

Holton, C. "Identifying disgruntled employee systems fraud risk through text mining: A simple solution for a multi-billion dollar problem." *Decision Support Systems* 46 (2009): 853-864.

Howells, F. M., D. J. Stein, and V. A. Russell. "Perceived mental effort correlates with changes in tonic arousal during attentional tasks." *Behavioral and Brain Functions* 6 (2010): 39.

Jenkins, J. G., and C. Haynes. "The persuasiveness of client preferences: An investigation of the impact of preference timing and client credibility." *Auditing: A Journal of Practice & Theory* 22 (2003): 143–154.

Joyce, E.J. and G. Biddle. “Are auditors' judgments sufficiently regressive?” *Journal of Accounting Research* 19 (1981): 323-349.

Kadous, K., J. Kennedy, and M. Peecher. "The effect of quality assessment and directional goal commitment on auditors' acceptance of client-preferred accounting methods." *The Accounting Review* 78 (2003): 759-778.

Kadous, K., J. Leiby, and M. Peecher. “How do auditors weight informal contrary advice? The joint influence of advisor social bond and advice justifiability.” *The Accounting Review* 88 (2013): 2061–2087.

Kadous, K., and A. Magro. "The effects of exposure to practice risk on tax professionals' judgements and recommendations." *Contemporary Accounting Research* 18 (2001): 451-475.

Kadous, K., A. Magro, B. Spilker. “Do effects of client preference on accounting professionals' information search and subsequent judgments persist with high practice risk?” *The Accounting Review* 83 (2008): 133-156.

Kahneman, D., and J. Beatty. "Pupil diameter and load on memory." *Science* 154 (1966): 1583-1585.

Keller, P.A. and L. Block. “Vividness effects: A resource-matching perspective.” *Journal of Consumer Research* 24 (1997): 295-304.

Knechel, W. R., and W. Messier. “Sequential auditor decision making: Information search and evidence evaluation.” *Contemporary Accounting Research* 6 (1990): 386–406.

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., and Gowda, S. M. “Standardization of automated analyses of oculomotor fixation and saccadic behaviors.” *IEEE Transactions on Biomedical Engineering*, *57*(2010), 2635-2645.

Kucewicz, M., J. Dolezal, V. Kremen, B. Berry, L. Miller, A. Magee, V. Fabian, and G. Worrell. “Pupil size reflects successful encoding and recall of memory in humans.” *Scientific Reports* 8 (2018): 1-7.

MacLeod, C. and L. Campbell. “Memory accessibility and probability judgments: An experimental evaluation of the availability heuristic.” *Journal of Personality and Social Psychology* 63 (1992): 890-902.

Marquardt, C., and C. Wiedman. “How are earnings managed? An examination of specific accruals.” *Contemporary Accounting Research* 21 (2004): 461-491.

Mathôt, S., and S. Van der Stigchel. "New light on the mind’s eye: The pupillary light response as active vision." *Current directions in psychological science* 24 (2015): 374-378.

McGill, A.L. and P. Anand. “The effect of vivid attributes on the evaluation of alternatives: The role of differential attention and cognitive elaboration.” *Journal of Consumer Research* 16 (1989): 188-196.

Nearon, B. “Foundations in auditing and digital evidence.” *The CPA Journal* 75 (2005): 32-34.

Nuthmann, A., and E. Van Der Meer. "Time's arrow and pupillary response." *Psychophysiology* 42 (2005): 306-317.

Orquin, J., and K. Holmqvist. 2018. Threats to the validity of eye-movement research in psychology. *Behavior Research Methods 50*(4): 1645-1656.

Partala, T., and V. Surakka. "Pupil size variation as an indication of affective processing." *International journal of human-computer studies* 59 (2003): 185-198.

Public Company Accounting Oversight Board (PCAOB). Auditing Standard (AS) No. 1105. Audit Evidence. August 5, 2010a. Available at: https://pcaobus.org/Standards/Auditing/Pages/AS1105.aspx

Public Company Accounting Oversight Board (PCAOB). Auditing Standard (AS) No. 2110. Identifying and Assessing Risks of Material Misstatement. August 5, 2010b. Available at: https://pcaobus.org/Standards/Auditing/Pages/AS2110.aspx

Querino, E., L. dos Santos, G. Ginani, E. Nicolau, D. Miranda, M. Romano-Silva, and L. Malloy-Diniz. “Cognitive effort and pupil dilation in controlled and automatic processes.” *Translational Neuroscience* 6 (2015): 168-173.

Rayner, Keith. "Eye movements and attention in reading, scene perception, and visual search." *The quarterly journal of experimental psychology* 62 (2009): 1457-1506.

Reimers, J. L., and M. Fennema. , M. G. “The audit review process and sensitivity to information source objectivity.” *Auditing: A Journal of Practice & Theory* 18 (1999): 117–123.

Reutskaja, E., R. Nagel, C. F. Camerer, and A. Rangel. "Search dynamics in consumer choice under time pressure: An eye-tracking study." *American Economic Review* 101 (2011): 900-926.

Rose, A., J. Rose, K. Sanderson, and J. Thibodeau. “When should audit firms introduce analyses of Big Data into the audit process?” *Journal of Information Systems* 31 (2017): 81-99.

Rotaru, K., Schulz A.K-D., and Fehrenbacher, D.D. (2018). New Technologies for Behavioral Accounting Experiments, in T. Libby and L. Thorne (eds.), *Routledge Companion to Behavioral Accounting Research*, pp. 253-272, Routledge.

Shanteau, J., “Cognitive heuristics and biases in behavioral auditing: Review, comments and observations.” *Accounting, Organizations and Society* 14 (1989): 165-177.

Shedler, J. and M. Manis. “Can the availability heuristic explain vividness effects?” *Journal of Personality and Social Psychology* 51 (1986): 26-36.

Sirois, S. and J. Brisson. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(6): 679-692.

Stanners, R. F., M. Coulter, A. W. Sweet, and P. Murphy. "The pupillary response as an indicator of arousal and cognition." *Motivation and Emotion* 3(1979): 319-340.

Thayer, J. “Determinants of investors' information acquisition: Credibility and confirmation.” *The Accounting Review* 86 (2001): 1-22.

Tan, H., and K. Jamal. “Do auditors objectively evaluate their subordinates’ work?” *The Accounting Review* 76 (2001): 99–110.

Tobii (2014). Tobii TX300 Eye Tracker User Manual. Tobii Technology. Available at: https://www.tobiipro.com/siteassets/tobii-pro/user-manuals/tobii-pro-tx300-eye-tracker-user-manual.pdf

Trotman, K.T. and W. Wright. “Triangulation of audit evidence in fraud risk assessments.” *Accounting, Organizations and Society* 37 (2012): 41-53.

Van der Wel, P., and H. van Steenbergen. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review *Psychonomic Bulletin & Review*, 1-11.

Van Steenbergen, H., and G. PH Band. "Pupil dilation in the Simon task as a marker of conflict processing." *Frontiers in human neuroscience* 7 (2013): 215.

Vasarhelyi, M., A. Kogan, and B. Tuttle. “Big Data in accounting: An overview.” *Accounting Horizons* 29 (2015): 381-396.

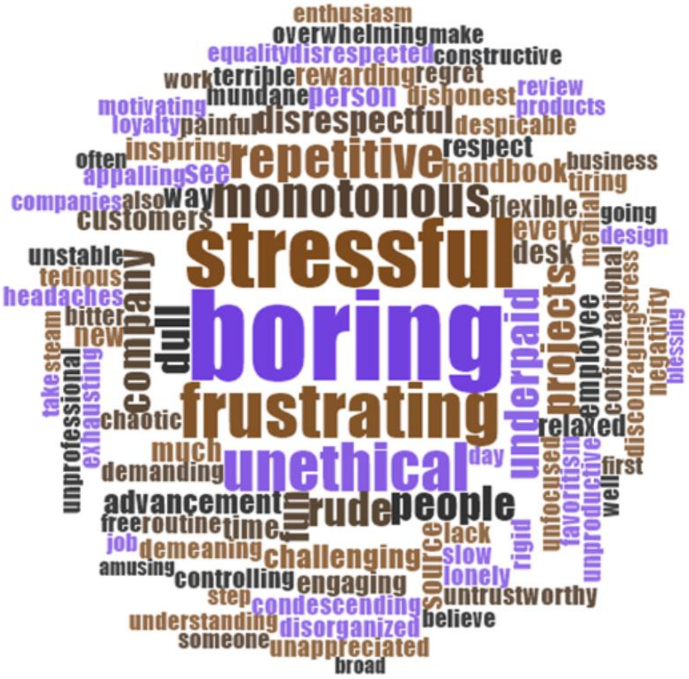
Verney, S. P., E. Granholm, and D. P. Dionisio. "Pupillary responses and processing resources on the visual backward masking task." *Psychophysiology* 38 (2001): 76-83.

Wedel, M., and R. Pieters. "A review of eye-tracking research in marketing." In *Review of marketing research*, pp. 123-147. Emerald Group Publishing Limited, 2008.

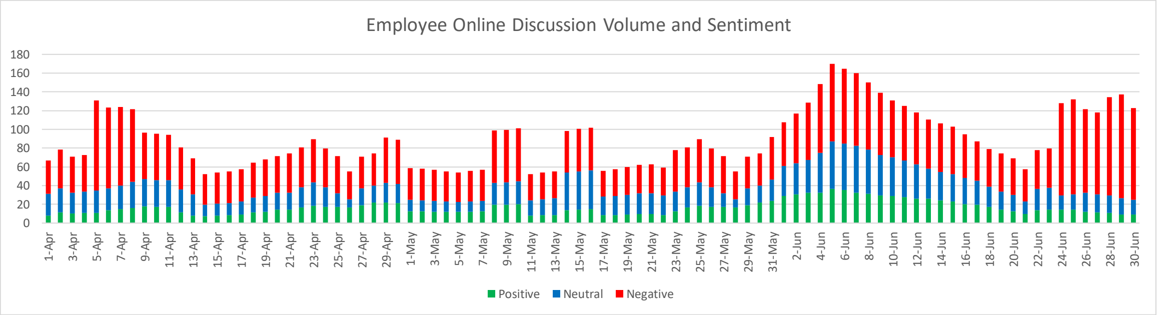
Yoon, K., L. Hoogduin, and L. Zhang. "Big Data as complementary audit evidence." *Accounting Horizons* 29 (2015): 431-438.

Appendix I – Data Visualizations

Word Cloud



Bar Graph



Source	Mentions	Positive	Positive%	Neutral	Neutral%	Negative	Negative%	Net Sentiment
Total	8117	1496	18%	2130	26%	4492	55%	-37%

Appendix II - Technical Specifications of the Eye Tracking and Pupil Measurements[8]

Core parameters	Parameter description	Parameter specifications adopted in the experiment
Apparatus	Sampling procedure	Binocular recording procedure was used (i.e. pupil dilation and eye tracking measures are based on the data acquired from both left and right eyes of the participants)
	Name and produce of the eye tracking device	Tobii TX300, Tobii (Sweden)
	Type of eye tracking device	Desk-mounted
	Sampling rate	300 Hz
	Sampling rate variability	0.3%
	Processing latency	1.0 – 3.3 ms
	Accuracy[9]	0.4^0 – at ideal conditions[10], 0.3^0 - at 25^0 gaze, 0.6^0 - at 30^0 gaze, 0.6^0 – at 1 lux[11], 0.4^0 – at 300 lux, 0.5^0 – at 600 lux, 0.5^0 – at 1000 lux.
	Precision	0.01^0 – with Stamper filter (for more details on the applied Stamper algorithm for noise reduction see Stamper, 1993)
	Eye tracking software used	Tobii Studio 3.4.5
	Chin rest used	Yes

[8] Informed by Orquin and Holmqvist (2018).

[9] The angular average distance from the actual gaze point to the one measured by the eye tracker.

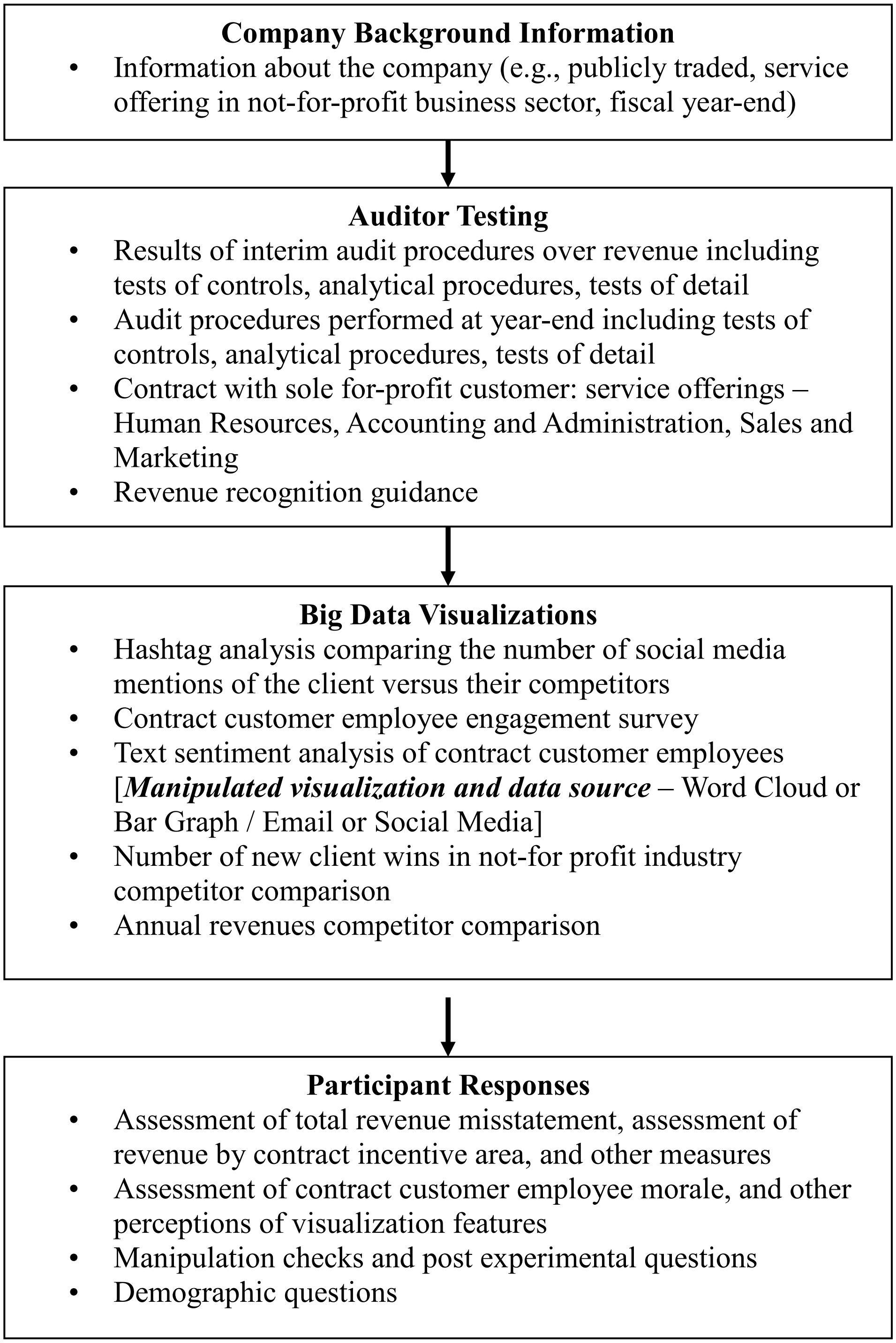
[10] The default experimental setup of this study conforms to the definition of ‘accuracy under ideal conditions’ outlined in Tobii (2014) as follows: (i) the head movement of the participant is fixed in a chinrest; and (ii) data collected immediately after calibration, in a controlled laboratory environment with constant illumination, with 9 stimuli points (related to the 9-point calibration procedure undertaken in this study) at gaze angle $\leq 18^0$.

.[11] Unit of illuminance and luminous emittance, measuring luminous flux per unit area. One *lux* is equal to one lumen per square metre.

Monitor	Screen size	23‴
	Screen resolution	1920 x 1080 pixel
	Distance between participant and screen	Operating distance: 50-80cm Default distance used in this study: 65cm
Calibration	How many points in calibration	9-point calibration
	Amount of recalibration	No recalibration used[12]
Areas of Interest (AOIs)	AOIs used for eye tracking data analysis	Visual representation of sentiment data using stacked bar graph and word cloud illustrated in Appendix I.
Exclusions	Number of trials excluded	None
	Number of participants excluded due to the missing eye tracking data	None
	Data quality threshold	A data quality threshold of 15% was used, i.e. at least 85% of the eye tracking data while attending two AOIs had to be present, otherwise the participant was excluded from the sample.
Event detection	What algorithm is used for event detection	The IV-T fixation filter (Komogortsev et al., 2010) was adopted via the selection of global settings in the eye tracking software (Tobii Studio 3.4.5). A 60ms threshold was selected as part of within IV-T Tobii filter parameters to define fixations.

[12] As the participants used chin rests, and the duration of the study was relatively short (approximately 15 minutes, including the time spent on calibration procedure, on familiarizing with the instructions, on attending the experimental treatments, and on answering the post-experimental questions) no recalibration of the eye tracking devices was required.

Figure 1
Sequence of Experimental Procedures



Company Background Information

- Information about the company (e.g., publicly traded, service offering in not-for-profit business sector, fiscal year-end)

Auditor Testing

- Results of interim audit procedures over revenue including tests of controls, analytical procedures, tests of detail
- Audit procedures performed at year-end including tests of controls, analytical procedures, tests of detail
- Contract with sole for-profit customer: service offerings – Human Resources, Accounting and Administration, Sales and Marketing
- Revenue recognition guidance

Big Data Visualizations

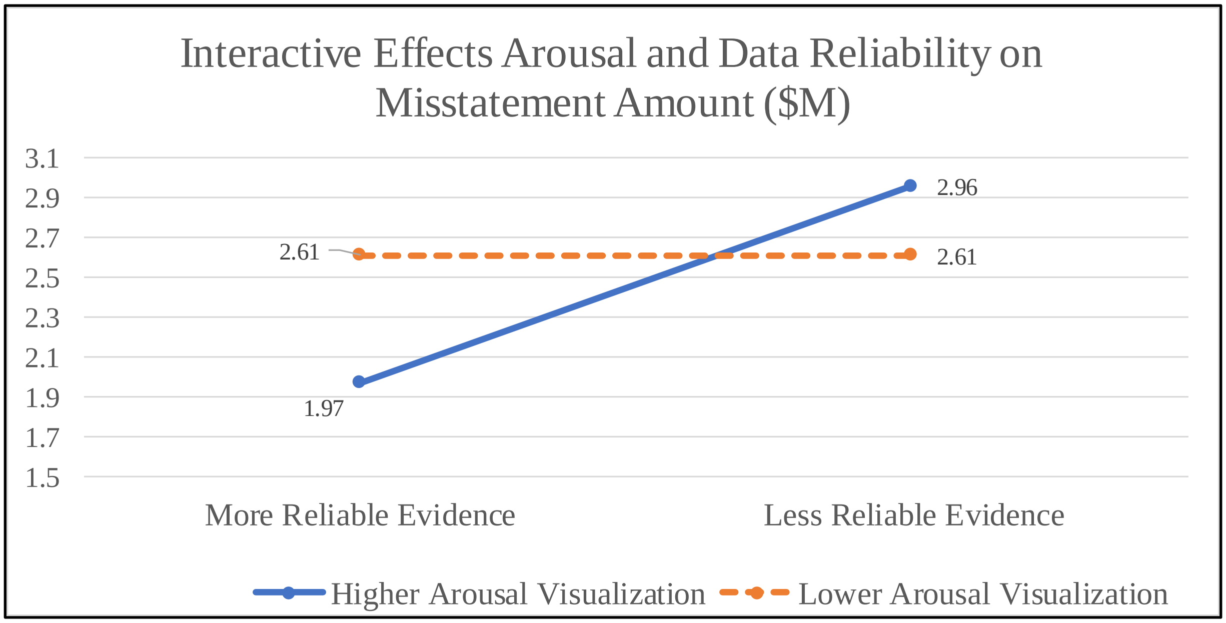
- Hashtag analysis comparing the number of social media mentions of the client versus their competitors
- Contract customer employee engagement survey
- Text sentiment analysis of contract customer employees [***Manipulated visualization and data source*** – Word Cloud or Bar Graph / Email or Social Media]
- Number of new client wins in not-for profit industry competitor comparison
- Annual revenues competitor comparison

Participant Responses

- Assessment of total revenue misstatement, assessment of revenue by contract incentive area, and other measures
- Assessment of contract customer employee morale, and other perceptions of visualization features
- Manipulation checks and post experimental questions
- Demographic questions

Figure 2

Experiment 1: Interactive Effects Arousal and Data Reliability on Misstatement Amount



Variable Definitions:

- Misstatement Amount = Amount auditors suggested should be recorded for Human Resource incentive bonus revenue. Participant response to the statement, "Assuming that you need to make a recommendation to the engagement partner regarding any need to adjust the client's revenue from the new for-profit contract with [Customer Name], what amount of revenue would you recommend to the partner as the appropriate amount to be recorded for the Human Resources performance incentive?" Participants were instructed to indicate an amount between $0 to $4 million.
- Arousal = Word Cloud visualization (*Higher Arousal*), Bar Graph visualization (*Lower Arousal*).
- Reliability = Reliability of source underlying Big Data visualization; Company email (*More Reliable*), Social media (*Less Reliable*).
- Arousal*Reliability = Interaction of *Arousal* and *Reliability* independent variables.

TABLE 1

Experiment 1: Descriptive Statistics and ANCOVA for Pupillometry Measures

Panel A: Pupillary Response Ratio (Standard Deviation) [Number of Participants]

Visualization Type	**Pupillary Response Ratio**
Word Cloud	1.023 (0.022) [74]
Bar Graph	0.979 (0.019) [74]

Panel B: Within-Subjects Contrasts for Repeated Measures ANCOVA (DV = Pupillary Response Ratio)

Factor	df	Mean Square	F-value	p-value[a]
Visualization Type	1	0.073	120.521	< 0.001
Visualization Type x Order	1	0.015	25.262	< 0.001
Error	72	0.001		

[a] p values are two-tailed

Panel C: Between-Subjects Effects of Order ANCOVA (DV = Pupillary Response Ratio)

Source	df	Mean Square	F-value	p-value[a]
Order	1	2.82E-05	0.969	0.328
Error	72	2.91E-05		

[a] p values are two-tailed

TABLE 2
Experiment 1: ANCOVA for Eye Tracking Measures

Panel A: Within-Subjects Contrasts for Repeated Measures ANCOVA (DV = Fixation Count)

Factor	df	Mean Square	F-value	p-value[a]
Visualization Type	1	941.84	1.022	0.315
Visualization Type x Order	1	385.95	0.419	0.520
Error	72	921.61		

[a] p values are two-tailed

Panel B: Within-Subjects Contrasts for Repeated Measures ANCOVA (DV = Fixation Duration)

Factor	df	Mean Square	F-value	p-value[a]
Visualization Type	1	39.87	0.948	0.334
Visualization Type x Order	1	63.39	1.507	0.224
Error	72	42.07		

[a] p values are two-tailed

TABLE 3
Experiment 2: Auditor Participant Demographics (n=120)

Demographic	**Mean** **(Standard Deviation)**
Gender: Male	59%
Years Audit Experience	2.7 (1.2)
Years Analytical Review Experience	2.0 (1.4)

TABLE 4
Experiment 2: Descriptive Statistics and ANCOVA for Misstatement Amount

Panel A: Mean in Millions $ (Standard Deviation) [Number of Participants]

	Higher Arousal Visualization	Lower Arousal Visualization	Total
More Reliable Evidence	1.97 (1.57) [28]	2.61 (1.55) [29]	2.29 (1.58) [57]
Less Reliable Evidence	2.96 (1.35) [33]	2.61 (1.42) [30]	2.79 (1.38) [63]
Total	2.51 (1.52) [61]	2.61 (1.47) [59]	2.56 (1.49) [120]

Panel B: ANCOVA Results for Misstatement Amount

Factor	Df	F-value	p-value[a]
Arousal	1	0.397	0.530
Reliability	1	3.733	0.056
Arousal*Reliability	1	3.690	0.057
Audit Experience	1	2.313	0.131
Error	115		

[a] p values are two-tailed

Panel C: Results for Simple Effects of Misstatement Amount

Factor	More Reliable	Less Reliable	t-value	p-value [a]
Higher Arousal	1.97	2.96	2.621	0.010
Lower Arousal	2.61	2.61	0.242	0.810

[a] p values are two-tailed

Variable Definitions:

- Misstatement Amount = Amount auditors suggested should be recorded for Human Resource incentive bonus revenue. Participant response to the statement, "Assuming that you need to make a recommendation to the engagement partner regarding any need to adjust the client's revenue from the new for-profit contract with [Customer Name], what amount of revenue would you recommend to the partner as the appropriate amount to be recorded for the Human Resources performance incentive?" Participants were instructed to indicate an amount between $0 to $4 million. Values shown in $millions, rounded to two decimal places.
- Arousal = Arousal of Big Data disconfirming evidence presentation; Word Cloud visualization (*Higher Arousal*), Stacked Bar Graph visualization (*Lower Arousal*).
- Reliability = Reliability of source underlying Big Data visualization; Company email (*More Reliable*), Social media (*Less Reliable*).
- Arousal*Reliability = Interaction of *Arousal* and *Reliability* independent variables.
- Audit Experience = Years of audit experience.

TABLE 5
Experiment 2: Descriptive Statistics and ANCOVA for Auditor Assessment of Budgeted Audit Hours

Panel A: Mean (Standard Deviation) [Number of Participants]

	Higher Arousal Visualization	Lower Arousal Visualization	Total
More Reliable Evidence	142.36 (43.81) [28]	130.95 (24.43) [29]	136.55 (35.45) [57]
Less Reliable Evidence	137.88 (30.47) [33]	127.67 (33.08) [30]	133.02 (31.90) [63]
Total	139.93 (36.93) [61]	129.28 (28.95) [59]	134.70 (33.54) [120]

Panel B: ANCOVA Results for Auditor Assessment of Budgeted Audit Hours

Factor	Df	F-value	p-value[a]
Arousal	1	4.614	0.034
Reliability	1	0.829	0.365
Arousal*Reliability	1	0.088	0.767
Audit Experience	1	16.611	< 0.001
Error	115		

[a] p values are two-tailed

Variable Definitions:

- Budgeted Audit Hours = Assessment of that audit hours budgeted to perform revenue substantive testing. Participant response to the statement, “Indicate how many hours you want to budget this year for substantive testing of the revenue account.”
- Arousal = Arousal of Big Data disconfirming evidence presentation; Word Cloud visualization (*Higher Arousal*), Stacked Bar Graph visualization (*Lower Arousal*).
- Reliability = Reliability of source underlying Big Data visualization; Company email (*More Reliable*), Social media (*Less Reliable*).
- Arousal*Reliability = Interaction of *Arousal* and *Reliability* independent variables.
- Audit Experience = Years of audit experience.