How Do Inspectors Evaluate Auditors’ Decisions? The Role of Prior Experience, Inspection Approach, and Wise Thinking

TIM BROWN, Assistant Professor
University of Illinois at Urbana-Champaign
tjb@illinois.edu

TRACIE M. MAJORS, Assistant Professor
University of Southern California
tmajors@marshall.usc.edu

MARK E. PEECHER, Deloitte Professor of Accountancy
University of Illinois at Urbana-Champaign
peecher@illinois.edu

December 2016

We thank Cassandra Estep and Sean Hillison for research assistance. We also thank participants from the 2016 Accounting, Behavior and Organizations conference and Donnie Young (discussant), as well as workshop participants at Nanyang Technological University and University of Southern California, for providing valuable feedback.
How Do Inspectors Evaluate Auditors’ Decisions? The Role of Prior Experience, Inspection Approach, and Wise Thinking

Abstract:

Research largely reports evidence of PCAOB inspections being useful (i.e., for investors and audit clients), yet auditors generally report dissatisfaction with the process. This study aims to shed insight reconciling these two findings, by examining how inspectors’ prior experience biases their evaluations of auditor judgment. We conduct an experiment in which participants design, and one week later, evaluate, an audit program. We manipulate relevance of experience by whether participants design an audit program for an account that is same as, or different from, the account they later evaluate. We also manipulate whether evaluations are framed under an inspection or reasonableness approach, and measure evaluators’ inherent propensity for wise thinking. Results show that evaluators with more relevant experience judge an audit program that differs from their own experience harshly, relative to evaluators with less relevant experience. A reasonableness approach mitigates this tendency in low wise thinkers, but instead amplifies bias in high wise thinkers. This study complements archival studies examining PCAOB inspections, as those studies typically rely on the content of inspection reports (i.e., the output), whereas our experimental approach enables us to examine the process of inspections. The PCAOB and audit firms should be interested in these results, as while excessive criticism is unlikely to manifest in ultimate inspection reports, it can pose the costs of diverting attention from the audit itself, as well as harm auditor morale.

Keywords: PCAOB inspections; wise thinking; attributes; prior experience; audit procedure choice
I. INTRODUCTION

The Public Company Accounting Oversight Board (PCAOB) is a regulator responsible for overseeing the conduct of auditors, created by the U.S. Congress in the aftermath of a series of financial reporting scandals and as a result of the Sarbanes-Oxley Act. The PCAOB conducts inspections of audit firms, in which inspectors review audit programs and procedures performed to ensure that audits are performed in accordance with PCAOB standards. While some studies cast doubt on PCAOB inspection reports being informative (Lennox and Pittman 2010), the majority of (later) studies show that investors (Robertson and Houston 2010; Gipper, Leuz, and Maffett 2015) and audit clients (Daugherty, Dickins, and Tervo 2011; Abbott, Gunny, and Zhang 2013; Robertson, Stefaniak, and Houston 2014; Aobdia 2016; Aobdia and Shroff 2016) appear to incorporate the content of PCAOB inspection reports in their decisions. Further, most studies find that audit quality has improved as a result of the inspection process (Carcello, Hollingsworth, and Mastrolia 2011; Gramling, Krishnan, and Zhang 2011; DeFond and Lennox 2015; Lamoreaux 2016; Aobdia 2016). It is puzzling, then, that auditors seem to generally report dissatisfaction with the inspection process (Houston and Stefaniak 2013; Johnson, Keune, and Winchel 2016). For example, if investors and client managers generally view reported deficiencies as informative, and consequential enough to affect their decisions, then why would auditors contend that inspectors are excessively critical? We develop and test theory in an experiment that inspectors’ previous experience is one factor that may help to solve this puzzle.

Because the PCAOB is charged with holding auditors of public companies to a standard of high quality, it often hires experienced auditors to serve as inspectors (PCAOB 2014, 2016). Inspectors’ experience provides relevant knowledge allowing them to understand how audit procedures provide assurance, and make critical determinations about audit effectiveness. Nonetheless, we propose that this previous experience might also affect how inspectors view an
auditor’s procedures, and therefore, the communications between inspectors and auditors. Individuals naturally process information based on their experience (Libby & Luft 1993), as it provides a framework to understand complex tasks like auditing. We posit that, if an auditor performs procedures that are different from an inspector’s experience, the inspector may believe that they are less effective than procedures that match their experience, simply by virtue of being different. This is likely to be true even if the alternative procedures provide sufficient assurance, as psychology research shows that the biasing effect of experience is very difficult to overcome (Dane 2010). If inspectors are excessively critical (e.g., pointing out trivial problems), these criticisms may manifest in informal communications between inspectors and audit team members – or even if documented in the inspector’s report, are likely to be “negotiated away” during negotiations between audit partners and the PCAOB. Accordingly, archival studies examining the content of inspection reports are unlikely to find evidence of such excessive criticism. Our study aims to fill this gap by showing evidence about what is likely going on during “behind the scenes” during inspections, which is important because time spent on (potentially) trivial, or excessively critical, deficiencies can impose meaningful costs, such as diverting auditors’ attention from performing the audit and harming auditors’ morale.

We also examine two moderating factors that likely influence the effect of experience: the framing of the approach the inspector is asked to employ and the inspector’s inherent propensity for wise thinking. Although PCAOB inspectors are trained to detect issues and ensure compliance during their inspections (PCAOB 2014), i.e., an “inspection approach,” they could also adopt a reasonableness approach, recognizing that differences in professional judgment are natural and do not necessarily indicate a deficiency in judgment. We note that both the professional community (Treasury 2008, Turley 2007) and the research community (Peecher,
Solomon and Trotman 2013) have explored how this type of change (i.e., a “reasonableness approach”) might improve auditor effectiveness. The inspector’s tendency to engage in “wise thinking” may also alter the effects of experience. Wise thinkers are more reflective, and tend to more deeply consider the tasks assigned to them (Ardelt 2004, Meacham 1983; 1990). We predict that these factors will jointly change the effect of experience, such that propensity for engaging in wise thinking will influence how a reasonableness approach is interpreted.

We test our hypotheses in a $2 \times 2 \times 2$ mixed experiment conducted over two stages. In the first stage, all participants are asked to design an audit program for a specific account – either Accounts Payable or Inventory. In the second stage (one week later), all participants are asked to evaluate an audit program for Inventory. This audit program is customized, such that participants who designed an audit program over Inventory see a program that is relatively inconsistent with the program that they designed in the first stage (but still achieves assurance over the highest risk assertion for the account). This is our condition of more relevant experience, as these participants designed procedures for the same account in the previous stage.

By contrast, the other half of participants (those who designed an audit program for Accounts Payable in the first stage) are in our control condition, which represents evaluating an audit program with less relevant experience. Each of these participants is yoked with a participant in the other condition, such that s/he views the same program that we customized for the other matched participant. However, participants in this condition have less relevant experience because they designed an audit program for a different account in the first stage. These subjects serve as a control condition, to investigate how the same audit programs are evaluated with less relevant experience (i.e., to provide an unbiased benchmark of program quality). It is important to note that all our participants were asked to design audit procedures, and thus possess
equivalent levels of experience – we simply manipulate if that experience is more versus less relevant to the procedures they are asked to evaluate in the second stage.

We also alter the framing of the evaluation approach participants are asked to employ during the second stage. Half are tasked with an “inspection” approach, which informs them that their review is an important check designed to detect deficiencies in judgments. The other half are tasked with a “reasonableness” approach, which informs them that disagreements are natural and should be considered against a reasonableness standard. Our primary dependent variable is participants’ evaluation of the judgment proficiency of the individual who selected the audit procedures in the audit program they are evaluating. Participants complete the experiment by answering a number of debriefing questions, including a measure of their propensity for wise thinking. We use this measure as our third experimental variable, separating subjects into two groups based on their responses.

Our results show a significant effect of relevance of prior experience, such that participants with more relevant prior experience—who are presented with procedures fairly different from what they designed—provide a much more negative judgment than participants who did not possess directly relevant experience (despite seeing the same sets of audit procedures). The comparison between participants in the two conditions shows strong evidence for the biasing effect of more relevant experience on inspector evaluations.

A more detailed analysis of the results shows an interactive effect. Participants who are less wise thinking show less bias related to their prior experience when they are presented with the reasonableness approach, compared to the inspection approach. Framing the evaluation in this way likely causes these participants to consider how reasonable people may disagree, which results in less negative judgments. In contrast, participants who are more wise thinking show
more bias related to their prior experience when they are presented with the reasonableness approach. We believe that these participants carefully considered the meaning of the approach, and determined that their own experience was the best indicator about which procedures were reasonable. Process testing shows that these effects are all mediated by evaluators’ perceptions of the competence of the person who designed the audit program, which provides additional evidence that our participants use the cues provided to them to assess factors related to the preparer (i.e., the preparer’s competence, which ultimately impacts their assessments of the preparer’s judgment proficiency), rather than attributing the preparer’s performance to external factors. In total, the results show that more relevant experience can bias assessments of another person’s work, demonstrating that important internal and external factors can alter how this bias influences judgments.

This study contributes to auditing literature and practice by identifying a likely contributing factor to auditor criticisms about the PCAOB inspections process. More specifically, our results highlight that, since inspectors invariably possess significant knowledge about how audits are performed (and typically have audit experience themselves), they may exhibit a bias against alternative methods that are different from what they have done in the past, even if they still achieve sufficient assurance. This bias may cause inspectors to be excessively critical in their interactions with auditors during the inspection process, which can incur significant costs, such as reducing auditor morale, reducing auditor satisfaction with the inspection process, or even diverting limited auditor resources from the audit itself to the inspection process. By providing new evidence on the likely process, or “behind the scenes” perspective, of inspections, our study complements research examining auditors’ perceptions of the inspection process (Houston and Stefaniak 2013; Johnson et al. 2016), as well as research examining the output of the inspection
process via the content of PCAOB inspection reports (see Abernathy, Barnes, and Stefaniak 2013 for a review).

This study’s results also illustrate a problem that auditors should consider when developing new auditing procedures. As technology evolves, innovative procedures can produce more effective and efficient audits (Peecher, Solomon and Trotman 2013). Nonetheless, our results suggest that these new procedures might be judged especially harshly if they do not fit the prior experience of inspectors, and that steps should be taken to ensure that new procedures are properly evaluated by outside inspectors. In light of our findings, the PCAOB could consider assigning inspectors to areas where they have less relevant experience (but still requisite knowledge), and can therefore be less biased in their evaluations – or alter the approach inspectors are asked to take, as an alternative way to mitigate this bias. Finally, this study’s results speak to the broader literature on reviews in auditing. Reviewers in audit settings, such as those performing AICPA peer reviews, review of other teams within the firm (such as concurring review partners), or superiors reviewing a subordinate’s work, are all likely approaching the task through the lens of their prior experience. Accordingly, practitioners may want to take proactive steps to ensure that previous experience does not influence outcomes in an unintended way.

In Section II, we provide background for our study and develop theory and hypotheses. In Section III, we describe our experiment and the method by which we test our hypotheses. In Section IV, we provide the results of our experiment. We conclude in Section V.

II. THEORY AND HYPOTHESES DEVELOPMENT

The PCAOB Inspection Process

Although auditing has historically been a self-regulating industry, the Sarbanes-Oxley Act of 2002 (SOX, U.S. House of Representatives) drastically changed the regulatory landscape. More specifically, Section 104 of SOX requires the Public Company Accounting Oversight Board
(“PCAOB”) to conduct inspections of all registered firms. The inspection process requires the PCAOB to evaluate the quality control systems of audit firms, perform tests of quality control procedures, and perform specific inspections of audit and review engagements (PCAOB 2004). This study focuses on the engagement inspection process, in which an inspector evaluates audit procedures performed after the fact to ensure that they achieved sufficient assurance. The goal of this inspection process is to ensure that auditors comply with PCAOB standards, and to improve overall audit quality (PCAOB 2014).

While some research suggests that PCAOB inspection reports are not very informative (Glover et al. 2009; Lennox and Pittman 2010), the majority of studies report evidence instead supporting the efficacy of the inspection process, and that inspection reports are considered to be informative (see Abernathy, Barnes and Stefaniak 2013). More specifically, archival studies have identified market penalties (i.e., dismissals from audit clients) for audit firms as a function of PCAOB identified deficiencies (Abbott et al. 2013; Daugherty et al. 2011).¹ In an experiment with corporate executives, however, Robertson et al. (2014) find that inspection reports, generally (i.e., irrespective of content), appear to reduce perceived audit quality and increase likelihood of auditor switching. Aobdia and Shroff (2016) examine audit firms globally, and find that audit firms inspected by the PCAOB (as part of the international inspections program) gain client market share after inspection reports are made public, but audit firms with high levels of deficiencies do not, suggesting that the inspections are informative. In further support, Gipper et al. (2015) report archival evidence that the PCAOB inspection process causes investors to view financial reporting as more credible, and Robertson and Houston (2010) find in an experiment

---

¹ Interestingly, however, Johnson et al. (2016) report evidence from interviews with auditors that some clients seem to question the value of the inspection process (i.e., that they question the value for the audit of incremental procedures required to comply with the PCAOB).
that investors believe PCAOB inspections improve the credibility of future audit opinions (especially for smaller audit firms, and when inspection reports contain high deficiencies and firms respond by conceding). Research also suggests that audit quality has improved over time as a result of the inspection process (Carcello et al. 2011; Gramling et al. 2011; Gunny and Zhang 2012; DeFond and Lennox 2015; Lamoreaux 2016). Finally, using a proprietary PCAOB data set (i.e., in which issuer clients’ identities and audit hours per issuer can be identified), Aobdia (2016) finds that identified deficiencies in inspection reports tend to increase auditor effort and increase clients’ likelihood of switching auditors. Overall, while there is some mixed evidence, it appears that investors and audit clients view PCAOB inspections as informative, and that inspection process improves audit quality.

Given these findings, it is puzzling that research eliciting views of audit firms (as well opinions generally expressed in the popular press) seem to suggest that auditors are dissatisfied with the inspection process. Houston and Stefaniak (2013) survey 107 audit partners from large accounting firms, asking them to compare the PCAOB inspection process to their firm’s own internal quality review. The authors find a low overall level of satisfaction with the inspection process, and that auditors perceive that it lacks procedural fairness. Auditors in the study report that inspectors lack a fundamental understanding of the audit firm’s methodology, and view the inspection process as more focused on finding deficiencies than improving audit quality. Johnson et al. (2016) conduct semi-structured interviews of auditors about the PCAOB inspection process. They find that the majority of auditors they interview, while acknowledging some benefits of PCAOB oversight, indicate doubt that the additional procedures they perform to comply with the PCAOB are valuable, and also describe unfavorable perceptions about the purpose and value of the process (e.g., an excessive focus on documentation; excessive
questioning of professional judgments). Further, the auditors seemed to sense a disconnect, whereby the detailed nature of the inspections seemed inconsistent with the purpose of regulatory oversight (Johnson et al. 2016). For example, one quoted auditor indicated that:

*I have friends that work at other annually-inspected firms, and I will say that they are equally annoyed with, and frustrated with, the PCAOB...that the common theme is that their inspections as of late have been very unreasonable and very, very, very detailed. They seem to have forgotten that part of executing audits requires judgment and sampling and the evaluation of a reasonable outcome and a reasonable financial statement presentation and disclosure.*

We believe that psychological factors related to the inspectors themselves may play an important role in reconciling these seemingly inconsistent findings (i.e., that, despite the documented benefits of PCAOB inspections, auditors appear to report dissatisfaction). As explained next, we posit that inspectors’ prior experience predisposes them to be (potentially excessively) critical when evaluating the quality of auditors’ work.

**The Role of Experience**

To ensure that inspections are completed competently, the PCAOB hires inspectors with significant experience within the audit industry. A recent post on the PCAOB careers website requests a candidate with “Six to fifteen years of progressively responsible experience having attained the Senior Manager or Manager level in the audit of companies traded on the U.S. markets” (PCAOB 2016). This experience requirement is undoubtedly important, as previous research shows that direct experience with a task increases performance, particularly when the task is complex (see, e.g., Bonner 1990 and Abdolmohammadi and Wright 1987) and it closely matches prior experience (Haerem and Rau 2007). When considering how inspections are

---

2 Consistent with this auditor’s observation of the importance of considering reasonableness, we discuss in a later section our theory about how adopting a reasonableness approach, instead of an inspection approach focused on finding problems, could improve the inspection process.
performed, the knowledge and experience that inspectors possess should allow them to more fully understand the audit process and how to evaluate it during an inspection (Libby and Luft 1993; Nelson and Tan 2005).

Notwithstanding the benefits, experience can have unintended consequences when individuals are asked to evaluate the performance of other people who may perform a task differently. That is, experience carries a trade-off, in which individuals are highly effective in performing their tasks, but less able to adapt to new conditions within the given domain (Dane 2010; Canas, Quesada, Antoli, and Fajardo 2003; Sternberg and Frensch 1992). This trade-off is related to how experience changes the way a task is viewed (Dane 2010). Experience creates expectations about what actions lead to success (Allen and Marquis 1964; Littlepage, Robison and Reddington 1997), and individuals with these expectations have difficulty changing their view on how the task should be performed. In the inspection setting, these expectations should include specific audit techniques that proved successful during inspectors’ audit careers (Libby and Luft 1993).

But audits are complex, and auditors often have considerable flexibility in determining the appropriate mix of substantive tests to achieve assurance over an account (Messier, Glover, and Prawitt 2013). When inspectors possess relevant experience when reviewing a set of audit procedures, it could cause them to react negatively to procedures that do not match what they did in the past. Their prior experience may bias them against doing things a different way, even if the different procedures provide a reasonable level of assurance. Johnson et al. (2016) find evidence consistent with this premise in their semi-structured interviews of auditors. As noted by one auditor interviewed by Johnson et al. (2016):

…I felt like they [the inspectors] had their agenda, and they had probably already made up their minds before the fieldwork even began. It was almost like they knew what they
didn’t like, but they needed to dig into the workpapers, and have the conversations with us to be able to better articulate and have a better rationale for why they didn’t like those things...I felt like they just dug in, and they weren’t going to listen.

Importantly, this effect may be especially important as the audit environment becomes more complex, and different types of procedures are performed (Peecher, Solomon, and Trotman 2013), such as “big data,” predictive analytics, or other innovative approaches. Simple “differences of opinion” between inspectors and auditors may be unlikely to materialize as deficiencies in published inspection reports, but these interactions can impose significant costs, such as reductions in auditor morale and diverting resources from the audit itself to the inspection process, which could potentially harm audit quality. Further, archival research examining the efficacy of the inspection process is unable to observe this element of the process of inspections (versus the observable output, in inspection reports), enabling our experimental approach to provide novel insight, and complement findings from survey and interview-based research on the inspection process (Houston and Stefaniak 2013; Johnson et al. 2016).

Formally, we predict the following hypothesis:

**H1:** Inspectors with more relevant prior experience will evaluate the judgment of a preparer of an audit program (that is inconsistent with the inspector’s judgments during their prior experience) as more deficient than inspectors with less relevant prior experience.

**The Inspection Task & Inspector Characteristics**

We also investigate how this effect changes based on the way the inspector approaches the inspection task and the inherent characteristics of the inspector. The PCAOB conducts inspections “for the purpose of assessing compliance with certain laws, rules, and professional standards in connection with a firm's audit work” (PCAOB 2014). A review of a recent Staff Inspection Brief (PCAOB 2015) reveals a focus on deficiencies, in which inspectors document instances in which auditors failed to perform sufficient testing or did not document procedures performed. We believe that these types of comments reflect a penalty frame, where the
inspector’s role is to search for problems and bring them to light, and is consistent with the approach occurring during the inspection process that auditors in Houston and Stefaniak (2013) describe. We call this the “inspection” approach – focusing purely on compliance with standards and seeing the inspection process as a way to uncover flawed judgments. The auditors interviewed by Johnson et al. (2016) seem to view inspectors as taking this approach.

But this type of approach may not always be appropriate. The audit task is complex, and individuals may make different reasonable decisions under identical circumstances (Turley 2007, Treasury 2008; Peecher et al. 2013). Adopting a “reasonableness” approach, in which differences in professional judgement are recognized and accepted, rather than the inspection approach, could produce better outcomes. This approach could be particularly important as audit practice evolves, and new, more innovative procedures become available (EY 2015). But this leaves an open question that we also seek to answer in this study: how might the characteristics of the inspector influence how they conduct their task?

Previous research within accounting has established that inherent attributes can have a significant impact on individuals’ decisions (Majors 2015; Murphy 2012; Hales, Hobson and Resutek 2012; Brewster, Peecher, and Solomon 2016). We focus on an individuals’ propensity for wise thinking, which is defined as “a disposition to use reflective thinking that balances the acquisition of new knowledge with the discovery of new uncertainties” (Ardelt 2004, Meacham 1983; 1990). A propensity for wise thinking makes an individual more receptive to new evidence (Baron 2008), less influenced by preexisting beliefs (Stanovich and West 2008), and more willing to adjust their approach based on the demands of the task (Kunzman and Bates 2005).

Within accounting research, recent research has established that wise thinking results in more sensitivity to audit evidence (as well as being less likely to heuristically respond to skepticism
prompts), and is related but distinct from skepticism (Brewster et al. 2016). We propose that inspectors who are more prone to engage in wise thinking will approach the inspection task differently from those who are less prone to engage in wise thinking. High wise thinking inspectors will carefully consider how to evaluate the procedures performed by others and engage in more reflection before providing a judgment. Bringing these concepts together, we make an interactive prediction:

**H2:** The effect of the relevance of prior experience on an inspector’s evaluation of a preparer’s judgment will depend on both the framing of the evaluation approach and the inspector’s propensity for wise-thinking.

More specifically, when inspectors take an “inspection” approach, they are asked to focus on compliance with standards and uncovering flawed judgments. We believe that low wise thinking inspectors will allow their previous experience to bias their judgments in this case and provide unfavorable judgments about procedures that are different from what they performed. However, we propose that high wise thinking inspectors will carefully consider the inspection task, and naturally come to the conclusion that reasonable individuals can disagree as a result of their tendency to engage in reflective thinking. This effect should reduce the biasing effect of experience when performing an inspection. Stated in hypothesis form:

**H2a:** When the evaluation is framed as an “inspection standard,” the relationship between the relevance of prior experience and evaluations of judgment deficiency will be more pronounced for inspectors that are less wise-thinking.

However, when the inspector takes a “reasonableness” approach, in which differences in professional judgement are recognized and accepted, we expect a different result. If inspectors are asked to adopt a reasonableness approach, it could be interpreted as a cue to avoid “myside bias,” which is a tendency to evaluate evidence or test hypotheses in a way that is biased towards one’s own opinion (Baron 2008). Previous research has shown that this bias is prevalent in a wide variety of contexts (see, e.g. Nussbaum & Kardash 2005; Toplak and Stanovich 2003;
Stanovich and West (2007). We expect that inspectors will hold an opinion that the procedures they selected were very appropriate, and examine how wise thinking might interact with a prompt according to this theory. More specifically, Macpherson and Stanovich (2007) ask participants to engage in a logical reasoning task and provide participants with instructions that are designed to reduce myside bias. These authors find that individuals with higher cognitive ability react less to the instructions, and do not reduce the amount of bias shown, compared with participants with lower cognitive ability. This experiment indicates that individuals with different abilities or attributes might react to prompts very differently.

In our experiment, low wise thinking inspectors are likely to see the reasonableness prompt as a cue to be lenient in their judgments, which could remove the bias associated with prior experience. In contrast, high wise thinking inspectors are likely to carefully consider what “reasonable” really means when performing their inspection. In line with Macpherson and Stanovich (2007), high wise thinking inspectors are more likely to deeply process the cue and carefully consider what a reasonable set of audit procedures should look like, rather than simply taking it as a cue for more leniency (i.e. less myside bias). After processing the cue, these high wisdom inspectors might decide that their prior experience is their best available indicator of what a reasonable procedure really is. This expectation is also consistent with findings from Brewster et al. (2016), who show that high wise thinking auditors are less likely to heuristically change risk assessments in accordance with a skepticism prompt. In total, our theory predicts that a reasonableness prompt would increase the biasing effect of experience for high wise thinking inspectors, as they carefully process the cue and determine that their own experience is their best, most relevant example of reasonable procedures. In hypothesis form:
H2b: When the evaluation is framed as a “reasonableness standard,” the relationship between the relevance of prior experience and evaluated judgment deficiency will be more pronounced for inspectors that are more wise-thinking.

In total, we expect that the effects of framing a procedure with different guidance will vary based on an inspector’s inherent propensity toward wise thinking. Reflecting on the true purpose of the task, and the real meaning of the frame, will drastically change the biasing effect of previous experience.

III. METHOD AND DESIGN

Overview and Participants

We test our predictions in a two-stage experiment with Masters of Accounting Students from a large public university (n = 105). All participants had completed one auditing course and were currently enrolled in (and almost completed) a second, graduate-level auditing course. 91 percent of participants had completed an accounting internship. All participants complete both stages of the experiment. In Stage 1, participants design an audit program to audit a specific account. In Stage 2 (one week later), participants view another audit program and are asked to evaluate the judgment proficiency of the preparer of the audit program. As explained in greater detail shortly, this experimental structure enables us to examine the impact of participants’ experience on their judgments in the evaluation task (which they perform in Stage 2), by manipulating how closely their experience relates to the evaluation task (in Stage 1).

We utilize Masters of Accounting students for several reasons. First, our intent is to test theory about the effect of prior experience and evaluation approach (manipulated variables), as well as wise-thinking (an inherent attribute), on inspectors’ evaluations of preparer judgment deficiency. As noted by Peecher and Solomon (2001), students are an appropriate proxy for professionals when it is unlikely that there is a factor varying between students and professionals that would interact with the study’s findings. As we carefully match both tasks in Stage 1 and
Stage 2 to our participants’ level of auditing knowledge (and the individual attribute of wise-thinking that we examine is an inherent trait possessed by students and professionals alike), we do not expect that professionals’ experience would change the nature of our expected results. Our manipulation of relevance of experience is likely to be less powerful than the effects of experience in the inspection setting, and represents a relatively conservative test of our effects. Second, we note that practicing auditors would actually be less suitable proxies for inspectors, due to the experiences and pre-conceptions that they would bring about the inspection process. Students provide a subject pool without these pre-conceptions, allowing us to maximize internal validity by isolating our construct of interest, while also economizing on the use of more experienced professionals (see Libby, Bloomfield, and Nelson 2002).

**Experiment Sequence of Events**

**Stage 1 – procedure selection task.** During Stage 1, participants are tasked with designing an audit program of procedures over a specific account (as explained shortly in the manipulations subsection, inventory or accounts payable depending on experimental condition). Participants view a list of ten possible audit procedures that they could select and are asked to select five, along with providing a rank order of the five procedures they select. All participants are informed that the audit team has already performed the procedure that would be considered most diagnostic, such that there is greater variation in possible reasonable procedures that participants can select.³

We also provide participants with information about the risk levels their audit team manager has communicated to them exist for each of five assertions: existence, completeness, completeness.

---

³ More specifically, participants in the inventory (accounts payable) conditions are told that the audit team has already decided to observe the physical count of inventory and compare test counts and tag control information with the inventory compilation (search for unrecorded liabilities by examining post-balance sheet transactions).
valuation, rights and obligations, and presentation and disclosure. All participants see that presentation and disclosure is low risk. Of the remaining four assertions, three are moderate risk and one is high risk. To enhance generalizability, we include a supplemental between-participants manipulation, varying whether the one high risk assertion is existence, completeness, valuation, or rights and obligations. After participants have selected their audit program, we ask them to provide justification for each procedure they selected as well as their audit program as whole (i.e., why they thought the procedure was important, what factors they considered when choosing the procedure, and how the procedure will provide assurance over the account). Participants then complete a post-experimental questionnaire and are dismissed, with instructions to return in one week for the second stage of the experiment.

**Stage 2 – procedure evaluation task.** During Stage 2 (one week after Stage 1), participants are tasked with evaluating an audit program over a specific account. All participants evaluate an audit program of procedures over the inventory account. Accordingly, by design, participants who designed an audit program of procedures for inventory (accounts payable) in Stage 1 have more (less) relevant experience to the evaluation task in Stage 2.

All participants first are reminded of the audit program that they designed in Stage 1. That is, each participant views the list of possible audit procedures they viewed in Stage 1, and we reproduce each participant’s five selected procedures exactly as the participant indicated in Stage 1 (i.e., numbering his/her first choice as 1, second choice as 2, etc.) After viewing this list, participants then continue to the evaluation task in the experiment. Participants receive evaluation instructions (during which we manipulate the evaluation approach, as explained in more detail shortly) and view a list of ten possible procedures to audit inventory. This list of possible procedures is identical to the list of possible procedures for the inventory condition in
Stage 1, and also similar to Stage 1, participants are informed that the audit team already chose to observe the physical count of inventory and compare test counts and tag control information with the inventory compilation.

Participants are told that they are viewing a portfolio of procedures that another person selected.\(^4\) Seven of these ten audit procedures are selected (indicated via a checkmark). Like Stage 1, we inform participants of the level of risk over each of five assertions (and these risk levels are identical to the particular risk levels that participants viewed in Stage 1). Between Stage 1 and Stage 2, we carefully customize—*for each participant who participated in the inventory condition in Stage 1*—which of the seven procedures are selected. That is, the selected seven procedures that each participant in the inventory condition views *does not* include the top three procedures that the participant selected in Stage 1. This means that we check all procedures in the list except for the participant’s top three procedures from Stage 1. In essence, this design choice (i.e., algorithm) causes participants to evaluate a portfolio of procedures that roughly “mismatches” the portfolio they selected in Stage 1. After reviewing the portfolio of procedures, participants are asked to evaluate the judgment of the person who selected this portfolio of procedures—on a scale from -5 to 5, in which -5 is “Very Deficient Judgment” and 5 is “Very Proficient Judgment.” After completing a series of post-experimental questions, participants are dismissed and provided $15.00 in cash compensation, along with course credit.

We take several steps to ensure that the “mismatching” audit program does not equate to an inferior audit program relative to the procedures that the participant selected in Stage 1. First, recall that the most diagnostic procedure over inventory— involving the physical count—has

\(^4\) Note that we are vague in this statement to avoid deception of any kind (i.e., we do not say “another person in this experiment.”) The other person designing the audit program is one of the researchers, two of whom performed the procedure of checking the seven procedures.
already been selected by the audit team. Second, recall that participants view the same assertion risk levels that they viewed in Stage 1 (with one of the five assertions being identified as high risk). We ensure that following our design algorithm described above resulted in the high risk assertion being covered by at least one selected procedure in the portfolio of seven selected procedures (and this was the case, without exception).

Third, recall that half of our participants selected a portfolio of audit procedures over a different account—accounts payable—in Stage 1. These participants represent our inspectors who have less relevant experience to the evaluation task, as they selected procedures over a different account than the account they are now evaluating (i.e., inventory). We match each participant in the accounts payable condition with a participant in the inventory condition, such that the participant in the accounts payable condition views an audit program that is identical to those viewed by his/her matched participant in the inventory condition. This design choice ensures that each customized portfolio of procedures presented to participants in the inventory condition is also evaluated by a participant in the accounts payable condition. This provides us with a benchmark of an “unbiased” evaluator (i.e., who is not influenced by the audit program selected in Stage 1).

We also took steps to ensure that the overall quality of the procedures presented to participants did not vary across experimental conditions. Since each participant was presented with a customized set of audit procedures in Stage 2, it is possible that the quality of the procedures varied across experimental conditions. Although our yoked design, and random assignment of participants to experimental conditions, helps reduce this risk, we asked two former audit partners to review the audit programs presented to our participants. These partners rated the overall quality of each individual program on a 5-point scale. We find that the quality
of audit procedures does not vary based on experimental condition (all $p > 0.14$, two-tailed), and controlling for these evaluations when performing statistical tests does not change the significance of any of our effects. In total, this evidence shows that our audit programs are of consistent quality for all participants in all experimental conditions, and our unique design allows us to isolate and test the effects of experience.

**Independent Variables**

**Relevance of prior experience manipulation.** As discussed earlier, we manipulate between-participants how relevant participants’ prior experience is to the evaluation task by varying the specific account that they select procedures for in Stage 1. Half of our participants are randomly assigned to design an audit program for *inventory* in Stage 1, while the other half of our participants are randomly assigned to design an audit program for *accounts payable* in Stage 1. Because all participants design an audit program of procedures for *inventory* in Stage 2, this design by construction prompts participants randomly assigned to the first experimental condition to have experience that is more relevant to the evaluation task. However, all participants have the general relevant experience of selecting a portfolio of audit procedures – their experience just differs in the degree of relevance.

**Evaluation approach framing manipulation.** We manipulate the framing of the evaluation approach between-participants by varying the evaluation instructions provided to participants. Specifically, in the “Inspection Approach” condition, participants view the following instructions:

*Note that an auditor’s professional judgments are fallible, and should be subject to an inspection test. The inspection test recognizes that external inspectors can reach very different conclusions when examining a portfolio of audit procedures in light of the assessed risk of material misstatement. This is true even when presented with the same risk assessments for specific assertions. So when disagreements between an auditor and*
inspector occur, they can represent an important aspect of the inspector’s evaluation of whether a deficiency in the auditor’s judgment exists.

Further, although all participants complete the same scale to evaluate the preparer’s judgment in designing the audit program (i.e., a scale from -5 to 5, ranging from very deficient judgment to very proficient judgment), the task is labeled “your inspection test” in this condition.

Alternatively, participants in the “Reasonableness Approach” condition view the following instructions:

Note that an auditor’s professional judgments are fallible, and should be subject to a reasonableness test. The reasonableness test recognizes that disagreement on the particulars is common amongst experts when examining a portfolio of audit procedures in light of the assessed risk of material misstatement. This is true even when presented with the same risk assessments for specific assertions. So when disagreements about particulars between an auditor and evaluator occur, they often reveal little about whether a deficiency in the auditor’s judgment exists.

Correspondingly, the task in this condition is labeled “Your reasonableness test.”

**Measurement of wise-thinking.** To measure the extent of participants’ wise-thinking as an attribute, we utilize the scale developed by Brewster et al. (2016). The scale includes 22 statements (from two validated psychology scales) about everyday life, for which participants indicate their level of agreement on a scale of five categories ranging from “Strongly Agree” to “Strongly Disagree.” The scale includes the seven statements from the reflective dimension of Ardelt’s (2003) wisdom scale. The other 15 statements are from the open-minded thinking scale developed by Stanovich and West (1997). These questions capture constructs like cognitive flexibility, tolerance for ambiguity, belief in simple versus certain knowledge, and willingness to consider contradictory evidence (e.g., Schömmer 1990; Cacioppo, Petty, Feinstein and Jarvis, 1996). Following Brewster et al. (2016) —and recognizing that linear coding of psychological scales can artificially truncate scale ends (see McClelland 2000) —we code the values as 0, 1, 2, 3, and 4 and then square them, resulting in values of 0, 1, 4, 9, and 16.
IV. RESULTS

Primary Analysis of Results

Our primary dependent variable is participants’ evaluation of the judgment of the individual who selected the audit procedures that they viewed in Stage 2. This dependent variable (Judgment Proficiency Evaluation) ranges from -5 (“Very Deficient Judgment”) to 5 (“Very Proficient Judgment”). Table 1 reports descriptive results for this variable by experimental condition for the between-participants manipulations of prior experience (more versus less relevant) and framing of evaluation approach (inspection versus reasonableness). For ease of interpretation, we classify participants as lower or higher on the wise thinking scale using a median split, though we revisit the continuous measure in later tests of the process. Specifically, participants scoring at or above the median (7.9) are classified as “high wise thinkers,” whereas participants scoring below the median are classified as “low wise thinkers.” Results are displayed graphically in Figure 1.

For the primary analysis, we conduct an Analysis of Variance (ANOVA) with Judgment Proficiency Evaluation as a dependent variable, and the following terms as independent variables (with all related interaction terms): a dichotomous variable Prior Experience indicating whether the participant was assigned to the more versus less relevant experience condition; a dichotomous variable Evaluation Approach indicating whether the participant was assigned to the inspection versus reasonableness condition; and a dichotomous variable Wise Thinker, indicating whether the participant has more or less propensity for wise thinking. Table 2 reports the results of these statistical tests.

<INSERT TABLE 1 and FIGURE 1 HERE>

<INSERT TABLE 2 HERE>
Prior experience has a significant main effect on evaluations of the preparer’s judgment, supporting our prediction in H1 that more relevant prior experience will cause evaluators to assess judgment as more deficient when procedures do not match their prior experience (F = 12.13; one-tailed p < .01). However, recall that the overarching prediction in H2 is that the effect of prior experience will depend on the frame of the evaluation approach as well as the evaluator’s propensity for wise thinking. In support of this overarching prediction, we also observe a significant three-way Prior Experience × Evaluation Approach × Wise Thinker interaction (F = 4.96; one-tailed p = 0.01). To perform specific tests of our hypotheses within this data, we split our results based on the Evaluation Approach independent variable and conduct more specific analyses.

H2a predicts that, when the evaluation is framed as an inspection, low wise thinkers will be more susceptible to the effect of prior experience when making their evaluations of preparer judgment than high wise thinkers. In support of H2a, we observe a marginally significant Prior Experience × Wise Thinker interaction within the inspection approach condition (F = 2.13; one-tailed p = 0.07). Panel B of Table 2 describes specific tests of this hypothesis. A planned interaction contrast test (e.g. Buckless and Ravenscroft 1990) representative of our predicted interactive hypothesis, with the values +1 (for high wise thinkers / more relevant experience), +1 (for high wise thinkers / less relevant experience), +1 (for low wise thinkers / less relevant experience), and -3 (for low wise thinkers / more relevant experience) is also significant (F = 7.64, two-tailed p = 0.008), providing additional support for our hypothesis. Simple effects tests also corroborates our expected pattern of results. High wise thinker evaluators show little sensitivity to prior experience, assessing similar (relatively high) evaluations of judgment proficiency (t = 0.17; one-tailed p = 0.34), whereas low wise thinkers show a significant
reduction in evaluated judgment proficiency with prior experience ($t = 2.50$; one-tailed $p = 0.01$).

In total, the results show that, under an inspection approach, low wise thinker inspectors are more biased by their prior experience than high wise thinker inspectors.

By contrast, and in support of H2b, we see that this pattern of results *reverses* when participants are asked to take a reasonableness approach in their evaluation. The two-way interaction between *Prior Experience* and *Wise Thinker* is significant ($F = 2.89$; one-tailed $p = 0.05$). Panel C of Table 2 describes specific tests of this hypothesis. A planned interaction contrast test (e.g. Buckless and Ravenscroft 1990) representative of our predicted interactive hypothesis, with the values +1 (for low wise thinkers / more relevant experience), +1 (for low wise thinkers / less relevant experience), +1 (for more wise thinkers / less relevant experience), and -3 (for more wise thinkers / more relevant experience) is also significant ($F = 12.09$, two-tailed $p = 0.001$), providing additional support for our hypothesis. Simple effects tests also support our expected pattern of results. Low wise thinkers show little sensitivity to prior experience, assessing similar (relatively high) evaluations of judgment proficiency ($t = 0.93$; one-tailed $p = 0.18$), whereas high wise thinkers now show a significant reduction in evaluated judgment proficiency with prior experience ($t = 3.01$; one-tailed $p < 0.01$). In total, this result reveals a flip from the pattern under an inspection approach: under a reasonableness approach, while low wise thinker evaluators’ propensity to be biased by their experience (under an inspection approach) is reduced, high wise thinkers actually become biased by their prior experience.

**Process Testing for Primary Analysis**

We conduct a test of the process to examine the mechanism through which prior experience interacts with evaluation approach, and propensity for wise thinking, to impact
inspectors’ evaluations of preparers’ judgment proficiency. We expect prior experience to bias inspectors’ evaluations by altering their perceptions of the preparers’ competence, and the variables of evaluation approach and wise thinking to jointly moderate that relation. To capture a measure of competence (Perceptions of Competence), we ask evaluators the following question: “How competent do you believe the person was that selected these procedures?” on a scale of 0 to 10, with 0 being “Not At All Competent” and 10 being “Extremely Competent.” Note that for this analysis, we use the squared continuous measure for Wise Thinker, instead of the dichotomous variable used previously.\(^5\)

<INSERT FIGURE 2 HERE>

Figure 2 displays the theoretical model. In a full regression model, with all independent variables and related interactions (and the mediator of Perceptions of Competence as the dependent variable), we see a significant three-way interaction between Prior Experience, Evaluation Approach, and Wise Thinker on Perceptions of Competence (t = -1.85; one-tailed p = 0.03). We also observe a negative effect of Prior Experience on Perceptions of Competence (t = 1.79; one-tailed p = 0.03), an interaction between Prior Experience and Approach (t = 1.65; one-tailed p = 0.05), and a marginally significant interaction between Approach and Wise Thinker (t = -1.85; one-tailed p = 0.06). Using the Preacher, Rucker, and Hayes (2007) moderated-mediation method (i.e., a bias-corrected bootstrapping technique) enables formally testing the significance of the indirect effect of Perceptions of Competence mediating the relation between Prior Experience and Judgment Proficiency Evaluation, and how this mediating relation is

---

\(^5\) Note that, to conduct a clean test of the process, we exclude participants who were in the experimental condition in which completeness was the high risk assertion for inventory, resulting in 73 total participants for this analysis. As explained shortly, this condition affords the opportunity to show evidence that, for an assertion that is unlikely to be particularly relevant for inventory, participants are not affected by their prior experience.
moderated jointly by both Evaluation Approach and Wise Thinker. We now turn to discussing results of this analysis to test the full theoretical model displayed in Figure 2.

Process results provide strong support for the role of Evaluation Approach and Wise Thinker jointly moderating the mediating effect of Perceptions of Competence, as evidenced by the confidence interval (-2.03, -0.06). Examining estimates of the mediating effect of Perceptions of Competence at low (one standard deviation below the mean), moderate (the mean), and high (one standard deviation above the mean) levels of Wise Thinker – separately within the inspection and reasonableness approach conditions – provides further insight and support. Specifically, we observe that, for less wise thinking evaluators, there is a significant, negative indirect mediating effect of Perceptions of Competence under an inspection approach (-2.92, -0.73), but not under a reasonableness approach (-2.56, 0.09). For moderately wise thinking evaluators, we observe a significant, negative indirect mediating effects of Perceptions of Competence in both an inspection approach (-1.74, -0.23) and a reasonableness approach (-2.74, -0.82). However, for high wise thinking evaluators, as expected, we observe exactly the opposite result as for low wise thinking evaluators. That is, there is a significant, negative indirect mediating effect of Perceptions of Competence under a reasonableness approach (-3.88, -1.15), but not under an inspection approach (-1.65, 0.98). In the full process model, there remains a significant direct effect of Prior Experience on Judgment Proficiency Evaluation (t = -2.00; one-tailed p = 0.03). Taken together, the process testing results provide strong support for our theory.

Robustness Test

Recall that we include a supplemental, between-participants manipulation varying whether existence, completeness, rights and obligations, or valuation is considered a high risk assertion. In addition to enhancing generalizability, this manipulation also affords the
opportunity to further corroborate that our participants have adequate knowledge to understand the audit program design task (in Stage 1) and audit program evaluation task (in Stage 2). That is, auditing standards prescribe that existence is typically relevant and higher risk for inventory, as managers have incentives to overstate asset accounts. Further, due to the risk of inventory on consignment being incorrectly included in an entity’s inventory balance, rights and obligations is also typically relevant. Finally, the valuation of inventory on the balance sheet as lower of cost or market makes valuation typically a relevant assertion. Conversely, completeness is rarely considered a high risk assertion for inventory, as managers have limited incentives to understate inventory balances.

<INSERT FIGURE 3 HERE>

Figure 3 plots how prior experience impacts participants’ evaluations of the preparer’s judgment proficiency across these four supplemental experimental conditions. Table 3, Panel A, reports descriptive statistics related to our primary dependent variable, collapsed across all independent variables other than relevance of prior experience and partitioning the data based on the assertion that was labeled high risk. Panel B of Table 3 reports an overall analysis of variance for our data when partitioned this way, and reveals a significant effect of experience (F = 15.03; p < 0.001), and a marginally significant interaction between experience and high-risk assertion (F = 2.40; p = 0.073, two-tailed). Panel C of Table 3 reports specific comparisons between experience conditions for each high-risk assertion. In support of participants’ understanding of the task, we see little effect of prior experience in the condition in which completeness was the high risk assertion (t = 0.27; p = 0.793), but strong effects of experience for all other assertion conditions (all p < 0.019). Put differently, in the condition in which the task was inconsistent with (or less relevant to) their understanding of the setting, their experience was no longer
relevant to their decisions.\(^6\) We believe that this is strong evidence that our participants understood the setting and had the appropriate level of knowledge to complete the task.

**V. CONCLUSION**

The PCAOB was created to serve as an unbiased, independent regulator of audit firms in the aftermath of the financial crisis. It performs inspections of registered accounting firms to ensure that audits are held to a standard of high quality. To accomplish this goal, the PCAOB hires inspectors with significant experience in the auditing field to perform inspections. In our study, we examine how this experience could bias the conclusions of an inspector, when the procedures they are asked to evaluate do not match what the inspector has done in the past. A person’s previous experience shapes how they perform new tasks (Libby and Luft 1993; Dane 2010), which could bias evaluations performed during an inspection. We also examine how the framing of the approach inspectors are asked to employ and the inspector’s inherent propensity to engage in wise thinking influence the effect of experience. There have been repeated calls to modify how auditors are evaluated by regulators (Treasury 2008; Turley 2007; Peecher, Solomon and Trotman 2013), and this type of change might alter the effect of experience. Moreover, propensity for wise thinking (more versus less) might influence how different types of evaluation approaches are interpreted and applied by inspectors.

Results of this study show that possessing directly relevant experience produces an unintended consequence: biasing inspectors against procedures that are different from the ones they performed in the past. We also show that this effect can vary based on the approach taken during the inspection and the level of wise thinking exhibited by the inspector. Low wise

---

\(^6\) We note that excluding participants who were in the condition in which completeness was the high risk assertion results in inferentially equivalent results in the primary analyses, consistent with these participants simply adding noise to the overall results. As noted earlier, we exclude these participants from process testing.
thinkers are less biased when asked to employ a reasonableness approach, relative to an inspection approach. We believe that this result is driven by a simple reading of the approach and interpreting it as a call for more leniency. In contrast, high wise thinkers are more biased when asked to employ a reasonableness approach. We believe that these evaluators carefully considered what reasonableness meant to them, and concluded that their best gauge of reasonableness was their previous experience. In total, the results show that experience can significantly bias inspection results, but that this bias can be mitigated under certain conditions – depending on the combination of the inspectors’ propensity toward wise thinking and approach they are asked to employ.

This study has implications for regulators and auditors. Despite the general tenor of research finding value from the PCAOB inspection process (e.g., Abernathy, Barnes, and Stefaniak 2013; DeFond and Lennox 2015; Lamereaux 2016), auditors appear to report significant dissatisfaction with the inspection process (Houston and Stefaniak 2013; Johnson et al. 2016). Our results identify one potential root cause of this dissatisfaction, thereby contributing evidence reconciling this puzzle: inspectors may be unduly biased against procedures that do not closely match what the inspector did in the past. This type of bias can be difficult to overcome, and could lead to dissatisfying inspection outcomes that do not truly address audit quality. We also note that the audit landscape is continuously changing, as new technologies and procedures become available that can provide effective and efficient assurance (Peecher, Solomon and Trotman 2013), but auditors should be aware of the effects of experience when implementing new procedures. Even if the procedures are effective, the previous experience of regulators might bias the inspectors against them.
Our results also suggest avenues for future research. We examine an inspection setting, in which the participant is placed in the role of an outside inspector and asked to evaluate procedures selected by another person. However, peer review is ubiquitous within the auditing profession, and represents another related setting where our effect could apply. Future research could examine how our result applies to within-firm reviews, and examine how other factors could mitigate or exacerbate the underlying bias when examining this related but distinct setting. Other research could also examine how the construct of propensity for wise thinking might alter audit outcomes, potentially extending the findings of Brewster et al. (2016) that auditors’ propensity for wise thinking influences how responsive their risk assessments are to more versus less corroborative audit evidence, as well as their interpretation of prompted skepticism guidance. Thinking reflectively about important audit items could aid judgment in many ways, and future research could examine how the audit environment might encourage wise thinking or bring more wise thinkers to bear on difficult problems.
REFERENCES


The adjacent graphs display descriptive statistics on evaluations of preparers’ judgment proficiency made by participants in our experiment. All participants are presented with a set of audit procedures and asked to evaluate the judgment of the preparer on an 11-point Likert scale anchored on -5 = “Very Deficient Judgment” and 5 = “Very Proficient Judgment.” We manipulate the presence of relevant experience by altering what our participants are asked to do in the first part of the experiment – half perform a more relevant task and half perform a less relevant task. We manipulate evaluation approach by altering the instructions provided to participants to focus on an “inspection” approach or a “reasonableness” approach. Wise thinking is measured using a scale developed by Brewster, Peecher, and Solomon (2016) administered at the end of the experiment. We perform a median split of these scores to classify participants as low or high wise thinkers.
Figure 2 displays the process model that we test to provide support for the mechanism underlying the interactive effect of *Prior Experience, Wise Thinker, and Evaluation Approach* on evaluations of preparers’ judgment proficiency. We utilize the moderated-mediation (bias-corrected bootstrapping) method proposed by Preacher et al. (2007).
Figure 3 plots evaluators’ assessments of preparers’ judgment proficiency based on the assertion they were told was high risk (a between-participants, supplemental manipulation in the design). They were told the other three assertions were moderate risk.
### TABLE 1. Descriptive Statistics – Judgment Evaluations

<table>
<thead>
<tr>
<th>Wise Thinking Experience</th>
<th>Inspection Approach</th>
<th>Reasonableness Approach</th>
<th>Collapsed Across Wise Thinking</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low Wise Thinker</td>
<td>High Wise Thinker</td>
<td>Collapsed Across Wise Thinking</td>
</tr>
<tr>
<td>More Relevant Experience</td>
<td>0.133 (2.588) n=15</td>
<td>1.591 (1.959) n=11</td>
<td>0.750 (2.414) n=26</td>
</tr>
<tr>
<td>Less Relevant Experience</td>
<td>2.143 (1.658) n=14</td>
<td>1.917 (1.832) n=12</td>
<td>2.2038 (1.708) n=26</td>
</tr>
<tr>
<td>Collapsed Across Experience</td>
<td>1.103 (2.381) n=29</td>
<td>1.761 (1.858) n=23</td>
<td>1.394 (2.170) n=52</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>More Relevant Experience</td>
<td>1.583 (1.165) n=12</td>
<td>-0.273 (2.284) n=11</td>
<td>0.696 (1.987) n=23</td>
</tr>
<tr>
<td>Less Relevant Experience</td>
<td>2.273 (2.195) n=11</td>
<td>2.333 (2.024) n=15</td>
<td>2.308 (2.055) n=26</td>
</tr>
<tr>
<td>Collapsed Across Experience</td>
<td>1.913 (1.730) n=23</td>
<td>1.231 (2.471) n=26</td>
<td>1.551 (2.161) n=49</td>
</tr>
</tbody>
</table>

This table presents descriptive statistics on evaluations of a preparer’s judgment for participants in our experiment. All participants are presented with a set of audit procedures and asked to evaluate the judgment of the preparer on an 11-point Likert scale anchored on -5 = “Very Deficient Judgment” and 5 = “Very Proficient Judgment.” We manipulate the presence of relevant experience by altering what our participants are asked to do in the first part of the experiment – half perform a more relevant task and half perform a less relevant task. We manipulate evaluation approach by altering the instructions provided to participants to focus on an “inspection” approach or a “reasonableness” approach. Wise thinking is measured using a scale developed by Brewster, Pecher, and Solomon (2016) administered at the end of the experiment. We perform a median split of these scores to classify participants as low or high wise thinkers.
TABLE 2. Analysis of Variance and Specific Tests – Judgment Evaluations

Panel A – Overall Analysis of Variance

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F-Stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior Experience</td>
<td>49.21</td>
<td>1</td>
<td>49.21</td>
<td>12.13</td>
<td>&gt;0.001</td>
</tr>
<tr>
<td>Evaluation Approach</td>
<td>0.02</td>
<td>1</td>
<td>0.02</td>
<td>0.01</td>
<td>0.935</td>
</tr>
<tr>
<td>Wise Thinking</td>
<td>0.49</td>
<td>1</td>
<td>0.49</td>
<td>0.12</td>
<td>0.728</td>
</tr>
<tr>
<td>Prior Experience × Evaluation Approach</td>
<td>1.43</td>
<td>1</td>
<td>1.43</td>
<td>0.35</td>
<td>0.554</td>
</tr>
<tr>
<td>Prior Experience × Wise Thinking</td>
<td>0.08</td>
<td>1</td>
<td>0.08</td>
<td>0.02</td>
<td>0.886</td>
</tr>
<tr>
<td>Evaluation Approach × Wise Thinking</td>
<td>14.22</td>
<td>1</td>
<td>14.22</td>
<td>3.50</td>
<td>0.064</td>
</tr>
<tr>
<td>Prior Experience × Evaluation Approach × Wise Thinking</td>
<td>20.12</td>
<td>1</td>
<td>20.12</td>
<td>4.96</td>
<td>0.028</td>
</tr>
<tr>
<td>Error</td>
<td>377.39</td>
<td>93</td>
<td>4.06</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Panel B – Analysis of Variance and Specific Tests for the “Inspection” Evaluation Approach Condition

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F-Stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior Experience</td>
<td>17.46</td>
<td>1</td>
<td>17.46</td>
<td>4.09</td>
<td>0.049</td>
</tr>
<tr>
<td>Wise Thinking</td>
<td>1.14</td>
<td>1</td>
<td>1.14</td>
<td>1.14</td>
<td>0.291</td>
</tr>
<tr>
<td>Prior Experience × Wise Thinking</td>
<td>2.13</td>
<td>1</td>
<td>2.13</td>
<td>2.13</td>
<td>0.151</td>
</tr>
<tr>
<td>Error</td>
<td>204.77</td>
<td>48</td>
<td>4.27</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tests of H2a

| Low Wise Thinkers / More Relevant Experience vs. All Other Cells | $F = 7.64, \ p = 0.008$ |
| Simple Effect of Relevant Experience on Low Wise Thinkers       | $t = 2.51, \ p = 0.009$ |
| Simple Effect of Relevant Experience on High Wise Thinkers      | $t = 0.41, \ p = 0.343$ |
Panel C – Analysis of Variance and Specific Tests for the “Reasonableness” Evaluation Approach Condition

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F-Stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior Experience</td>
<td>32.73</td>
<td>1</td>
<td>32.73</td>
<td>8.53</td>
<td>0.005</td>
</tr>
<tr>
<td>Wise Thinking</td>
<td>9.72</td>
<td>1</td>
<td>9.72</td>
<td>2.53</td>
<td>0.119</td>
</tr>
<tr>
<td>Prior Experience × Wise Thinking</td>
<td>11.07</td>
<td>1</td>
<td>11.07</td>
<td>2.89</td>
<td>0.096</td>
</tr>
<tr>
<td>Error</td>
<td>172.61</td>
<td>45</td>
<td>3.84</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tests of H2b

<table>
<thead>
<tr>
<th>Test</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Wise Thinkers / More Relevant Experience vs. All Other Cells</td>
<td>F = 12.09, p = 0.001</td>
<td></td>
</tr>
<tr>
<td>Simple Effect of Relevant Experience on Low Wise Thinkers</td>
<td>t = 0.96, p = 0.184</td>
<td></td>
</tr>
<tr>
<td>Simple Effect of Relevant Experience on High Wise Thinkers</td>
<td>t = 3.01, p = 0.003</td>
<td></td>
</tr>
</tbody>
</table>

This table presents descriptive statistics on evaluations of a preparer’s judgment for participants in our experiment. All participants are presented with a set of audit procedures and asked to evaluate the judgment of the preparer on an 11-point Likert scale anchored on -5 = “Very Deficient Judgment” and 5 = “Very Proficient Judgment.”. We manipulate relevant experience and evaluation approach between subjects, and perform a median split on the data based on each participant’s propensity for wise thinking. Panel A presents an overall analysis of variance for our experiment. Panels B and C perform analyses of variance and simple effects tests for the inspection and reasonableness evaluation approach conditions, respectively. All p-values are one-tailed, except for those marked with a star.
TABLE 3. Supplemental Analysis for Judgment Evaluations – Split By High-Risk Assertions

Panel A. Descriptive Statistics

<table>
<thead>
<tr>
<th>Assertion Experience</th>
<th>Existence</th>
<th>Valuation</th>
<th>Rights and Obligations</th>
<th>Completeness</th>
<th>Collapsed Across Assertion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experience</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Existence</td>
<td>Valuation</td>
<td>Rights and Obligations</td>
<td>Completeness</td>
<td>Collapsed Across Assertion</td>
</tr>
<tr>
<td>More Relevant</td>
<td>0.091</td>
<td>0.923</td>
<td>-0.046</td>
<td>1.643</td>
<td>1.643</td>
</tr>
<tr>
<td>Experience</td>
<td>(2.468)</td>
<td>(2.019)</td>
<td>(2.413)</td>
<td>(1.781)</td>
<td>(1.781)</td>
</tr>
<tr>
<td></td>
<td>n=11</td>
<td>n=13</td>
<td>n=11</td>
<td>n=14</td>
<td>n=14</td>
</tr>
<tr>
<td>Less Relevant</td>
<td>2.538</td>
<td>2.917</td>
<td>1.923</td>
<td>1.429</td>
<td>1.429</td>
</tr>
<tr>
<td>Experience</td>
<td>(1.854)</td>
<td>(0.900)</td>
<td>(1.705)</td>
<td>(2.441)</td>
<td>(2.441)</td>
</tr>
<tr>
<td></td>
<td>n=13</td>
<td>n=12</td>
<td>n=13</td>
<td>n=14</td>
<td>n=14</td>
</tr>
<tr>
<td>Collapsed Across</td>
<td>1.417</td>
<td>1.880</td>
<td>1.020</td>
<td>1.536</td>
<td>1.536</td>
</tr>
<tr>
<td>Experience</td>
<td>(2.448)</td>
<td>(1.856)</td>
<td>(2.248)</td>
<td>(2.099)</td>
<td>(2.099)</td>
</tr>
<tr>
<td></td>
<td>n=24</td>
<td>n=25</td>
<td>n=24</td>
<td>n=28</td>
<td>n=28</td>
</tr>
</tbody>
</table>

Panel B – Analysis of Variance

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F-Stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior Experience</td>
<td>60.09</td>
<td>1</td>
<td>60.09</td>
<td>15.03</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>High-Risk Assertion</td>
<td>12.36</td>
<td>3</td>
<td>4.12</td>
<td>1.03</td>
<td>0.383</td>
</tr>
<tr>
<td>Prior Experience × High-Risk Assertion</td>
<td>28.74</td>
<td>1</td>
<td>9.58</td>
<td>2.40</td>
<td>0.073</td>
</tr>
<tr>
<td>Error</td>
<td>371.77</td>
<td>93</td>
<td>3.40</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Panel C – Simple Effects Tests

<table>
<thead>
<tr>
<th>Tests of Experience Effects for Specific High-Risk Assertions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect of Relevant Experience when <em>Existence</em> Is the Highest Risk Assertion</td>
<td>$t = 2.71, \ p = 0.007$</td>
</tr>
<tr>
<td>Effect of Relevant Experience when <em>Valuation</em> Is the Highest Risk Assertion</td>
<td>$t = 3.23, \ p = 0.003$</td>
</tr>
<tr>
<td>Effect of Relevant Experience when <em>Rights &amp; Obligations</em> Is the Highest Risk Assertion</td>
<td>$t = 2.27, \ p = 0.018$</td>
</tr>
<tr>
<td>Effect of Relevant Experience when <em>Completeness</em> Is the Highest Risk Assertion</td>
<td>$t = 0.27, \ p = 0.793^*$</td>
</tr>
</tbody>
</table>

This table presents descriptive statistics (Panel A) on evaluators’ assessments of preparers’ judgment proficiency based on the assertion they were told was high risk (a between-participants, supplemental manipulation in the design). They were told the other three assertions were moderate risk. Panel B presents an analysis of variance for this supplemental analysis. Panel C reports simple effects tests. All $p$-values are one-tailed, except for those marked with a star.