Do Stronger Wise-Thinking Dispositions Facilitate Auditors’ Objective Evaluation of Evidence When Assessing and Addressing Fraud Risk?*

Billy E. Brewster
Texas State University

Alex J. Johanns
University of Illinois at Urbana–Champaign

Mark E. Peecher†
University of Illinois at Urbana–Champaign

Ira Solomon
Tulane University

April 2021

* Accepted by Jeff Cohen. We thank the anonymous reviewers, Sanaz Aghazadeh, Clif Brown, Tim Brown, Tina Carpenter, Shana Clor-Proell, Mary Curtis, Christine Earley, Brooke Elliott, Cassandra Estep, Brian Gale, Shawn Gordon, Richard Hatfield, Kamber Hetrick, Kathryn Kados, Tracie Majors, Bill Messier, Mark Nelson, Ray Pfeiffer, Chad Proell, Aaron Saiewitz, Mary Stanford, Hun-Tong Tan, Ken Trotman, Elaine Wang, Amanda Winn, and workshop participants at Louisiana State University, Mississippi State University, the New England Behavioral Accounting Research Seminar, Texas Christian University, Texas State University, University of Kentucky, University of Nevada at Las Vegas, and University of North Texas for helpful comments on earlier versions of this paper. We also thank PCAOB Board Member Duane Desparte and PCAOB inspection staff for their helpful comments during a 2018 workshop presentation held there.

† Corresponding author.
Do Stronger Wise-Thinking Dispositions Facilitate Auditors’ Objective Evaluation of Evidence When Assessing and Addressing Fraud Risk?

ABSTRACT

The objective evaluation of evidence is imperative for audit effectiveness and the proper exercise of professional skepticism. However, numerous studies suggest that auditors fail to evaluate evidence objectively when assessing or addressing the risk of material misstatement due to fraud. We develop theory to predict that auditors do evaluate evidence objectively but only when they have stronger wise-thinking dispositions (WTDs), a construct that is new to the audit literature. We define WTDs as the tendency of individuals to naturally engage in the balanced revision of beliefs and doubts about target phenomena by thinking openly and reflectively about evidence. We report prediction-consistent results from two experiments that measure the strength of participants’ WTDs and manipulate whether the underlying evidence is less or more indicative of fraud. The experimental results also document that auditors vary considerably in WTD strength and collectively demonstrate the reproducibility of audit judgment-quality benefits of stronger WTDs. We further validate the WTD construct in auditing using confirmatory bi-factor analyses to show that it has one higher-order general factor along with several subfactors. Overall, our theory and results advance the literature by identifying WTDs as a determinant of auditors’ ability to objectively evaluate evidence. Additionally, our findings have implications for standard setters and audit firms as quality control standards and audit working paper review processes might benefit from revisions that take into account that auditors do not objectively evaluate evidence unless they have stronger WTDs.

Keywords: fraud risk, earnings management, risk of material misstatement, professional skepticism, wise-thinking disposition
I. Introduction

Audit standards prescribe that auditors should objectively evaluate evidence when assessing and addressing the risk of fraud (e.g., AICPA AU-C 200, 240, 330, 500; PCAOB AS 1015, 1101, 1105, 2110, 2401). Failure to objectively evaluate evidence heightens investors’ exposure to loss and audit firms’ exposure to adverse inspections and litigation (Knechel 2013; Rapoport 2018; Kowsmann et al. 2020; Rothenberg 2020). Despite these adverse consequences, numerous experimental studies suggest that auditors often fail to objectively evaluate evidence (e.g., Nelson and Tan 2005; Nelson 2009; Knechel et al. 2013; Bauer et al. 2020). Concerns about auditor failure to evaluate evidence objectively have triggered numerous calls by regulators and commentators for auditors to heighten their exercise of professional skepticism—that is, do better at critically evaluating evidence with a questioning mind (Nagarajan 2020; Nicodemus 2020).\(^1\) Academics also have called for research that identifies characteristics of individual auditors that facilitate or inhibit objective evaluation of evidence (e.g., Brown et al. 1999; Kachelmeier et al. 2014; Nolder and Kadous 2018).

We introduce wise-thinking dispositions (WTDs), a new construct to the audit judgment literature, and develop theory predicting that auditors objectively evaluate evidence of fraud when they have stronger, as opposed to weaker, WTDs. Thinking dispositions are behavioral capacities and inclinations to engage in particular patterns of cognitive behavior (Perkins et al. 1993), or higher-level cognitive styles (Stanovich 2001). They guide the application of decision-makers’ intellect, knowledge, and abilities (Fuller and Kaplan 1994; West and Sá 1999; Sternberg and Grigorenko 1997; Perkins et al. 2000; Sternberg 2001). We define stronger WTDs as dispositions

---

\(^1\) Questions about auditor objectivity often arise when alleged financial statement frauds are exposed (e.g., WorldCom and Wirecard). To the extent that auditors fail to objectively evaluate “red flag” evidence indicative of fraud during the audit, auditor detection of fraud is less likely (Davies 2020; Tokar and Davies 2021).
of individuals to naturally engage in wise thinking. Wise thinking entails the balanced revision of beliefs and doubt about target phenomena by thinking openly and reflectively about available evidence (Meacham 1983; Kitchener and Brenner 1990; Ardelt 2003; Staudinger and Glück 2011a). We develop the WTD construct by drawing on literatures that conceptualize thinking dispositions, wisdom, and epistemic rationality (e.g., Meacham 1983; Stanovich and West 1997, 1998; West and Sá 1999; Ardelt 2003, 2004; Birren and Svensson 2005; Stanovich and West 2007, 2008; Staudinger and Glück 2011a; Glück, et al. 2013). One substantive contribution of this paper is introducing and articulating the WTD construct and demonstrating its amenability to low-cost, valid empirical measurement.

An open question which we address herein is, “To what extent do audit professionals vary in WTD strength?” On the one hand, if auditors with weaker WTDs consistently fail to objectively evaluate evidence, as we predict, audit firms may directly or indirectly prevent such auditors from entry and promotion if they do not possess or develop strong WTDs. On the other hand, research suggests that auditors often are not rewarded solely for objectively evaluating evidence of financial statement fraud (e.g., Peecher et al. 2013; Hobson et al. 2017). Further, auditors seemingly are rewarded for attaining management-preferred goals and for seeing the world through management’s eyes (e.g., Kadous et al. 2003; Koch and Salterio 2017; Bhaskar et al. 2019). Thus, our second contribution is providing initial empirical evidence on variation in auditors’ WTD strength.

A third contribution of the paper is providing theory-based evidence from two experiments showing that auditors with relatively stronger WTDs objectively evaluate and respond to audit evidence when assessing and addressing fraud risk. Our experiments use different audit contexts for enhanced confidence in the reproducibility of the predicted judgment benefits of stronger
WTDs (Hail et al. 2020). Importantly, the dependent variables in the two experiments are different in a key respect. Experiment 1 participants assess fraud risk after management makes last-minute reduction to warranty expense that allows earnings to narrowly beat analysts’ forecasts. In Experiment 2 participants also assess fraud risk, but in a goodwill impairment setting. Moreover, Experiment 2 participants also specify planned actions to address fraud risk. This latter dependent variable is important because, unless auditors undertake different actions to address heightened fraud risk, the chance of fraud detection does not change.

Another key difference is that while all Experiment 1 conditions include a skepticism prompt, Experiment 2 excludes this manipulation. This design difference allows a fourth contribution: investigation of whether the presence or absence of skepticism prompts attenuates the judgment benefits of stronger WTDs. That is, even if stronger WTDs enable auditors to evaluate evidence objectively when coupled with a skepticism prompt, the question of whether stronger WTDs help unprompted auditors’ objectivity remains unaddressed. Thus, reporting the results of a second experiment re-examining whether stronger WTDs enable auditors to objectively evaluate evidence given the absence of any skepticism prompt strengthens our theory-testing (Asay et al. 2021).

Findings in both experiments indicate that the strength of auditors’ WTDs varies considerably and provide support for our prediction. In Experiment 1, only auditors with stronger WTDs assess fraud risk to be significantly higher (lower) when Evidence is more (less) indicative of fraud, and they do so regardless of skepticism Prompt condition. In Experiment 2 only auditors

---

2 Consistent with Hail et al. 2020, we use reproducibility to mean replication in the wide sense, which occurs when hypothesis-consistent findings in an initial study obtain in a new study that employs variations of theory-relevant tasks, participants, and contexts. Replication in a narrow sense occurs when hypothesis-consistent findings obtain in a new study that exactly repeats the prior study’s experimental design. While both have value, reproducibility more strongly addresses concerns about theory generalizability.
with stronger WTDs are significantly more likely to take actions that more persuasively address the heightened risk of fraud when the underlying Evidence is more indicative rather than less indicative of fraud. Finally, we examine WTDs within a confirmatory bi-factor model as well as use a third experiment to examine the divergence of WTDs from Hurtt Trait Skepticism (HTS) scores. Results confirm that WTDs are comprised of one higher-order general factor as well as three subfactors and that WTDs readily diverges from HTS. We conclude the paper by discussing implications and by identifying numerous opportunities for future research.

2. Background and hypotheses

Auditors’ “on average” failure to objectively respond to evidence

Audit standards (e.g., AU-C 240; PCAOB AS 1105; 2110; 2301) and audit research (e.g., Bell et al. 2005; Peecher, et al. 2007) prescribe that auditors should objectively evaluate evidence when assessing and addressing the risk of material misstatement due to error or fraud. The objective evaluation of evidence necessitates a “balanced assessment of all relevant circumstances” when reaching audit conclusions (e.g., IIA, Code of Ethics 2021), and it is paramount for the proper exercise of professional skepticism (PCAOB AS 1101.08-.11; 1105). Nolder and Kadous (2018, 5) observe how audit standards’ description of professional skepticism calls for “… receptivity, openness, or alertness to new information and an objective and unbiased assessment of the merits of evidence …” Auditors should increase (decrease) the assessed risk of fraud when evidence is objectively more (less) indicative of material misstatement due to fraud. They should also objectively alter their audit plans to address higher or lower fraud risk. However, numerous studies find that, on average, auditors often fail to adapt their plans (e.g., Brown et al. 1999; Waller and Zimbelman 2003; Nelson and Tan 2005; Nelson 2009; Hammersley et al. 2010; Kochetova-
Kozloski et al. 2011; Trotman and Wright 2012; Trompeter et al. 2013; Knechel et al. 2013; Hammersley et al. 2011; Kachelmeier et al. 2014; Griffith et al. 2015).

Zimbelman (1997), for example, finds that audit seniors and managers fail to alter assessed misstatement risk based on the presence versus absence of incentives for management to materially misstate assets. Similarly, audit seniors diagnosing the cause of an unexpected increase in income fail to prioritize evidence having greater diagnostic value (Brown et al. 1999). Hoffman and Zimbelman (2009) find that, unless prompted to think strategically, auditors fail to appropriately revise planned audit procedures to address seeded fraud cues. Trotman and Wright (2012) find that auditors fail to respond to external performance indicators that objectively heighten misstatement risk despite seemingly positive internal performance indicators under management’s control. In an abstract setting, Kachelmeier et al. (2014) find that participants in an auditor role fail to distinguish between evidence that is more versus less indicative of intentional misstatement. Overall, auditors appear on average to have difficulty objectively evaluating evidence.3

The cult of the average, thinking dispositions, and wise-thinking dispositions

Social psychologists warn against “the cult of the average” when studying determinants of human behavior. Achor (2010, 9-10), as an example, notes:

The typical approach to understanding human behavior has always been to look for average behavior….This misguided approach has created what I call ‘the cult of the average’ in the behavioral sciences….If someone asks a question such as ‘How fast can a child learn how to read in a classroom? science changes that to ‘How fast does the average child learn to read in the classroom?’ We ignore the children who read faster or slower and tailor the classroom toward the ‘average’ child….That’s the first mistake psychology makes.4

3 Despite numerous studies showing that auditors have difficulty in responding to evidence appropriately, auditors do assess the risk of material misstatement to be higher when given salient evidence of low management integrity (e.g., Peecher 1996; Goodwin 1999; Kizirian et al. 2005; Quadackers et al. 2014; Dikolli et al. 2020). This finding may reflect humans’ adaptive tendency to interact more cautiously with people who exhibit signs of lower integrity (e.g., Kim et al. 2004).

4 As another example, Ben-Shahar (2007, 137-38) remarks, “Psychologists who argue that the depth of our happiness is fixed make the ‘error of the average’—they derive their conclusions from what most people do while ignoring those
With few exceptions (e.g., fluid intelligence, as in Brown et al. 1999), the audit literature has yet to identify auditor-specific constructs that result in “above average” objectivity in evaluating evidence. To address this gap, we introduce the construct of a wise-thinking disposition and explain that this construct is a way to distinguish among auditors in terms of their propensity for objectively evaluating audit evidence. Thinking dispositions are behavioral capacities and inclinations to engage in particular patterns of cognitive behavior (Perkins et al. 1993), or higher-level cognitive styles (Stanovich 2001). They capture “…variation in peoples’ goal management, epistemic values, and epistemic self-regulation—differences in the operation of the reflective mind” (Stanovich et al. 2018, 26). Thinking dispositions also guide the application of decision-makers’ intellect, knowledge, and abilities (West and Sá 1999; Sternberg and Grigorenko 1997; Perkins et al. 2000; Sternberg 2001), which helps calibrate how strongly they believe something to be true in light of available evidence (Stanovich et al. 2018). A useful analogy is to compare individuals’ intellect, knowledge, and abilities to the engines of cars while thinking dispositions are the drivers of cars (Stanovich 2009).

Stanovich (2009, 35) also notes that “many different studies involving thousands of participants have indicated that measures of intelligence display only moderate to weak correlations (usually less than 0.30) with some thinking dispositions (for example, actively open-minded thinking, need for cognition) and near zero correlations with others (such as conscientiousness, curiosity, diligence).” Thinking dispositions also differ from knowledge in that while knowledge improves rapidly with instruction, thinking dispositions “are stable traits that help explain intellectual performance …” This stability results in them not being easily influenced who don't fit the norm.” He goes on to argue that “Often, it is those outside the norm, the exceptional ones, who point to the truth of what is possible.” We similarly argue that stronger WTD auditors will demonstrate that auditors can objectively evaluate evidence, despite prior findings that “average” auditors fail to objectively evaluate evidence.
by situational factors (Perkins et al. 2000, 269). Thinking dispositions likely can improve, but only over the longer term, with certain instructional approaches. For example, compared to a traditional knowledge-transmission model of instruction, an enculturation model better reinforces thinking dispositions; one that regularly asks learners to exercise the desired disposition by thinking out loud or by listening to others do so and then reflecting on verbalized thoughts (Tishman et al. 1993; Perkins et al. 2000).

The disposition on which we focus concerns individuals’ disposition to engage in wise thinking. Wise thinking is rooted in the construct of wisdom in both sociology (Ardelt 2003; 2004; 2015) and psychology (Sternberg 2003; Staudinger and Glück 2011a).6 Until recently, few studies characterized the wisdom construct in ways that enabled empirical measurement (Staudinger et al. 1992; Birren and Svensson 2005). Several recent studies, however, find that reflective thinking, which balances the dialectic between knowing and doubting in light of uncertainty, is the paramount taproot of wisdom (Ardelt 2003; Stanovich and West 2008; Hall 2010; Staudinger and Glück 2011b). Birren and Fisher (1990, 325) also focus on this dialectic by describing wisdom as a “metacognitive style that involves the ability to make sound judgments; knowing that one does not know everything, seeking the truth to the extent that it is knowable.” Our working definition of stronger WTDs, therefore, is dispositions by which individuals naturally engage in wise thinking, which is the balanced revision of beliefs and doubts about target phenomena by thinking openly and reflectively about available evidence (Birren and Fisher 1990; Meacham 1983; Kitchener and Brenner 1990; Ardelt 2003; Staudinger and Glück 2011a). Thus, stronger WTDs likely improve auditors’ propensity to form different beliefs and take different actions given an

---

5 We thank sociologist Monika Ardelt for her suggestion to use wise thinking instead of wisdom to describe our dispositional construct. As she notes, “Wisdom entails more than a thinking disposition, even though wise thinking is certainly worthwhile to investigate in itself” (Ardelt, personal e-mail communication, November 30, 2015).
objective evaluation of different states of evidence (Baron 1993, 2008; Meacham 1990; Tishman et al. 1993; Ardelt 2003; Stanovich 2009). For example, compared to weaker WTD auditors, stronger WTD auditors would be more concerned about an auditee management explanation that a spike in its inventory simply reflects a forthcoming season of favorable holiday sales when the auditee’s competitors are writing down similar inventory items (Bell et al. 2005). Formally, we predict:

**Hypothesis 1.** Auditors’ fraud risk assessments and planned actions to address fraud risk will reflect a higher (lower) risk of fraud when evidence is more (less) indicative of fraud, but only when they have stronger wise-thinking dispositions.

Implicit in this hypothesis is the assumption that sufficient variation exists in WTD strength across auditors. While prior studies demonstrate that adults vary in wisdom and epistemic rationality, no study of which we are aware examines these two phenomena in professionals charged with a responsibility to be objective. Adults who self-select into the audit profession may do so in part because they have and enjoy applying relatively strong WTDs. It also could be that professions have developed screening processes to weed out individuals with weaker WTDs. This possibility is not easily reconciled, however, with prior experimental findings that, at least on average, auditors fail to objectively respond to audit evidence. Such findings suggest either our hypothesis is true but auditors generally have weaker WTDs, or that our hypothesis is false in that even auditors who have stronger WTDs fail to objectively evaluate evidence. As such, we first conduct an initial analysis in each experiment to investigate the following research question (RQ):

**RQ1.** To what extent do auditors vary in the strength of their WTDs?

In addition, we investigate whether the presence and the nature of external professional skepticism prompts constitute boundary conditions for our hypothesis. Audit standards implicitly assume that being mindful of the duty to exercise professional skepticism throughout the audit will
help ensure the objective, good-faith evaluation of audit evidence (e.g., PCAOB AS 1015.07 - 0.09). Skepticism prompts are not designed to be evidence of fraud in and of themselves, but they remind auditors of particular attitudes to take when evaluating evidence (e.g., Peecher 1996; Brown et al. 1999; Turner 2001; Peytcheva 2014; Grenier 2017). Thus, we investigate whether skepticism prompts moderate our hypothesis. We do so by including two skepticism prompts in Experiment 1 and by excluding skepticism prompts in Experiment 2. Experiment 1’s skepticism prompts capture elements of the ongoing evolution in audit standards’ prescriptions about the exercise of professional skepticism. Until the late 1990’s, audit standards had emphasized that professional skepticism requires a neutral attitude towards evidence (e.g., SAS No. 53). Today’s audit standards, by contrast, have evolved towards an attitude of presumptive doubt (Bell et al. 2005; Nelson 2009; Cohen et al. 2017). Current standards have removed some of the neutrality language and added language that requires auditors to set aside prior beliefs about management’s honesty during fraud brainstorming sessions (e.g., SAS No. 99). To capture this evolution, we employ both a less doubtful prompt (neutral) that reminds auditors to be skeptical if signs emerge

These studies provide mixed evidence that skepticism prompts can improve auditor judgment. In brief, compared to auditors prompted to be objective, auditors prompted to be skeptical of management’s explanations for unusual account-balance fluctuations do not change how likely they assess these explanations as the cause of the fluctuations, nor do they generate more misstatement explanations for the fluctuation (Peecher 1996). Similarly, as compared to a prompt that expresses concern about how auditors approach management explanations, a prompt to be sufficiently skeptical does not result in auditors searching for more evidence, spending a longer time searching for evidence, or shifting away from a management-directed search strategy (Turner 2001). Of some concern, this type of skepticism prompt results in auditors becoming information prone by ascribing high value to evidence without regard to its diagnosticity (Brown et al. 1999). On perhaps a more positive note, skepticism prompts do improve the cognitive performance of audit students (but not auditors) in Peytcheva (2014), and skepticism prompts increase audit specialists’ generation of fraud explanations for unusual account-balance fluctuations as well as increase their assessment of the likelihood of fraud in Grenier (2017).

SAS No. 53 ¶16 (AICPA 1988) states, “The auditor neither assumes that management is dishonest nor assumes unquestioned honesty.” It further explains, “A presumption of management dishonesty … would be contrary to the accumulated experience of auditors. Moreover, if dishonesty were presumed, the auditor would potentially need to question the genuineness of all records and documents obtained from the client …. An audit conducted on these terms would be unreasonably costly and impractical” (SAS No. 53, ¶17 AICPA 1988).

SAS No. 99 ¶13 (AICPA 2006) also notes, “In exercising professional skepticism in gathering and evaluating evidence, the auditor should not be satisfied with less-than-persuasive evidence because of a belief that management is honest.” This prescription is tenuous normatively, as how much evidential matter one needs to be sufficiently persuasive itself differs depending on whether the source is trustworthy or untrustworthy.
that management lacks integrity and a more doubtful prompt (doubtful) that reminds them to be skeptical even if no signs emerge that management lacks integrity.

A priori, both the neutral and the doubtful levels of skepticism Prompt could be boundary conditions of our hypothesis that auditors objectively evaluate evidence only when they have stronger WTDs. One, auditors with stronger WTDs could opt to override their natural thinking disposition to objectively evaluate evidence of fraud if prompted to be doubtful. That is, when prompted to be doubtful, auditors with stronger WTDs could assess and address fraud risk in the way we predict that auditors with weaker WTDs naturally do so—without regard to the underlying evidence of fraud. Two, prompting neutrality could motivate auditors with weaker WTDs to become more aware of whether or not red-flag signs of fraud exist. If prompted to be neutral, auditors with weaker WTDs could assess and address fraud risk in the way we predict that auditors with stronger WTDs do so naturally—with objective regard to the underlying evidence of fraud.

Thus, we investigate the following additional RQs:

**RQ2.1** Does a doubtful prompt impair the propensity for auditors with stronger WTDs to reach judgments and decisions that reflect a higher (lower) risk of fraud when the underlying audit evidence is more (less) indicative of fraud?

**RQ2.2** Does a neutral prompt enhance the propensity for auditors with weaker WTDs to reach judgments and decisions that reflect a higher (lower) risk of fraud when the underlying audit evidence is more (less) indicative of fraud?

**RQ2.3** Does the propensity for auditors with stronger WTDs to respond objectively to evidence that is more indicative versus less indicative of fraud (Hypothesis 1) hold in both the presence and the absence of skepticism prompts?
3. Experiment 1

Participants and task

Participants are 87 auditors from a Big Four firm with an average audit experience of 44.5 months (standard deviation = 15.5). Auditors at this level commonly evaluate management’s nonrecurring journal entries. To verify that our participants had the requisite task experience, we asked them to report the number of engagements in which they had reviewed nonrecurring journal entries. The mean (standard deviation) of their response is 10.1 (19.7), so we conclude that our participants have the requisite task experience.

In Part I, participants first read a randomly assigned skepticism prompt (see below) and then they read about a fictitious audit client which manufactures and sells auto parts. They receive summary financial performance, a description of a prior favorable relationship with management, and a mock Forbes magazine interview by the CFO emphasizing integrity and a commitment to high-quality financial reporting. Participants next evaluate a nonrecurring journal entry that reduces warranty expense. Earnings fell short of analysts’ forecasts before this entry but beat forecasts afterwards. The CFO explains that lower warranty expense reflects higher product quality and lower future product returns. The Evidence manipulation (see below) either corroborates or undermines management’s explanation. Participants then assess the likelihood that management is trying to commit fraud, engage in within-GAAP earnings management, or use high-quality

9 We obtained Institutional Review Board approval for each reported study.
10 Neither months of experience ($t_{1,85} = 0.44, p = 0.663$) nor number of prior engagements with nonrecurring entries ($t_{1,73} = 0.39, p = 0.699$) differ across auditors who have stronger versus weaker WTDs. We did not collect information on age, gender, or ethnicity.
11 Specifically, the CFO says, “… the whole mentality behind earnings management is unhealthy…..” And, after being asked, “So integrity trumps earnings management in your eyes?” responds, “The two terms are essentially contradictory…. Let the short-term periodic earnings fall where they may…. If you were to give me a choice between high-quality earnings that barely miss analysts’ forecasts versus low-quality earnings that beat their forecasts, I’d take the high-quality earnings because that’s best for the long term.”
accounting. After Part I, participants place materials in an envelope and begin Part II, providing demographics and completing scales for WTD and Tacit Knowledge.\textsuperscript{12}

**Experimental design**

We use a $2 \times 2 \times 2$ (×3) mixed design. The (×3) within-subjects part of the design entails the assessed likelihoods of Attempted Fraud, Within-GAAP Earnings Management, and High-Quality Attempt. The $2 \times 2$ entails between-subjects manipulations of Evidence (more indicative vs. less indicative of fraud) and skepticism Prompt (doubtful vs. neutral) as well as a median-split measure of participants’ WTDs (stronger vs weaker).

A standard way to reveal if participants rely on a given factor in reaching judgments or decisions is to manipulate that factor and observe whether or not it influences their responses (e.g., Tversky and Kahneman 1982; Toplak et al. 2011). The approach we use is as follows: In the less indicative Evidence condition, two key performance indicators (KPIs), lower product returns and less rework, corroborate management’s explanation for reducing warranty liability (see Exhibit 1, panel A). In the more indicative Evidence condition (Exhibit 1, panel B) two additional KPIs cast doubt on management’s higher product quality story: assembly-worker overtime and production-line utilization rates increased dramatically in the 4th quarter relative to the prior year, likely leading to higher worker fatigue and reduced product quality, thereby increasing future product returns.

To implement the skepticism Prompt, we use an FBI double-agent vignette in both levels. This vignette is unrelated to the client but captures the fact that auditors may face a strategic

\textsuperscript{12} We included tacit knowledge questions to rule out the possibility that auditors with stronger WTDs also tend to have higher tacit knowledge, in which case tacit knowledge could also be driving our results. However, the correlation (untabulated) between these two measures is not significant ($r = +0.011$, $t = 0.14$, $p = 0.891$). Further, untabulated analyses show that Tacit Knowledge does not influence auditors’ risk assessments in isolation or in interaction with other independent variables (lowest $p = 0.307$).
opponent who tries to commit and conceal fraud (e.g., Bowlin 2011). The doubtful level emphasizes being skeptical even if no signs appear to exist that management lacks integrity (see, e.g., SAS No. 99, para 13). The less doubtful neutral level emphasizes being skeptical if red-flag signs emerge that management lacks integrity (see, e.g., SAS No. 99, paras 31-40). The doubtful vignette describes an FBI double agent who provided no outward cues of treason (see Exhibit 2, panel A). The neutral vignette describes the same double agent but reveals red flags of treason—that is, a lot of cash inexplicably found in his home (see Exhibit 2, panel B).

To measure auditors’ WTDs, we construct a survey with 22 statements from everyday life (see Exhibit 3). The survey adapts statements from Ardelt’s (2003) Three-Dimensional Wisdom Scale (i.e., 3D-WS) and from Stanovich and West’s (2007) Actively Open-Minded Thinking (AOMT) scale. The three dimensions of 3D-WS are reflective, cognitive, and affective, with the reflective dimension being paramount (Ardelt 2003). As such, 5 of the WTD survey statements adapted from 3D-WS relate to its reflective dimension, and two WTD survey statements are adapted from the cognitive dimension (see Exhibit 3). Our remaining 15 WTD survey statements are reflective and cognitive items from the AOMT. A disposition to use active open-minded thinking is foundational for wise thinking, as prior literature notes (Stanovich 2001; Stanovich and West 2007, 226; Stanovich and West 2008, 130; Stanovich 2009). High AOMT scores, for example,

13 Like our scale, statements used in these two scales originate from other sources, including other thinking disposition scales. Questions from Ardelt’s 2003 3D-WS scale originate from Perspective-Taking Scale of the Interpersonal Reactivity (Davis 1980), Need for Cognition (Cacioppo et al. 1996), Dogmatism (Rokeach 1960), Ambiguity Tolerance (MacDonald 1970), Intolerance of Ambiguity (King and Hunt 1975), and Personal Problem-Solving Inventory (Heppner and Peterson 1982). Questions from Stanovich and West’s (2007) composite AOMT that we use also come from earlier-developed scales: Cognitive Flexibility (Stanovich and West 1997), Constructive Thinking (Epstein and Meier 1989), Dogmatism (Paulhus and Reid 1991), Willingness to Consider Contradictory Evidence (Stanovich and West 1997), Tolerance for Ambiguity (e.g., Epstein and Meier 1989), Need for Cognition (e.g., Cacioppo et al. 1996), Counterfactual Thinking (e.g., Stanovich and West 1997); and Belief Identification (e.g., Sá et al. 1999).

14 While the affective dimension is relevant for interpersonal contexts involving empathy and compassion (Ardelt 2003, 2004), it is less relevant to reflective, epistemically rational thought needed to evaluate evidence objectively. Future research can investigate whether adding the affective dimension improves measurement of WTDs.
“indicate openness to belief change and cognitive flexibility, whereas low scores indicate cognitive rigidity and resistance to change” (Stanovich and West 1997, 347).  

Participants express agreement, neutrality, or disagreement with each statement, using two categories of agreement and disagreement (e.g., strongly agree, agree, neutral, disagree, or strongly disagree). For each statement, the wisest response is strongly agree or strongly disagree. We score WTDs using the sum of the squared deviations: we record a deviation of 0, 1, 2, 3, or 4 and then square these deviations—that is, 0, 1, 4, 9, or 16, going from most-wise to least-wise responses. We then sum the squared deviations across the 22 statements in Exhibit 3 with lower (higher) values indicative of stronger (weaker) WTD.

Using the sum of squared deviations is a common way to score forms of practical intelligence. For example, tacit knowledge scales in psychology and accounting studies use this approach (e.g., Wagner 1987; Tan and Libby 1997; Bol et al. 2018). Using the sum of squared deviations helpfully increases the systematic variation in a measured construct relative to using the sum of linear deviations, but also is more appropriate than using the sum of linear deviations for such scales, for the following reasons: (i) It exacts less of a relative penalty for responses that directionally agree with the best response; (ii) It better captures the intuition that responses that run in the opposite direction of the best response are qualitatively worse than being neutral about what is the best response; (iii) It more heavily penalizes strong dissent from the best response; and (iv) It more heavily penalizes responses that contain more random variation across the 22 items.

15 Using samples drawn from student and lay adult populations, some studies examine demographic correlates of Ardelt’s 3D-WS and Stanovich and West’s AOMT. Ardelt (2009), for example, reports that 3D-WS scores do not significantly differ by gender among undergraduates, but that older adults outscore undergraduates on the cognitive dimension. Plews-Ogan et al. (2016) find 3D-WS scores vary across physicians but not by physicians’ age or gender. Some evidence indicates that males outscore females on the AOMT (Stanovich et al. 2018), although the effect size is small (0.27 < Cohen’s D < 0.36).

16 Greater systematic variation using the sum of squared deviations manifests in a larger number of unique WTD strength scores across participants. A sum of linear deviations approach (untabulated) yields only 20 (23) unique
**Dependent measures**

We aimed to capture the degree of participants’ concern about potential fraud in light of the fact that earnings distortion activities fall along a “continuum that ranges from complete legitimacy at one extreme to fraud at the other” (POB 2000, 78). As shown in Exhibit 4, we asked participants to assess the likelihood that management is attempting to engage in each of the following:

- **Fraud** or **Near Fraud** – An intentional attempt to distort business reality that clearly violates GAAP or that really “pushes the envelope” of GAAP.
- **Earnings Management** – An intentional attempt to “window dress” business reality that is acceptable under GAAP.
- **High-Quality Attempt** – A good-faith attempt by management to use accounting that fairly and neutrally reflects business reality. This attempt could be successful, resulting in *High Quality Accounting*, or unsuccessful, resulting in *Erroneous Accounting*.

We include the near fraud category to capture instances in which auditors are concerned about potential fraud but reluctant to believe management is committing out-and-out fraud (e.g., Hobson et al. 2017). A factor analysis (untabulated) on these five categories returns three factors, but with the first factor being the strongest. The first factor explains 95% of the common variance, equally weights **Fraud** (loading = 0.78) and **Near Fraud** (loading = 0.74), and has an eigenvalue of 1.36. Eigenvalues for the two other factors are both below 1.0, with the second factor equally weighting **High-Quality Accounting** (loading = 0.54) and **Erroneous Accounting** (loading = 0.57). **Earnings Management** is the third factor (loading = 0.25). Given this factor structure, we use two dependent measures to test our hypothesis. One dependent measure is **Attempted Fraud**—the average of **Fraud** and **Near Fraud**—as the dependent variable. The second dependent measure is a difference score between **Attempted Fraud** and the average of the two non-misstatement categories of **Within-GAAP Earnings Management** and **High-Quality Attempt**.¹⁷

¹⁷ A difference score approach allows for examination of the *relative* likelihood of fraud versus non-misstatement.
Check on prompt manipulation

Right after reading the Prompt manipulation, participants assess whether the main point of the FBI double-agent narrative is that auditors should be skeptical when cues emerge that signal management lacks integrity (boldface in original). On a 7-point scale (1=completely disagree; 4 no opinion; 7=completely agree), participants’ evaluations fall below the midpoint in the doubtful condition but above the midpoint in the neutral condition (3.03 < 5.80; t75 = -6.23, p < 0.001, one-tailed), consistent with a successful Prompt manipulation. 18

Research Question 1: Does the strength of auditors’ WTDs vary?

In untabulated analyses, we find considerable variation in the strength of auditors’ WTDs, ranging from 31 to 135 (mean = 68.53, standard deviation = 19.82). Further, we find no evidence that greater general or task experience significantly correlates with stronger WTDs. Neither months of experience (t1,85 = 0.44, p = 0.663) nor number of prior engagements with nonrecurring entries (t1,73 = 0.39, p = 0.699) differ for auditors who have stronger versus weaker WTDs. 19 Consistent with random assignment, WTD strength does not differ across the Prompt manipulation (F 1, 83=1.08, p=0.302), the Evidence manipulation (F 1, 83=0.99, p=0.323), or their interaction (F 1, 83=1.34, p=0.250). Last, we use a median split to classify auditors as having stronger or weaker WTDs to enable use of planned contrasts. 20 Means (standard deviations) for auditors with stronger WTDs are 51.95 (9.66) and for auditors with weaker WTDs are 83.32 (13.93).

18 We perform no manipulation check on Evidence as theory predicts that auditors with weaker WTDs are not likely to notice differences in Evidence while auditors with stronger WTDs likely would notice these differences. Findings discussed hereafter show that stronger WTD auditors reacted significantly to the Evidence manipulation, consistent with a successful manipulation of Evidence (see Sigall and Mills 1998, 221). Last, we did capture auditors’ misstatement risk expectations before presentation of Evidence. Separate ANOVAs (untabulated) reveal no statistically significant differences on expected likelihood of fraud or near fraud due to Prompt, WTDs, Evidence, or their interactions.

19 All p-values throughout are two-tailed unless otherwise noted. We use an alpha of 0.05 for statistical significance.

20 As Iacobucci et al. (2015) note, an advantage of median splits is parsimony and qualitatively similar results as continuous specifications would yield so long as the continuous variable subjected to the median split is not
Hypothesis 1 Test

Operationally, Hypothesis 1 predicts an Evidence effect, but only for stronger WTD auditors which would manifest as a two-way Evidence × WTD interaction. That is, we predict relatively higher likelihood assessments of fraud given Evidence that is more indicative as compared to less indicative of fraud, but only for stronger WTD auditors. Using the language of Guggenmos et al. (2018, 232), the hypothesis predicts a “Pac-Man” shaped interaction. Table 1 presents means, standard errors, and the number of observations for assessed likelihoods of Attempted Fraud, Within-GAAP Earnings Management, and High-Quality Attempt in columns (1), (2), and (3), respectively. Column (4) reports mean difference scores of Attempted Fraud minus the average likelihood of the two non-misstatement categories—that is, Within-GAAP Earnings Management and High-Quality Attempt. For example, the top row of panel A reports that stronger WTD auditors in the more indicative Evidence, doubtful skepticism Prompt condition assessed average likelihoods of 61.22, 67.67, and 34.22 for Attempted Fraud, Within-GAAP Earnings Management, and High-Quality Attempt, respectively. As such, the mean difference score Fraud vs. Not Misstated equals 61.22 − (67.67 + 34.22)/2 = +10.28. We infer objective evaluation of evidence if auditors assess significantly higher likelihoods of fraud when Evidence is more indicative of fraud than when it less indicative of fraud. We infer an absence of objective evaluation of evidence if assessed likelihoods of fraud do not differ significantly across Evidence levels or if they differ significantly, but in the direction opposite to what the Evidence warrants.

significantly correlated with experimentally manipulated factors. Nonetheless, median-splits reduce power and can add classification error as compared to continuous specifications (McClelland et al. 2015). We address this concern by examining whether our results are robust to use of continuous specifications.

21 The appropriate weights are +1 and -1 for conditions that refer to Pac-Man’s open mouth and 0 for the other two cells; Pac-Man looks like this, 🍂 (Wikimedia Commons 2015).
We test our hypothesis by using visual fits, planned contrasts, semi-omnibus tests, and contrast variance residuals, $q^2$ (Guggenmos et al. 2018). Figure 1 allows for visual inspection. The predicted “Pac-Man” shapes clearly manifest for both dependent variables. The contrast weights are $[0, +1, 0, -1]$ corresponding to the labels [A, B, C, D] in Figure 1’s graphs. Table 2 presents a three-way ANOVA and our planned contrast. The planned contrast is highly significant using Attempted Fraud ($p_{(one-tailed)} = <0.001$; see panel A) and using Fraud vs. Not Misstated ($p_{(one-tailed)} = <0.001$; see panel B). As hypothesized, auditors assess a significantly higher (lower) risk of material misstatement due to fraud when Evidence is more (less) indicative of fraud, but only if they have a stronger WTD. In addition, the semi-omnibus F-test for the residual between-cell variance is not significant for Attempted Fraud ($p=0.158$; see panel A) or for Fraud vs. Not Misstated ($p = 0.118$; see panel B). Finally, there is evidence that our independent factors influenced auditors judgments in ways other than hypothesized in that the contrast variance residual is nontrivial both for Attempt Fraud ($q^2 = 42.3\%$) and for Fraud vs. Not Misstated ($q^2 = 44.3\%$). This is not surprising as the ANOVA models in Table 2 panels A and B indicate a significant main effect for Prompt, which is not part of our hypothesis, but which we investigate further below. Overall, these results support Hypothesis 1: auditors assess a significantly higher (lower) risk of material misstatement due to fraud when Evidence is more (less) indicative of fraud, but only if they have stronger WTDs.22

Investigation of RQ2

RQ2 has two parts addressed by Experiment 1, with RQ2.1 focusing on stronger WTD participants and RQ2.2 focusing on weaker WTD participants. RQ2.1 asks: “Does a doubtful

22 Our inferences are robust to using a continuous rather than median-split specification based on a measure of WTD strength that uses the sum of squared deviations. They also are robust to both a continuous and median-split specification based on a measure of WTD strength that uses the sum of linear deviations.
prompt impair the propensity for auditors with relatively stronger WTDs to make judgments and decisions that reflect a higher (lower) risk of fraud when the underlying audit evidence is more (less) indicative of fraud?” Table 3 reports a two-way ANOVA with Evidence, Prompt and their interaction as independent factors on stronger WTD auditors’ risk assessments. Panel A of Table 3 examines the assessed likelihood of fraud (Attempted Fraud), and panel B of Table 3 examines the differences between the assessed likelihood of Attempted Fraud and the average of the assessed likelihoods of Within-GAAP Earnings Management and High-Quality Attempt (Fraud vs. Not Misstated).

In both two-way ANOVAs, Evidence is highly significant (both ps=0.001), and neither Prompt nor Evidence×Prompt is close to being significant. Further, in both two-way ANOVAs, the Evidence main effect accounts for at least 97.9 percent of the total between-cell variance while the Prompt main effect and the Prompt×Evidence interaction account for a total of 2 percent or less of the total between-cells variance.23 To further investigate RQ2.1, we examine all the simple main effects associated with this ANOVA, reporting p-values for the Fisher’s Least-Significant-Difference Test (which does not control for family-wise error rates) and for Tukey’s Honestly-Significant-Difference Test (which is more conservative in that it controls for family-wise error rate). The simple effect of Evidence is statistically significant for both the neutral and the doubtful levels of skepticism Prompt for the Fisher’s LSD (highest p(one-tailed)=0.013) and at least moderately statistically significant using the conservative Tukey’s HSD test (highest p(one-tailed) =0.057). These results indicate that, regardless of skepticism Prompt level, stronger WTD auditors’ express greater (lesser) concern about fraud risk when Evidence is more (less) indicative of fraud.

23 Specifically, in panel A, Table 3, the SS for Evidence 4978.63 while the total between-cells variance SS is 5083.35, and 4978.63/5083.35 = 97.9%. In panel B, Table 3 the SS for Evidence is 7231.88, while the total between-cells variance SS is 7328.89, and 7231.88/7328.89 = 98.7%.
Next, we investigate RQ2.2 which asks, “Does a neutral prompt enhance the propensity for auditors with relatively weaker WTDs to reach judgments and decisions that reflect a higher (lower) risk of fraud when the underlying audit evidence is more (less) indicative of fraud? Table 4 reports a two-way ANOVA with Evidence, Prompt and their interaction as independent factors on weaker WTD auditors’ risk assessments. Panel A of Table 4 examines the assessed likelihood of fraud (Attempted Fraud), and panel B of Table 4 examines the differences between the assessed likelihood of Attempted Fraud and the average of the assessed likelihoods of Within-GAAP Earnings Management and High-Quality Attempt (Fraud vs. Not Misstated).

In both panel A and B of Table 4, Prompt is the only factor that is significant (highest p=0.014). Evidence is not significant (Attempted Fraud p=0.190; Fraud vs. Not Misstated p=0.488), and neither is the Evidence × Prompt interaction (Attempted Fraud p=0.230, Fraud vs. Not Misstated p=0.231). Further, in both two-way ANOVAs, the Evidence main effect accounts for only 18.2% (Attempted Fraud) and 4.4% (Fraud vs. Not Misstated) of the total between-cell variance. To further investigate RQ2.2, we examine all of the simple main effects associated with this ANOVA, again reporting p-values for the Fisher’s Least-Significant-Difference Test as well as for Tukey’s Honestly-Significant-Difference Test. While the simple effect of Evidence clearly fails to attain statistical significance at an alpha of 0.05 for both dependent variables at the doubtful level of the skepticism Prompt for Fisher’s LSD (lowest p=0.717), it approaches significance at the neutral level of the skepticism Prompt for Attempted Fraud in panel A (p=0.078). However, this finding is diminished in that Evidence fails to obtain significance at

24 Specifically, in panel A, Table 4, the SS for Evidence is 625.80 while the total between-cells variance SS is 3444.98, and 625.8/3444.98 = 18.2%. In panel B, Table 4 the SS for Evidence is 245.04 while the total between-cells variance SS is 5532.65, and 245.04/5532.65 = 4.4%.
conventional levels for the *Fraud vs. Not Misstated* dependent measure in panel B (p=0.183). Further, using the Tukey HSD test, no *Evidence* effect is significant (lowest p=0.286).

What we observe instead is that the skepticism *Prompt* affects how highly auditors with weaker *WTDs* assessed fraud risk, as they assess it higher when reminded to be doubtful than when reminded to be neutral. Notably, the *Prompt* main effect is significant in each of the two-way ANOVAs without any *Evidence*×*Prompt* interaction. The exploratory simple effects analysis, however, reveals the only simple effect of *Prompt* that attains significance is when weaker *WTD* auditors encountered less indicative *Evidence*. Overall, these results provide only limited evidence that our neutral skepticism prompt can be helpful to auditors with weaker WTDs.

5. Experiment 2

In Experiment 1, we find that, as hypothesized, auditors evaluate *Evidence* objectively, but only when they have stronger *WTDs*. However, in all of the Experiment 1 conditions auditors were prompted to exercise professional skepticism. Thus, an alternative interpretation of Experiment 1 findings is that the objective evaluation of evidence requires a *combination* of stronger WTDs and a skepticism prompt, rather than stronger WTDs only. Experiment 2 thus tests our hypothesis without use of skepticism prompts. We also designed the new experiment such that its dependent measures go beyond the assessment of fraud risk (i.e., a judgment) to include action plans to address fraud risk (i.e., a choice). We wanted to do so because, unless auditors undertake different actions to address fraud risk, the chance of fraud detection is unlikely to change.

In Experiment 2 the participants have less audit experience and there are new case materials and new dependent variables which span both the assessment and plans to address fraud risk, all of which subject our hypothesis to a reproducibility test. Graduate students with an average
(standard deviation) age of 26.9 (8.2) from a large state university participated. These participants had completed an audit internship and/or were enrolled in a masters-level auditing course. We adapt experimental materials from two prior studies (Griffith et al. 2015; Kadous and Zhou 2019). Participants assess and address the risk that management is using unreasonable (overly aggressive) revenue projections to support their assertion that goodwill is properly valued (not impaired). Participants view the following audit evidence: summary of discussions with management regarding the revenue projections, historical revenue projections, peer company revenue projections, a sensitivity analysis, and market and industry information. Next, participants assess the reasonableness of management’s revenue growth projections and choose an action to address concerns about its reasonableness. Last, participants complete a post-test, including WTD items.

We use a 2×2 between-subjects design and manipulate Evidence at more indicative vs. less indicative of overly aggressive revenue projections. In the more indicative condition, for example, management’s projections are more favorable than those of peer firms, while in the less indicative condition, they are in line with those of peer firms (see Exhibit 5 for other details). The second factor is a measured variable—that is, a median split of participants’ WTDs (stronger vs. weaker).

For dependent variables, participants assess the reasonableness of management’s revenue projections and address misstatement risk arising from management’s revenue projections by choosing a Next Action. Our focus is whether the participants are sufficiently unsure about the reasonableness of revenue projections to warrant a discussion of their concerns with their senior

---

25 We collected Experiment 2 data in 2019, and Experiment 1 data in 2004. The latter are still relevant today as theory provides no reason to expect the date of data collection to alter the WTD x Evidence interaction, and audit studies during the intervening period continue to demonstrate that, on average, auditors struggle to evaluate evidence objectively (e.g., Bauer et al. 2020; Griffith et al. 2015; Glover et al. 2017; Joe et al. 2017).

26 The auditee (an electronics manufacturer) uses a discounted cash flow model to estimate the fair value of its United States reporting unit. Participants are informed that other members of the audit team have evaluated and determined to be reasonable the operating expense projections, capital expenditure projections, and the model’s discount rate.
associate. Following prior research (Griffith et al. 2015; Kadous and Zhou 2019), we code chosen actions as more appropriate or less appropriate. Figure 2 lists the five alternative Next Actions from which participants chose. Note that these listed Next Actions provide additional insight on how highly participants assess the risk of intentional misstatement and also their choice from alternative actions to address this risk. Two of these responses are more appropriate for more indicative Evidence (i.e., the third or fourth option), and two others are more appropriate responses for less indicative Evidence (i.e., the first or second option). We infer more objective evaluation of evidence in assessing and addressing fraud risk if participants choose more appropriate responses. Hypothesis 1 predicts that auditors will choose objectively more appropriate actions based on Evidence, but only if they have stronger WTDs. We perform no manipulation check on Evidence as the findings below demonstrate a successful manipulation (Sigall and Mills 1998, 221).

**RQ1 variation in the strength of auditors’ WTD**

We first investigate RQ1, again observing considerable variation in the strength of auditors’ WTDs, ranging from 32 to 115 (untabulated mean 66.30, standard deviation = 19.33). Of the 71 participants, 39 (54.9%) have prior audit experience, and of these 39, there is an average (standard deviation) of 0.9 (2.1) years of audit experience. Audit experience is not correlated with participants’ WTD score (untabulated r= +0.023, p=0.849). Further, the average WTD strength

---

27 The fifth option involves concluding that the revenue projections are unreasonable without indicating that a follow up discussion with the senior associate is necessary. Consistent with our expectation, few participants (two) chose this option. We code this as a more (less) appropriate action for more (less) indicative Evidence given that respondents who choose this option express concern about the reasonableness of the revenue projections. However, untabulated analysis shows that statistical inferences would not change were we to instead code this option as less appropriate for more indicative Evidence.

28 We asked participants to report age and GPAs. Neither age nor GPA affect our inferences when included in our hypothesis tests. Untabulated analysis shows age is weakly positively correlated with WTD strength (Pearson r = +0.25, Bartlett Chi-Square(1) = 4.08, p = 0.043), but GPA is not significantly correlated with WTD strength (Pearson r = -0.13, Bartlett Chi-Square(1) = 1.18, p = 0.277). We did not ask participants to identify their gender or ethnicity.
does not differ statistically with respect to Experiment 1’s audit senior participants and the less experienced auditors in this experiment (untabulated 68.53 versus 66.30, t(156)=0.711, p=0.478). We again use a median split to classify auditors as having stronger WTDs or weaker WTDs to enable the use of planned contrasts. Consistent with random assignment, WTD strength does not differ across Evidence conditions (untabulated t(69) = 0.23, p=0.817). Descriptive statistics (untabulated) for auditors with stronger WTDs are mean=50.24 and standard deviation=7.87, and for auditors with weaker WTDs, mean=81.05 and standard deviation 14.24, in line with means and standard deviations in Experiment 1.

**Test of Hypothesis 1’s reproducibility**

Table 5 depicts how Evidence and WTD strength jointly influence auditors’ Revenue Projection Reasonableness Assessments. Participants responded to the following question: “Based on your evaluation, how likely is it that the five-year revenue projections are reasonable?” on an 11-point scale (0 – Not likely at all, 5 – As likely as not, 10 – Extremely likely). We observe no significant effects in the two-way ANOVA (lowest p=0.272). Instead, most participants express high uncertainty about the reasonableness of these projections. When asked “… how likely is it that the five-year revenue projections are reasonable?” the average response falls between 5 and 6. In fact, over half (80%) of the participants respond within 1 point (2 points) of the midpoint 5 “As Likely as not” on the 11-point scale.

The more critical dependent variable asks participants how they would address concerns about the reasonableness of management’s revenue projections. Figure 2 and Table 5 illustrate and provide results indicating a significant interaction between WTD strength and Evidence (Z = 2.07,
Panel B of Table 5 also presents the planned contrast using weights of 0 less indicative Evidence, weaker WTD; 0 more indicative Evidence, weaker WTD; -1 less indicative Evidence, stronger WTD; and +1 more indicative Evidence, stronger WTD. As in Experiment 1, these contrast weights fit the predicted pattern: auditors respond to Evidence in their choice of additional procedures but only when they have a stronger WTD. The visual fit predominantly supports Hypothesis 1, as does the significant contrast test (Z = 1.84, p = 0.033, one-tailed), the insignificant semi-omnibus test (untabulated $\chi^2 = 0.54$, p = 0.763), and relatively small contrast variance residual ($q^2$) of 26.9% (Guggenmos et al. 2018).

Importantly, simple effects at the bottom of panel B of Table 5 show that stronger WTD auditors follow up with their senior associate more often when Evidence suggests that management’s revenue projections are overly aggressive (p = 0.020, one-tailed), but that weaker WTD auditors do not do so (p = 0.331, two-tailed). From an audit effectiveness perspective, it is notable that the simple effect of stronger vs. weaker WTD is significant given Evidence that is more indicative of fraud (p = 0.020, one-tailed) Collectively, these findings demonstrate the reproducibility of our hypothesis, provide evidence that it holds in the absence of skepticism prompts (RQ2.3), and strengthen the validity of the WTD construct. Having shown that auditors

---

29 Untabulated analysis shows that the interaction remains significant when using a continuous measure of the strength of participants’ WTDs (p=0.009, one-tailed). Also, the same inferences hold if we include participants’ assessed reasonableness of revenue projections as a covariate.

30 If anything, weaker WTD auditors follow up with their senior associate less often facing Evidence suggesting that management’s revenue projections are overly aggressive, and this dysfunctionality increases the size of the contrast variance residual, $q^2$. However, this simple effect is not statistically significant, which does match our planned contrast weights (0 less indicative Evidence, weaker WTD; 0 more indicative Evidence, weaker WTD). Clearly, auditors respond less appropriately to Evidence when they have weaker WTDs versus stronger WTDs.

31 Untabulated analysis shows our inferences are robust to using a continuous rather than median-split specification based on a measure of WTD strength that uses the sum of squared deviations. They also are robust to both a continuous and median-split specification based on a measure of WTD strength that uses the sum of linear deviations.
(fail to) objectively evaluate evidence if they have stronger (weaker) WTDs, we next probe whether WTD coheres as a multifaceted construct and diverges from Hurtt Trait Skepticism (HTS).

6. Further analyses to assess WTD construct and divergent validity

To further validate the multifaceted nature of our WTD construct, we perform a confirmatory bi-factor analysis. A bi-factor model is a hierarchical factor model that determines whether a construct includes a higher-order general factor that captures a commonality threading among two or more lower-order subfactors. In selecting each of our 22 scale items, we anticipated that they would entail a higher-order general wise thinking factor in addition to a small number of subfactors (see, e.g., Chen et al. 2012). Of these 22 items, we categorize 10 into a subfactor that we call reflective thinking, 10 into one that we call evidence-based beliefs & doubts, and 2 into a subfactor that we call counterfactual thinking. Other subfactors could exist, but we keep our model simple, as is appropriate for new constructs (e.g., Churchill 1979).

Because empirical assessment of construct validity using bi-factor analyses is more powerful and stable with larger sample sizes, we use 298 participants by pooling the 87 audit seniors in Experiment 1, the 71 graduate audit students in Experiment 2, and an additional 140 undergraduate participants in Experiment 3, who completed an exploratory study during fall 2015 in which they fill out both the WTD and HTS questionnaires. Pooling across these three groups is appropriate conceptually and pragmatically. Conceptually, the WTD construct is invariant across different adults; its meaning does not differ between audit seniors and audit staff. Indeed, nearly identical WTD means and standard deviations obtain across these groups (untabulated 68.19 [19.82] Experiment 1; 66.81 [19.34] Experiment 2; and 66.51 [19.03] Experiment 3).

As reported in supporting information in the online Appendix, the model fit measures exceed conventional cutoffs for a good fit (Boateng et al. 2018; Kenny 2020), and 11 of our 22
items (50%) load on a higher-order wise thinking (general) factor. Additionally, 7 of the 10 items categorized as reflective thinking load on that subfactor, 8 of the 10 items categorized as evidence-based beliefs items load on that subfactor, and one of the two items categorized as counterfactual thinking items loads on that subfactor. Overall, the confirmatory bi-factor analysis supports the conclusion that our scale’s 22 items collectively capture a multifaceted wise-thinking disposition construct consisting of a general factor as well as three subfactors.32

Last, having already shown WTD strength only is very weakly correlated with tacit managerial knowledge (see footnote 12), we next examine if a distinction can be drawn between the WTD construct and the Hurtt Trait Skepticism (HTS) construct. Prior research shows that certain advantages are associated with having higher HTS. McAllister et al. (2020) find that having a higher HTS auditor on an audit team helps other team members consider case information they otherwise overlook. Further, HTS scores are positively correlated with a longer auditing career trajectory (Cohen et al. 2017). Nonetheless, trait skepticism does not fully capture professional skepticism in that it largely omits situational skepticism. Situational skepticism captures the idea that auditors should heighten their skepticism in some situations as compared to others (e.g., Nolder and Kadous 2018). Counter to this idea, Quadackers et al. (2014) find that regardless of their HTS score, auditors assess fraud risk to be higher in response to a stark manipulation by which management either: aggressively manages earnings, frequently gets into disputes with external auditors, and emphasizes productivity; or strives to report accurately, works harmoniously with external auditors, and emphasizes integrity. Auditors with higher versus lower HTS scores

32 We do not rely on Cronbach’s alpha as a metric for reliability for WTD. Rodriguez et al. (2016, 140) emphasize that, “…When the data are multidimensional, as in a bifactor model, alpha … loses its appropriateness as an indicator of how well a total or subscale score reflects a single latent variable.” For completeness, our Cronbach’s alpha is 0.90, well above standard cutoffs. A study of wisdom measures, including the Ardelt scale, reports Cronbach alphas between 0.47 and 0.88 (Glück et al. 2013).
also assess fraud risk to be higher *regardless* of which management type they face. In other studies, higher HTS scores reportedly have no significant effect on auditors’ evaluation of evidence (see, e.g., Rasso 2015, 47-49). In short, we know of no study that reports findings consistent with higher HTS scores enabling auditors to more objectively evaluate evidence. Thus, the observed effects of higher HTS on auditors’ evidence evaluation diverges from that which we observe in two separate experiments herein. We find that auditors objectively evaluate Evidence that is less versus more indicative of potential fraud when assessing fraud risk and planning follow-up procedures to address fraud risk, but *only* when they have stronger WTDs. As such, higher HTS scores and stronger WTD scores would appear to be capturing decidedly different constructs.

Nonetheless, to further investigate the possibility that HTS and WTD capture essentially identical constructs we perform further untabulated analyses using undergraduate senior participants from a large state university from the aforementioned Experiment 3. They completed our 22-item WTD scale and the 30-item HTS scale in a post-test. Four participants failed to fill out all measures, leaving 136 observations. To begin, we simply regress participants’ HTS scores on their WTD scores. Unsurprisingly, HTS scores explain some of the variation in WTD ($R^2=21\%$, $p < 0.001$), but the larger point is that they fail to explain most of the variation in WTD (i.e., 79\%, $1 - R^2$).\(^{33}\)

We next conduct a factor analysis on 52 items, 22 from the WTD scale and 30 from the HTS scale. We use promax rotation, which allows extracted factors to be correlated. Following

\(^{33}\) For this regression, we use the sum of linear deviations for both WTD and HTS to increase the chance of finding a stronger linear association. Hurtt (2010) explains that six factors underlie HTS: *Search for Knowledge, Suspension of Judgment, Interpersonal Understanding, Self-Determining, Self-Confidence*, and *Questioning Mind*. Even larger percentages of the variation in WTD scores are unexplained by each of the six factors that comprise HTS, as follows: *Search for Knowledge* 80\%, *Suspension of Judgment* 89\%, *Interpersonal Understanding* 90\%, *Self-Confidence* 92\%, *Self-Determining* 94\%, and *Questioning Mind* 96\%.
convention (Boateng et al. 2018; Nunnally 1978), we extract 10 factors as they have loadings of an absolute value of at least 0.40 (see online Appendix). Of these 10, 6 factors load exclusively on HTS items. Further, the remaining four factors load exclusively on WTD items. Finally, only 1 of the 24 inter-factor correlations between the 6 HTS factors and the 4 WTD factors is even moderately strongly correlated (i.e., $r \geq +0.40$, at +0.44). Of the 23 remaining inter-factor correlations, nine are weakly correlated (i.e., $+0.20 \leq r \leq +0.39$) and 14 are very weakly correlated ($+0 \leq r \leq +0.19$). We conclude that HTS and WTD readily diverge from one another empirically.

7. Concluding comments

Audit theory and standards prescribe that auditors should evaluate evidence objectively and describe objective evaluation as indispensable for the proper exercise professional skepticism. However, both anecdotal inspection findings and empirical findings from numerous prior experimental studies indicate that auditors, at least on average, fail to objectively evaluate evidence. This paper contributes to the literature by developing the hypothesis that auditors objectively evaluate evidence, provided that they have relatively strong WTDs. We find support for this hypothesis across two experiments that differ with regard to inclusion or exclusion of a prompt to exercise professional skepticism. The experiments also use different audit contexts that feature potential fraud as well as different dependent variables and different participants. In both experiments, auditors objectively evaluate whether evidence is more or less indicative of fraud only when they have stronger WTDs, demonstrating the reproducibility of the audit judgment-quality benefits of stronger WTDs. This finding has potential implications for standard setters and

34 The 6-factor structure of HTS discussed in the prior footnote largely replicates in our data. Factor one includes all 5 Self-Confidence items, factor two includes 5 of the 6 Search for Knowledge items, factor three includes all 5 Interpersonal Understanding items, factor four includes all 5 Self-Determining items and 2 Questioning Mind items, factor five includes all 5 Suspension of Judgment items, and factor ten is one Interpersonal Understanding item.

35 In Experiment 3, we use Experiment 1 case materials and manipulate skepticism Prompt, but hold Evidence constant. Since Evidence is not manipulated, this experiment is not designed to provide data useful for testing our hypothesis.
audit firms. In particular, quality control standards and audit working paper review processes might benefit from a revision that takes into account our theory-consistent findings that auditors do not objectively evaluate evidence unless they have stronger WTDs.

A second contribution is identification of WTDs as a new construct, defining it as a thinking disposition by which individuals naturally engage in balanced revision of prior beliefs and doubts about target phenomena by thinking openly and reflectively about available evidence (Meacham 1983; Kitchener and Brenner 1990; Ardelt 2003; Staudinger and Glück 2011a). We further contribute to the literature by showing that both inexperienced and experienced auditors vary considerably in terms of the strength of their WTD. Finally, we perform confirmatory bi-factor analyses to demonstrate that a general factor threads through our 22 items used to measure WTDs, and we differentiate WTD from Hurtt’s Trait Skepticism (HTS). These findings have potential practical implications in that weaker WTD auditors likely have difficulty responsibly exercising professional skepticism, as they struggle to objectively evaluate evidence.

We close by considering opportunities to extend this research. While we introduce the WTD construct to the accounting literature with an aim of showing that stronger WTDs enable auditors to objectively evaluate evidence, stronger WTDs likely also improve other facets of auditor task performance. Do stronger WTDs help auditors to allocate scarce cognitive or budgetary resources while performing complex tasks, reduce auditor overconfidence, or enable fuller learning from past deficiencies in professional judgments? As WTDs are one component of a rich cognitive architecture, researchers could examine how WTDs interact with other forms of auditor knowledge (e.g., technical or industry) or practical intelligence (e.g., tacit knowledge, systems thinking, emotional intelligence) to improve auditor judgments and decisions. To build on our finding that WTD strength varies significantly among audit staff and seniors, future studies
can examine how WTD strength varies within and across audit managers and partners, as well as within and across auditors receiving more favorable or less favorable performance evaluations. Consequently, to further our understanding of how audit teams interact, one could examine under what conditions, if any, that auditors with stronger WTDs compensate for other team members with weaker WTDs. Related future research could examine instructional or enculturation methods that strengthen auditors’ WTDs as well as endeavor to identify better measures of WTD strength. As an example, one could use a multi-method approach by pairing our WTD survey with participants’ reactions to scenarios depicting complicated life problems and solicit participants’ goals, beliefs, doubts, and possible courses of action. Trained judges could then code participants’ responses (Staudinger and Glück 2011b). While our studies conveniently provided auditors with evidence of fraud, future investigations could examine whether stronger WTDs improve auditors’ evidence search or mitigate over-aversion or over-eagerness to use AI-based audit evidence. Finally, future research could examine whether the failure of auditors with weaker WTDs to objectively evaluate evidence can be overcome by interventions that induce deliberative mindsets such as situational prompts or advisor roles (see Griffith et al. 2015 or Bauer et al. 2020, respectively).
REFERENCES

Achor, S. 2010. *The happiness advantage: How a positive brain fuels success in work and life*. New York: Crown Publishing Group.

AICPA. 1988. *Statement on Auditing Standard No. 53: The Auditor’s Responsibility to Detect and Report Errors and Irregularities*. New York: AICPA.

AICPA. 2006. *Statement on Auditing Standard No. 99: Consideration of Fraud in a Financial Statement Audit*. New York: AICPA.

AICPA. 2014. *Overall Objectives of the Independent Auditor*. AU-C Section 200. New York: AICPA.

AICPA. 2014. *Consideration of Fraud in a Financial Statement Audit*. AU-C Section 240. New York: AICPA.

AICPA. 2014. *Performing Audit Procedures in Response to Assessed Risks and Evaluating the Audit Evidence Obtained*. AU-C Section 330. New York: AICPA.

AICPA. 2014. *Audit Evidence*. AU-C Section 500. New York: AICPA.

Ardelt, M. 2003. Empirical assessment of a three-dimensional wisdom scale. *Research on Aging* 25 (3): 275-324.

Ardelt, M. 2004. Wisdom as expert knowledge system: A critical review of a contemporary operationalization of an ancient concept. *Human Development* 47 (5): 257-85.

Ardelt, M. 2015. Correlates of wisdom. In *Encyclopedia of Adulthood and Aging*, 1–5. John Wiley & Sons, Inc.

Asay, H. S., R. Guggenmos, K. Kadous, L. Koonce, and R. Libby. 2021. Theory testing and process evidence in accounting experiments. Working paper, University of Iowa, Cornell University, Emory University, and University of Texas at Austin.

Baron, J. 1993. Why teach thinking? – An essay. *Applied Psychology: An International Review* 42 (3): 191-237.

Baron, J. 2008. *Thinking and deciding*, 4th ed. New York: Cambridge University Press.

Bauer, T. D., S. Hillison, B. Pomeroy, and M. E. Peecher. 2020. Revising audit plans to address fraud risk: “A case of do as I advise, not as I do”? *Contemporary Accounting Research* 37 (4): 2558-89.
Bell, T. B., M. E. Peecher, and I. Solomon. 2005. *The 21st century public company audit: Conceptual elements of KPMG’s global audit methodology*. Montvale, New Jersey: KPMG LLP.

Ben-Shahar, T. 2007. *Happier: Learn the secrets to daily joy and lasting fulfillment*. New York: McGraw-Hill.

Bhaskar, L. S., P. E. Hopkins, and J. H. Schroeder. 2019. An investigation of auditor judgment when companies release earnings before audit completion. *Journal of Accounting Research* 57 (2): 355-90.

Birren, J. E. and L. M. Fisher. 1990. Conceptualizing wisdom: The primacy of affect-cognition relations. In *Wisdom: Its Nature, Origins, and Development*, edited by R.J. Sternberg, 317-32. New York: Cambridge University Press.

Birren, J. E. and C. M. Svensson. 2005. Wisdom in history. In *A Handbook of Wisdom: Psychological Perspectives*, edited by R.J. Sternberg and J. Jordan, 3-31. New York: Cambridge University Press.

Boateng, G. O., T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, and S. L. Young. 2018. Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health* 6 (149): 1-18.

Bol, J. C., C. Estep, F. Moers, and M. E. Peecher. 2018. The role of tacit knowledge in auditor expertise and human capital development. *Journal of Accounting Research* 56 (4): 1205-52.

Bowlin, K. 2011. Risk-based auditing, strategic prompts, and auditor sensitivity to fraud risk. *The Accounting Review* 86 (4): 1231-53.

Brown, C. E., M. E. Peecher, and I. Solomon. 1999. Auditors’ hypothesis testing in diagnostic inference tasks. *Journal of Accounting Research* 37 (1): 1-26.

Cacioppo, J. T., R. E. Petty, J. A. Feinstein, and W. B. G. Jarvis. 1996. Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin* 119 (2): 197-253.

Chen, F. F., A. Hayes, C. S. Carver, J. Laurenceau, and Z. Zhang. 2012. Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality* 80 (1): 219-51.

Churchill, G. A. 1979. A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research* 16 (1): 64-73.

Cohen, J. R., D. W. Dalton, and N. L. Harp. 2017. Neutral and presumptive doubt perspectives of professional skepticism and auditor job outcomes. *Accounting, Organizations and Society* 62: 1-20.
Davies, P. J. 2020. Wirecard CEO resigns after banks say missing $2 billion doesn’t exist. *Wall Street Journal*, June 19. https://www.wsj.com/articles/mystery-surrounding-wirecards-missing-2-billion-deepens-as-banks-say-they-don’t-have-it-11592562966

Davis, M. H. 1980. A multidimensional approach to individual differences in empathy. In *JSAS Catalog of Selected Documents in Psychology* 10. Washington, DC: American Psychological Association.

Dikolli, S. S., T. Keusch, W. J. Mayew, and T. D. Steffen. 2020. CEO behavioral integrity, auditor responses, and firm outcomes. *The Accounting Review* 95 (2): 61-88.

Epstein, S. and P. Meier. 1989. Constructive thinking: A broad coping variable with specific components. *Journal of Personality and Social Psychology* 57 (2): 332-50.

Fuller, L. R., and S. E. Kaplan. 2004. A note about the effect of auditor cognitive style on task performance. *Behavioral Research in Accounting* 16 (1): 131–43.

Glück, J., S. Koing, K. Naschenweng, W. Redzanowski, L. Dorner, I. Straber, and W. Wiedermann. 2013. How to measure wisdom: Content, reliability, and validity of five measures. *Frontiers in Psychology* 4 (405): 1-13.

Glover, S. M., M. H. Taylor and Y. Wu. 2017. Current practices and challenges in auditing fair value measurements and complex estimates: Implications for audit standards and the academy. *Auditing: A Journal of Practice & Theory* 36 (1): 63-84.

Goodwin, J. 1999. The effects of source integrity and consistency of evidence on auditors’ judgments. *Auditing: A Journal of Practice & Theory* 18 (2): 1-16.

Grenier, J. H. 2017. Encouraging professional skepticism in the industry specialization era. *Journal of Business Ethics* 142 (2): 241-56.

Griffith, E. E., J. S. Hammersley, K. Kadous, and D. Young. 2015. Auditor mindsets and audits of complex estimates. *Journal of Accounting Research* 53 (1): 49-77.

Guggenmos, R. D., M. Piercey, and C. Agoglia. 2018. Custom contrast testing: Current trends and a new approach. *The Accounting Review* 93 (5): 223-44.

Hail, L., M. Lang, and C. Leuz. 2020. Reproducibility in accounting research: Views of the research community. *Journal of Accounting Research* 58 (2): 519-43.

Hall, S. S. 2010. *Wisdom: From philosophy and neuroscience*. New York: Vintage Books.

Hammersley, J. S., E. M. Bamber, and T. D. Carpenter. 2010. The influence of documentation specificity and priming on auditors’ fraud risk assessments and evidence evaluation decisions. *The Accounting Review* 85 (2): 547-71.
Hammersley, J. S., K. Johnstone, and K. Kadous. 2011. How do audit seniors respond to heightened fraud risk? *Auditing: A Journal of Practice & Theory* 30 (3): 81-101.

Heppner, P. P., and C. H. Petersen. 1982. The development and implications of a personal problem-solving inventory. *Journal of Counseling Psychology* 29 (1): 66-75.

Hobson, J. L., W. J. Mayew, M. E. Peecher, and M. Venkatachalam. 2017. Improving experienced auditors’ detection of deception in CEO narratives. *Journal of Accounting Research* 55 (5): 1137-66.

Hoffman, V. B., and M. F. Zimbelman. 2009. Do strategic reasoning and brainstorming help auditors change their standard audit procedures in response to fraud risk? *The Accounting Review* 84 (3): 811-37.

Hurtt, K. R. 2010. Development of a scale to measure professional skepticism. *Auditing: A Journal of Practice & Theory* 29 (1): 149-71.

IIA (Institute of Internal Auditors). 2021. *Code of Ethics*, [https://na.theiia.org/standards-guidance/mandatory-guidance/Pages/Code-of-Ethics.aspx](https://na.theiia.org/standards-guidance/mandatory-guidance/Pages/Code-of-Ethics.aspx)

Iacobucci, D., S. S. Posavac, F. R. Kardes, M. J. Schneider, D. L. Popovich. 2015. Towards a more nuanced understanding of the statistical properties of a median split. *Journal of Consumer Psychology* 25 (4): 652-65.

Joe, J. R., S. D. Vandervelde, and Y. Wu. 2017. Use of high quantification evidence in fair value audits: Do auditors stay in their comfort zones? *The Accounting Review* 92 (5): 89-116.

Kachelmeier, S. J., T. Majors, and M. G. Williamson. 2014. Does intent modify risk-based auditing? *The Accounting Review* 89 (6): 2181-201.

Kadous, K., S. Kennedy, and M. E. Peecher. 2003. The effect of quality assessment and directional goal commitment on auditors’ acceptance of client-preferred accounting methods. *The Accounting Review* 78 (3): 759-78.

Kadous, K., and Y. Zhou. 2019. How does intrinsic motivation improve auditor judgment in complex audit tasks? *Contemporary Accounting Research* 36 (1): 108-31.

Kenny, D., 2020. Measuring model fit. [http://davidakenny.net/cm/fit.htm](http://davidakenny.net/cm/fit.htm)

Kim, P. H., D. L. Ferrin, C. D. Cooper, and K. T. Dirks. 2004. Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology* 89 (1): 104–18.

King, M. B. and R. A. Hunt. 1975. Measuring the religious variable: National replication. *Journal for the Scientific Study of Religion* 14 (1): 13-22.
Kitchener, K. S., and H. G. Brenner. 1990. Wisdom and reflective judgment. In *Wisdom: Its Nature, Origins, and Development*, edited by R.J. Sternberg, 212-29. New York: Cambridge University Press.

Kizirian, T. G., B. W. Mayhew, and D. Sneathen Jr. 2005. The impact of management integrity on audit planning and evidence. *Auditing: A Journal of Practice & Theory* 24 (2): 49-67.

Knechel, W. R. 2013. Do auditing standards matter? *Current Issues in Auditing* 7 (2): A1-A16.

Knechel, W. R., G. V. Krishnan, M. Pevzner, L. B. Shefchik, and U. K. Velury. 2013. Audit quality: Insights from the academic literature. *Auditing: A Journal of Practice & Theory* 32 (Supplement 1): 385-421.

Kochetova-Kozloski, N., W. F. Messier, and A. Eilifesen. 2011. Improving auditors’ fraud judgments using a frequency response mode. *Contemporary Accounting Research* 28 (3): 837-58.

Kowsmann, P., M. Maurer, and J. Yang. 2020. String of firms that imploded have something in common: Ernst & Young audited them. *Wall Street Journal*, October 16, https://www.wsj.com/articles/string-of-firms-that-imploded-have-something-in-common-ernst-young-audited-them-11602863319

Macdonald, A. P. 1970. Revised scale for ambiguity tolerance: Reliability and validity. *Psychological Reports* 26 (3): 791-98.

McAllister, M., A. Blay, and K. Kadous. 2020. Fraud brainstorming group composition in auditing: The persuasive power of a skeptical minority. *The Accounting Review*, https://doi.org/10.2308/TAR-2018-0027

McClelland, G. E., J. G. Lynch, J. R. Irwin, S. A. Spiller, and G. J. Fitzsimmons. 2015. Median splits, type II errors, and false-positive consumer psychology: Don’t fight the power. *Journal of Consumer Psychology* 25 (4): 679-89.

Meacham, J. A. 1983. Wisdom and the context of knowledge: Knowing that one doesn’t know. In *On the Development of Developmental Psychology*, edited by D. Kuhn and J. A. Meacham, 111-34. Basel, Switzerland: Karger.

Meacham, J. A. 1990. The loss of wisdom. In *Wisdom: Its Nature, Origins, and Development*, edited by R. J. Sternberg, 181-211. New York: Cambridge University Press.

Nagarajan, S. 2020. EY’s chairman expressed “regret” that German fintech Wirecard’s fraudulent practices were “not uncovered sooner”. *Business Insider*, September 15, https://markets.businessinsider.com/news/stocks/ey-chairman-expresses-regret-over-wirecard-audit-failures-2020-9-1029591270?op=1#
Nelson, M. W. 2009. A model and literature review of professional skepticism in auditing. *Auditing: A Journal of Practice & Theory*, 28 (2): 1-34.

Nelson, M. W. and H.-T. Tan. 2005. Judgment and decision making research in auditing: A task, person, and interpersonal interaction perspective. *Auditing: A Journal of Practice & Theory* 24 (s-1): 41-71.

Nicodemus, A. 2020. EY member firm fined $1.5M for failure to uncover $3B homebuilding fraud. *Compliance Week*, December 18, https://www.complianceweek.com/regulatory-enforcement/ey-member-firm-fined-15m-for-failure-to-uncover-3b-homebuilding-fraud/29856.article

Nolder, C. J. and K. Kadous. 2018. Grounding the professional skepticism construct in mindset and attitude theory: A way forward. *Accounting, Organizations and Society* 67: 1-14.

Nunnally, J. C. 1978. *Psychometric theory*, 2nd ed. New York: McGraw-Hill.

Paulhus, D. L. and D. B. Reid. 1991. Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology* 60 (2): 307-17.

Peecher, M. E. 1996. The influence of auditors’ justification processes on their decisions: A framework and experimental evidence. *Journal of Accounting Research* 34 (1): 125-40.

Peecher, M. E., R. Schwartz, and I. Solomon. 2007. It’s all about audit quality: Perspectives on strategic-systems auditing. *Accounting, Organizations and Society* 32 (4-5): 463-85.

Peecher, M. E., I. Solomon, and K. T. Trotman. 2013. An accountability framework for financial statement auditors and related research questions. *Accounting, Organizations and Society* 38 (8): 596-620.

Perkins, D. N., E. Jay, and S. Tishman. 1993. Beyond abilities: A dispositional theory of thinking. *Merrill-Palmer Quarterly* 39 (1): 1-21.

Perkins, D. N., S. Tishman, R. Ritchhart, K. Donis, and A. Andrade. 2000. Intelligence in the wild: A dispositional view of intellectual traits. *Educational Psychology Review* 12 (3): 269-93.

Peytcheva, M. 2014. Professional skepticism and auditor cognitive performance in a hypothesis-testing task. *Managerial Auditing Journal* 29 (1): 27-49.

PCAOB. 2010. *Due Professional Care in the Performance of Work*. Auditing Standard No. 1015.

PCAOB. 2010. *Audit Risk*. Auditing Standard No. 1101.

PCAOB. 2010. *Audit Evidence*. Auditing Standard 1105.

PCAOB. 2010. *Identifying and Assessing Risks of Material Misstatement*. Auditing Standard 2110.
PCAOB. 2010. *The Auditor’s Responses to the Risks of Material Misstatement*. Auditing Standard 2301.

PCAOB. 2010. *Consideration of Fraud in a Financial Statement Audit*. Auditing Standard 2401.

POB (Public Oversight Board). 2000. The panel on audit effectiveness: Report and recommendations. Stamford, CT: POB.

Quadackers, L., T. Groot, and A. Wright. 2014. Auditors’ professional skepticism: Neutrality versus presumptive doubt. *Contemporary Accounting Research* 31 (3): 639-57.

Rapoport, M. 2018. Judge says PricewaterhouseCoopers was negligent in Colonial Bank Failure: Ruling opens accounting firm to hundreds of millions of dollars in potential damages. *Wall Street Journal*, January 2, https://www.wsj.com/articles/judge-says-pricewaterhousecoopers-was-negligent-in-colonial-bank-failure-1514762610

Rasso, J. T. 2015. Construal instructions and professional skepticism in evaluating evidence. *Accounting, Organizations and Society* 46: 44-55.

Rodriguez, A., S. P. Reise, and M. G. Haviland. 2016. Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods* 21 (2): 137-50.

Rokeach, M. 1960. *The open and closed mind*. New York: Basic Books.

Rothenberg, N. R. 2020. Auditor reputation concerns, legal liability, and standards. *The Accounting Review* 95 (3): 371-91.

Sá, W. C., R. F. West, and K. E. Stanovich. 1999. The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology* 91 (3): 497-510.

Sigall, H., and J. Mills. 1998. Measures of independent variables and mediators are useful in social psychology experiments: But are they necessary? *Personality and Social Psychology Review* 2 (3): 218–26.

Stanovich, K. E. 2001. The rationality of educating for wisdom. *Educational Psychologist* 36 (4): 247-51.

Stanovich, K. E. 2009. *What intelligence tests miss: The psychology of rational thought*. New Haven, CT: Yale University Press.

Stanovich, K. E. and R. F. West. 1997. Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology* 89 (2): 342-57.
Stanovich, K. E. and R. F. West. 1998. Individual differences in rational thought. *Journal of Experimental Psychology: General* 127 (2): 161-88.

Stanovich, K. E. and R. F. West. 2007. Natural myside bias is independent of cognitive ability. *Thinking & Reasoning* 13 (3): 225-47.

Stanovich, K. E., and R. F. West. 2008. On the failure of cognitive ability to predict myside and one-sided thinking biases. *Thinking & Reasoning* 14 (2): 129-67.

Stanovich, K. E., R. F. West, and M. E. Toplak. 2018. *The rationality quotient: Towards a test of rational thinking*. Cambridge, MA: The MIT Press.

Staudinger, U. M., and J. Glück. 2011a. Intelligence and wisdom. In *The Cambridge Handbook of Intelligence*, edited by R. J. Sternberg and S. B. Kaufman, 827-46. New York: Cambridge University Press.

Staudinger, U. M., and J. Glück. 2011b. Psychological wisdom research: Commonalities and differences in a growing field. *Annual Review of Psychology* 62 (1): 215-41.

Staudinger, U. M., J. Smith, and P. B. Baltes. 1992. Wisdom-related knowledge in a life review task: Age differences and the role of professional specialization. *Psychology and Aging* 7 (2): 271-81.

Sternberg, R. J. 2001. Why schools should teach for wisdom: The balance theory of wisdom in educational settings. *Educational Psychologist* 36 (4): 227-45.

Sternberg, R. J. 2003. *Wisdom, intelligence, and creativity synthesized*. New York: Cambridge University Press.

Sternberg, R. J., and E. L. Grigorenko. 1997. Are cognitive styles still in style? *American Psychologist* 52 (7): 700-12.

Tan, H.-T., and R. Libby. 1997. Tacit managerial versus technical knowledge as determinants of audit expertise in the field. *Journal of Accounting Research* 35 (1): 97-113.

Tishman, S., E. Jay, and D. N. Perkins. 1993. Teaching thinking dispositions: From transmission to enculturation. *Theory into Practice* 32 (3): 147-53.

Tokar, D., and P. J. Davies. 2021. Wirecard red flags should have prompted earlier response, former executive says. *Wall Street Journal*, February 8, https://www.wsj.com/articles/wirecard-red-flags-should-have-prompted-earlier-response-former-executive-says-11612780200

Toplak, M. E., R. F. West, and K. E. Stanovich. 2011. The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory and Cognition* 39 (7): 1275-89.

Trompeter, G. M., T. D. Carpenter, N. Desai, K. L. Jones and R. A. Riley, Jr. 2013. A synthesis of fraud-related research. *Auditing: A Journal of Practice & Theory* 32 (Supplement 1): 287-321.

This article is protected by copyright. All rights reserved.
Trotman, K., and W. F. Wright. 2012. Triangulation of audit evidence in fraud risk assessments. *Accounting, Organizations and Society* 37 (1): 41-53.

Turner, C. W. 2001. Accountability demands and the nature of auditor’s evidence search strategy: The influence of reviewer preferences and the nature of the response (belief vs. action). *Journal of Accounting Research* 39 (3): 683-706.

Tversky, A., and D. Kahneman. 1982. Judgment and uncertainty: Heuristics and biases. In *Judgment under Uncertainty: Heuristics and Biases*, edited by D. Kahneman, P. Slovic, and A. Tversky, 3-22. New York: Cambridge University Press.

Wagner, R. K. 1987. Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology* 52 (6): 1236-47.

Waller, W. S., and M. F. Zimbelman. 2003. A cognitive footprint in archival data: Generalizing the dilution effect from laboratory to field settings. *Organizational Behavior and Human Decision Processes* 91 (2): 254-68.

West, R. F., and W. C. Sá. 1999. Thinking dispositions and decontextualized reasoning. In *Who is Rational? Studies of Individual Differences in Reasoning*, edited by K.E. Stanovich, 153-89. Mahwah, NJ: Lawrence Erlbaum Associates.

Wikimedia Commons. 2015. File:Pacman.svg, [https://commons.wikimedia.org/w/index.php?title=File:Pacman.svg&oldid=173501691](https://commons.wikimedia.org/w/index.php?title=File:Pacman.svg&oldid=173501691)

Zimbelman, M. F. 1997. The effects of SAS No. 82 on auditors’ attention to fraud risk factors and audit planning decisions. *Journal of Accounting Research* 35 (Supplement): 75-97.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of this article: **Online Appendix.** Supporting information
EXHIBIT 1

Evidence manipulation (Experiment 1)

Panel A: Less Indicative of Fraud

Audit team member’s rationale: Why change this reserve? AUP has increased this reserve steadily in the past four years. In response to my inquiry, management explains that the change reflects the fact that AUP’s production quality significantly surpassed its expectations this year. Management says re-work and returns have both been lower this year than both last year’s results and this year’s expectations. When I asked management whether other indicators would tell another story, management said, “No.”

While reviewing KPIs in two of AUP’s core business processes (production and customer fulfillment), I did find some corroborative evidence. Specifically, some KPIs support management’s statement in that this year’s returns and re-work were both lower (by about 30% compared to last year).

Still, I remain concerned that AUP’s real motivation for this entry may be to keep their earnings from falling short of analysts’ forecasts and that changing the reserve is simply a way to “generate” earnings.

Panel B: More Indicative of Fraud

Audit team member’s rationale: Why change this reserve? AUP has increased this reserve steadily in the past four years. In response to my inquiry, management explains that the change reflects the fact that AUP’s production quality significantly surpassed its expectations this year. Management says re-work and returns have both been lower this year than both last year’s results and this year’s expectations. When I asked management whether other indicators would tell another story, management said, “No.”

However, while reviewing KPIs in two of AUP’s core business processes (production and customer fulfillment), I found mixed evidence. While some KPIs support management’s statement in that this year’s returns and re-work were both lower (by about 30% compared to last year), other KPIs indicate that both assembly-worker overtime and production-line utilization rates increased this year. The increase in worker overtime was very high in the 4th quarter (187% of last year’s 4th quarter overtime).

My concern is that management’s story (inadvertently?) doesn’t tell the whole story. Instead, it omits any mention of increased overtime and utilization rates. The 20X3 decrease in returns could well be quite transitory. And, 20X4 returns could well increase to historic levels once enough time passes for the effects of worker fatigue (due to 4th Qtr. 20X3 overtime) to show up in reduced product quality (in the form of 20X4 returns). I am concerned that AUP’s real motivation for this entry may be to prevent their earnings from falling short of analysts’ forecasts and that changing the reserve is simply a way to “generate” earnings.
EXHIBIT 2

Skepticism Prompt Manipulation (Experiment 1 and 3)

Panel A: More Doubtful – Be Skeptical Even if No Signs Emerge that Management Lacks Integrity

Auditors should always be skeptical, and not just when they have had negative experiences with client management and not just when cues emerge that signal a lack of integrity.

The infamous case of FBI counterintelligence traitor Robert Hanssen provides a non-audit analogy that highlights how being insufficiently skeptical in the absence of cues that signal a lack of integrity increases one's exposure to devastating negative consequences. Hanssen's "home-for-dinner family-man" persona and other endearing qualities made him appear to be trustworthy; hence, he was a perfect spy. Over time, US intelligence officials became increasingly trusting of him and gave him access to secret and top-secret information. Hanssen exploited their misplaced trust by spying for the Russian government for 20 years. His espionage resulted in the transfer of a "store of national security secrets" and to the execution of three intelligence sources.

The Hanssen analogy clarifies why auditors always should be skeptical, even when no cues emerge that potentially signal that client-management lacks integrity.

Panel B: Less Doubtful (Neutral) – Be Skeptical if Signs Emerge that Management Lacks Integrity

Auditors should be skeptical when cues emerge that signal a lack of integrity, because it significantly constrains the effectiveness of an entity's controls.

The infamous case of FBI counterintelligence traitor Robert Hanssen provides a non-audit analogy that highlights how being insufficiently attentive to cues signalling a lack of integrity increases one's exposure to devastating negative consequences. Hanssen's 20 years of espionage resulted in the transfer of a "store of national security secrets" and to the execution of three intelligence sources. Yet, the FBI did not follow up on cues signaling that he was a Russian spy. As just one example, in 1990, eleven years before his arrest, Hanssen's family inexplicably found thousands of dollars in his home. Hanssen's brother-in-law, himself an FBI agent, became suspicious and told bureau members about the cash and that he suspected Hanssen was a Russian spy. Amazingly, the FBI failed to investigate.

The Hanssen analogy clarifies why auditors should be skeptical whenever cues emerge that potentially signal that client-management lacks integrity.
EXHIBIT 3
Wise-Thinking Disposition (WTD) Measurement Scale (Experiments 1, 2, and 3)

How strongly do you agree or disagree with the following statements?
1. My beliefs would not have been different if I had been raised by a different set of parents.
2. Changing your mind is a sign of weakness.
3. I always try to look at all sides of a problem.*
4. There is nothing wrong with being undecided about many issues.
5. Certain beliefs are just too important to abandon no matter how good a case can be made against them.
6. Intuition is the best guide in making decisions.
7. One should disregard evidence that conflicts with one’s established beliefs.
8. It is better simply to believe in something than to be confused by doubts about it.*
9. Abandoning a previous belief is a sign of strong character.
10. When I am upset at someone, I usually try to “put myself in his or her shoes” for a while.*
11. It is important to persevere in your beliefs even when evidence is brought to bear against them.
12. Good supervisors never leave their subordinates with doubts about the task at hand.*
13. A person should always consider new possibilities.
14. One usually can overcome difficulties by thinking about the problem, rather than through waiting for good fortune.
15. Before criticizing somebody, I try to imagine how I would feel if I were in her/his place.*
16. If I think longer about a problem, I will be more likely to solve it.
17. One should always revise beliefs in response to new information or evidence.
18. I sometimes find it difficult to see things from another person’s point of view.*
19. Coming to decisions quickly is a sign of wisdom.
20. People should always take into consideration evidence that goes against their beliefs.
21. Even if my environment (family, neighborhood, schools) had been different, I probably would have the same religious views.
22. I try to look at everybody’s side of a disagreement before I make a decision.*

Notes: To respond to the above statements, participants checked a box indicating that they strongly agree, agree, neutral, disagree, or strongly disagree. *Denotes statements that we adapt from Ardelt’s (2003) three-dimensional wisdom scale (3D-WS). Remaining statements are adapted from the Composite Active-Open-Minded-Thinking (AOMT) scale (Stanovich and West 2007).
EXHIBIT 4
Risk assessment dependent measures (Experiments 1 and 3)

Following are five categories that one could use to describe AUP’s preferred accounting treatment. Based on what you have learned from the case, interview, and audit team member’s rationale, please judge how likely it is that AUP’s preferred accounting treatment is a member of each category. (Place a vertical line at the appropriate points on each of the scales.)

| Category                                                                 | Extremely Unlikely | As Likely As Not | Extremely Likely |
|--------------------------------------------------------------------------|--------------------|------------------|------------------|
| Fraudulent Financial Reporting -- (intentional attempt to distort business reality and a clear violation GAAP) | 0  10  20  30  40  50  60  70  80  90  100 | | |
| “Near Fraud” -- (intentional attempt to distort business reality that really “pushes the envelope” of technical compliance with GAAP) | 0  10  20  30  40  50  60  70  80  90  100 | | |
| Earnings Management -- (intentional attempt to “window dress” business reality that is acceptable under GAAP) | 0  10  20  30  40  50  60  70  80  90  100 | | |
| High-Quality Accounting -- (good-faith attempt to fairly and neutrally reflect business reality that clearly is acceptable under GAAP) | 0  10  20  30  40  50  60  70  80  90  100 | | |
| Erroneous Accounting -- (good-faith attempt to fairly and neutrally reflect business reality but that unintentionally may violate GAAP) | 0  10  20  30  40  50  60  70  80  90  100 | | |

Notes: Excerpted in this exhibit are case materials participants used to record likelihood assessments regarding management’s intentions for the nonrecurring warranty-reducing journal entry. We combined “Fraudulent Financial Reporting” and “Near Fraud” into one overall measure. We further combined “High-Quality Accounting” and “Erroneous Accounting” into one measure. Combining these items simplifies without qualitatively affecting the reported results.
### EXHIBIT 5
Summary of *Evidence* manipulation in Experiment 2

| No. | Description |
|-----|-------------|
| 1   | **Benchmarking analysis**  
More indicative evidence condition: Projected revenue growth rate exceeds that of peer firms  
Less indicative evidence condition: Projected revenue growth rate in line with peer firms |
| 2   | **Sensitivity analysis**  
More indicative evidence condition: 1% decrease in revenue projections will result in fair value for the reporting unit falling below carrying value  
Less indicative evidence condition: 1% decrease in revenue projections will result in fair value that still exceeds carrying value |
| 3   | **Historical revenue projections**  
More indicative evidence condition: The company has consistently over-projected revenue in the past, which casts doubt on the current projections (more suggestive of bias)  
Less indicative evidence condition: The company has not consistently exceeded or fallen short in their projections (less suggestive of bias) |
| 4   | **Market factors**  
More indicative evidence condition: The projected revenue growth of Product C is suspect due to a new competing product and prior product delays  
Less indicative evidence condition: No competing product or prior product delays are present |
FIGURE 1
Experiment 1: Illustration of results relevant to our hypothesis

Panel A: The effect of Evidence and WTD strength on Attempted Fraud

Panel B: The effect of Evidence and WTD strength on Fraud vs. Not Misstated

Notes: This figure graphically depicts the observed mean values of Attempted Fraud (panel A) and Fraud vs. Not Misstated (panel B) from Experiment 1. In this experiment, we manipulate Evidence as either more or less indicative of fraud, and we use a median-split measure of the strength of participants’ WTDs. The mean values are reported and variables are defined in Table 1.
FIGURE 2
Experiment 2: Illustration of results relevant to our hypothesis

Based on your evaluation of the five-year revenue projections, what would you like to do at this point? Please select one of the options below.

- I conclude that the five-year revenue projections are reasonable. There is no need to initiate a discussion with my senior associate at this time.

- I am not ready to conclude at this point, but I will continue under the assumption that the five-year revenue projections are reasonable. There is no need to initiate a discussion with my senior associate at this time.

- I am not ready to conclude at this point, but I am not sure that the five-year revenue projections are reasonable. There are one or more issues I would like to discuss with my senior associate whenever they have some free time.

- I am not ready to conclude at this point, but I am not sure that the five-year revenue projections are reasonable. There are one or more issues I would like to discuss with my senior associate immediately.

- I conclude that the five-year revenue projections are unreasonable. There is no need to initiate a discussion with my senior associate at this time.

Notes: The figure graphically depicts participants’ choice of Next Action from Experiment 2. In this experiment, we manipulate Evidence as either more or less indicative of fraud, and we use a median-split measure of the strength of participants’ WTDs. Participants selected one response out of five possible responses as their Next Action. We code responses as one if
they selected an appropriate action, and zero otherwise. The proportion of appropriate follow-up Next Actions are reported in Table 5. The two appropriate responses when Evidence is more indicative of possible fraud are the third and fourth options. The boldfaced items shown above were also in the instrument.
TABLE 1
Experiment 1: Descriptive statistics

Panel A: Stronger WTD auditors, means (std. errors)

| Experimental condition(s) | (1) Attempted Fraud | (2) Within-GAAP Earnings Mgt. | (3) Not Misstated High-Quality Attempt | (4) Difference Fraud vs. Not Misstated |
|---------------------------|---------------------|--------------------------------|----------------------------------------|----------------------------------------|
| Doubtful Skepticism Prompt |                     |                                |                                        |                                        |
| More Indicative Evidence  | 61.22 (6.51)        | 67.67 (7.08)                  | 34.22 (6.13)                           | +10.28 (7.77)                          |
| n = 9                     |                     |                                |                                        |                                        |
| Less Indicative Evidence  | 39.71 (5.63)        | 72.92 (6.13)                  | 35.54 (5.31)                           | -14.52 (6.73)                          |
| n = 12                    |                     |                                |                                        |                                        |
| **Evidence Effect**       | +21.51**            | -5.25                         | -1.32                                  | +24.80**                               |

Neutral Skepticism Prompt

| More Indicative Evidence  | 58.78 (6.51)        | 68.00 (7.08)                  | 29.83 (6.13)                           | +9.86 (7.77)                           |
| n = 9                     |                     |                                |                                        |                                        |
| Less Indicative Evidence  | 35.86 (5.89)        | 60.00 (6.40)                  | 49.50 (5.54)                           | -18.89 (7.03)                          |
| n = 11                    |                     |                                |                                        |                                        |
| **Evidence Effect**       | +27.08**            | +8.00                         | -19.67                                 | +28.75**                               |

Collapsed across Prompt

| More Indicative Evidence  | 60.00 (4.60)        | 67.83 (5.00)                  | 32.03 (4.33)                           | +10.07 (5.49)                          |
| n = 18                    |                     |                                |                                        |                                        |
| Less Indicative Evidence  | 37.79 (4.07)        | 66.46 (4.43)                  | 42.52 (3.84)                           | -16.70 (4.87)                          |
| n = 23                    |                     |                                |                                        |                                        |
| **Evidence Effect**       | +22.21***           | +1.37                         | -10.49                                 | +26.77***                              |
### TABLE 1 (continued)

**Panel B: Weaker WTD auditors, means (std. errors)**

| Experimental condition(s) | (1) Attempted Fraud | (2) Not Misstated Within-GAAP Earnings Mgt | (3) High-Quality Attempt | (4) Difference Fraud vs. Not Misstated |
|---------------------------|---------------------|------------------------------------------|--------------------------|--------------------------------------|
| Doubtful Skepticism Prompt |                     |                                         |                          |                                      |
| More Indicative Evidence  | 57.35 (5.41)        | 71.54 (5.89)                            | 31.73 (5.10)             | +5.71 (6.47)                         |
| n = 13                    |                     |                                         |                          |                                      |
| Less Indicative Evidence  | 56.70 (6.17)        | 55.00 (6.71)                            | 40.10 (5.81)             | +9.15 (7.37)                         |
| n = 10                    |                     |                                         |                          |                                      |
| **Evidence Effect**       | **+0.65**           | **16.54**                               | **-8.37**                | **-3.44**                            |
| Neutral Skepticism Prompt |                     |                                         |                          |                                      |
| More Indicative Evidence  | 49.89 (5.41)        | 68.85 (5.89)                            | 43.42 (5.10)             | -6.25 (6.47)                         |
| n = 13                    |                     |                                         |                          |                                      |
| Less Indicative Evidence  | 35.65 (6.17)        | 73.80 (6.71)                            | 35.50 (5.81)             | -19.00 (7.37)                        |
| n = 10                    |                     |                                         |                          |                                      |
| **Evidence Effect**       | **+14.24***         | **-4.95**                               | **+7.92**                | **+20.67**                           |
| Collapsed across Prompt   |                     |                                         |                          |                                      |
| More Indicative Evidence  | 53.62 (3.83)        | 70.19 (4.16)                            | 37.58 (3.60)             | -0.27 (4.57)                         |
| n = 26                    |                     |                                         |                          |                                      |
| Less Indicative Evidence  | 46.18 (4.37)        | 64.40 (4.75)                            | 37.80 (4.11)             | -4.93 (5.21)                         |
| n = 20                    |                     |                                         |                          |                                      |
| **Evidence Effect**       | **+7.44**           | **+5.79**                               | **-0.22**                | **+4.66**                            |

**Notes:** Our hypothesis predicts higher assessed likelihoods of fraud given Evidence that is more indicative versus Evidence that is less indicative of fraud, but only for auditors with stronger WTDs. Exhibit 1 presents the more indicative and less indicative conditions of Evidence, and Exhibit 2 presents the doubtful versus neutral conditions of skepticism Prompt. WTD strength is measured as the median split of the sum of the squared deviations of each participant’s response as compared to the most-wise response for each of the 22 post-test questions in Exhibit 3. The scales in Exhibit 4 capture our dependent measures: participants assessed the likelihood of Attempted Fraud, Within-GAAP Earnings Management, or High-Quality Attempt, as shown above in columns (1), (2), and (3). Column (4) compares assessed likelihood of Attempted Fraud against the average of assessed likelihoods of Within-GAAP Earnings Management and High-Quality Attempt to determine Fraud vs. Not Misstated. Columns (1) and (4) are dependent variables for Hypothesis 1. Statistical significance of Evidence-effect differences greater than zero is indicated where *, **, and *** represent levels of 0.05, 0.01, and 0.001 respectively, using one-tailed tests.
### TABLE 2
Experiment 1: Hypothesis test

#### Panel A: 3-way ANOVA on Attempted Fraud (see column (1) in Table 1)

|                      | SS    | df | MS     | F-stat | p-value |
|----------------------|-------|----|--------|--------|---------|
| Evidence (E)         | 4688.10 | 1 | 4688.10 | 12.30  | 0.001   |
| Prompt (P)           | 1614.11 | 1 | 1614.11 | 4.24   | 0.043   |
| Wise-Thinking Disposition (WTD) | 21.42   | 1 | 21.42   | 0.06   | 0.813   |
| E×P                  | 299.43  | 1 | 299.43  | 0.79   | 0.378   |
| E×WTD                | 1163.57 | 1 | 1163.57 | 3.05   | 0.084   |
| P×WTD                | 658.17  | 1 | 658.17  | 1.73   | 0.193   |
| E×P×WTD              | 197.99  | 1 | 197.99  | 0.52   | 0.473   |
| **Total between-cells variance** | **8642.79** |   |         |        |         |
| **Error**            | 30107.78 | 79 | 381.11  |        |         |

Contrast ANOVA and Residual Between-Cell Variance Test

**Planned Contrast [0, +1, 0 -1]**

|                      | SS    | df | MS     | F-stat | p-value |
|----------------------|-------|----|--------|--------|---------|
| **4978.63**          | 1     | 1 | **4978.63** | **13.06** | **< 0.001** |
| Residual between-cells variance | 3664.16 | 6 | 610.69  | 1.60   | 0.158   |
| **Total between-cells variance** | **8642.79** |   |         |        |         |
| **Error**            | 30107.78 | 79 | 381.11  |        |         |

Contrast Variance Residual, $q^2$ 42.3%

#### Panel B: 3-way ANOVA on Fraud vs. Not Misstated difference (see column (4) in Table 1)

|                      | SS    | df | MS     | F-stat | p-value |
|----------------------|-------|----|--------|--------|---------|
| Evidence (E)         | 5265.93 | 1 | 5265.93 | 9.69   | 0.003   |
| Prompt (P)           | 2686.15 | 1 | 2686.15 | 4.94   | 0.029   |
| Wise-Thinking Disposition (WTD) | 11.05   | 1 | 11.05   | 0.20   | 0.887   |
| E×P                  | 540.46  | 1 | 540.46  | 1.00   | 0.322   |
| E×WTD                | 2607.85 | 1 | 2607.85 | 4.80   | 0.031   |
| P×WTD                | 1663.53 | 1 | 1663.53 | 3.06   | 0.084   |
| E×P×WTD              | 199.66  | 1 | 199.66  | 0.37   | 0.546   |
| **Total between-cells variance** | **12974.62** |   |         |        |         |
| **Error**            | 42930.55 | 79 | 543.43  |        |         |

Contrast ANOVA and Residual Between-Cell Variance Test

**Planned Contrast [0, +1, 0 -1]**

|                      | SS    | df | MS     | F-stat | p-value |
|----------------------|-------|----|--------|--------|---------|
| **7231.88**          | 1     | 1 | **7231.88** | **13.31** | **< 0.001** |
| Residual between-cells variance | 5742.75 | 6 | 957.13  | 1.76   | 0.118   |
| **Total between-cells variance** | **12974.62** |   |         |        |         |
| **Error**            | 42930.55 | 79 | 543.43  |        |         |

Contrast Variance Residual, $q^2$ 44.3%

**Notes:** *Planned contrast weights of [0, +1, 0 -1] correspond to [A, B, C, D] in Figure 1. p-values are two-tailed equivalent except for the **boldfaced** planned contrasts, which use one-tailed equivalents. For variable definitions, see Table 1.
**TABLE 3**  
Experiment 1: Investigation of RQ2.1: Stronger WTD participants

**Panel A: 2-way ANOVA on Attempted Fraud (see column (1) in Table 1)**

| Source of Variance     | SS     | df | MS      | F-stat | p-value |
|------------------------|--------|----|---------|--------|---------|
| Evidence (E)           | 4978.63| 1  | 4978.63 | 12.08  | 0.001   |
| Prompt (P)             | 99.77  | 1  | 99.77   | 0.24   | 0.627   |
| E×P                    | 4.95   | 1  | 4.95    | 0.01   | 0.914   |
| **Total between-cells variance** | **5083.35** |    |         |        |         |
| Error                  | 15328.89| 37 | 414.29  |        |         |

**Post Hoc Simple Main Effect Tests**

| Difference | Fisher’s LSD p-value<sup>a</sup> | Tukey HSD p-value<sup>b</sup> |
|------------|---------------------------------|------------------------------|
| More vs. Less Indicative Evidence (E) Given Doubtful Skepticism Prompt | -21.51 | 0.011<sup>*</sup> | 0.048<sup>*</sup> |
| More vs. Less Indicative Evidence (E) Given Neutral Skepticism Prompt | -22.91 | 0.009<sup>*</sup> | 0.038<sup>*</sup> |
| Doubtful vs. Neutral Skepticism Prompt (P) Given Less Indicative Evidence | 3.85 | 0.654 | 0.969 |
| Doubtful vs. Neutral Skepticism Prompt (P) Given More Indicative Evidence | 2.44 | 0.800 | 0.994 |

<sup>a</sup> The LSD test is used for post hoc comparisons of means when there are more than two groups.

<sup>b</sup> The Tukey HSD test is used for post hoc comparisons of means with equal variances.

---

This article is protected by copyright. All rights reserved.
### TABLE 3 (continued)

Experiment 1: Investigation of RQ2.1: Stronger WTD participants

**Panel B:** 2-way ANOVA on *Fraud vs. Not Misstated* difference (see column (4) in Table 1)

| Source                        | SS    | df | MS    | F-stat | p-value |
|-------------------------------|-------|----|-------|--------|---------|
| Evidence (E)                  | 7231.88 | 1  | 7231.88 | 12.23  | 0.001   |
| Prompt (P)                    | 57.68  | 1  | 57.68  | 0.10   | 0.757   |
| E × P                         | 39.33  | 1  | 39.33  | 0.07   | 0.798   |
| **Total between-cells variance** | 7328.89 |    |        |        |         |
| **Error**                     | 21877.92 | 37 | 591.29 |        |         |

**Post Hoc Simple Main Effect Tests**

| More vs. Less Indicative Evidence (E) Given Doubtful Skepticism Prompt | Difference | Fisher’s LSD p-value\(^a\) | Tukey HSD p-value\(^b\) |
|------------------------------------------------------------------------|------------|----------------------------|------------------------|
|                                                                        | -24.799    | 0.013*                     | 0.057*                 |

| More vs. Less Indicative Evidence (E) Given Neutral Skepticism Prompt | Difference | Fisher’s LSD p-value\(^a\) | Tukey HSD p-value\(^b\) |
|------------------------------------------------------------------------|------------|----------------------------|------------------------|
|                                                                        | -28.75     | 0.006*                     | 0.029*                 |

Doubtful vs. Neutral Skepticism Prompt (P) Given Less Indicative Evidence

| Doubtful vs. Neutral Skepticism Prompt (P) Given More Indicative Evidence | Difference | Fisher’s LSD p-value\(^a\) | Tukey HSD p-value\(^b\) |
|---------------------------------------------------------------------------|------------|----------------------------|------------------------|
|                                                                            | 0.48       | 0.971                      | 1.000                  |

**Notes:** *All p-values are two-tailed equivalents except for the **boldfaced** simple effects of Evidence given stronger WTD, which use a one-tailed equivalent as they pertain to our hypothesis. For variable definitions, see Table 1. *Fisher’s Least-Significant Difference reflects t-tests for all pairwise comparisons, with no adjustments to the observed significance levels for multiple comparisons. \(^b\)Tukey Honestly-Significant Difference test uses the largest value of the difference between all pairwise comparisons and the Studentized Range Distribution to control for experiment-wise alpha = 0.05. It is a conservative test of the significance of comparisons, which is useful for exploratory analyses.*
Experiment 1: Investigation of RQ2.2: Weaker WTD participants

**Panel A: 2-way ANOVA Attempted Fraud (see column (1) in Table 1)**

|                      | SS    | df  | MS    | F-stat | p-value |
|----------------------|-------|-----|-------|--------|---------|
| Evidence (E)         | 625.80| 1   | 625.80| 1.78   | 0.190   |
| Prompt (P)           | 2297.35| 1  | 2297.35| 6.53   | 0.014   |
| E×P                  | 521.83| 1   | 521.83| 1.48   | 0.230   |
| Total between-cells variance | 3444.98 |     |       |        |         |
| Error                | 14778.89| 42 | 351.88|        |         |

**Post Hoc Simple Main Effect Tests**

| Difference | Fisher’s LSD p-value<sup>a</sup> | Tukey HSD p-value<sup>b</sup> |
|------------|----------------------------------|-------------------------------|
| More vs. Less Indicative Evidence (E) Given Doubtful Skepticism Prompt | -0.646 | 0.935 | 1.000 |
| More vs. Less Indicative Evidence (E) Given Neutral Skepticism Prompt | -14.23 | 0.078 | 0.286 |
| Doubtful vs. Neutral Skepticism Prompt (P) Given Less Indicative Evidence | 21.05 | 0.016 | 0.073 |
| Doubtful vs. Neutral Skepticism Prompt (P) Given More Indicative Evidence | 7.46  | 0.316 | 0.742 |
TABLE 4 (continued)
Experiment 1: Investigation of RQ2.2: Weaker WTD participants

Panel B: 2-way ANOVA on *Fraud* vs. *Not Misstated* difference (see column (4) in Table 1)

|                  | SS     | df | MS     | F-stat | p-value |
|------------------|--------|----|--------|--------|---------|
| Evidence (E)     | 245.04 | 1  | 245.04 | 0.49   | 0.488   |
| Prompt (P)       | 4546.99| 1  | 4546.99| 9.07   | 0.004   |
| E×P              | 740.62 | 1  | 740.62 | 1.48   | 0.231   |
| **Total between-cells variance** | 5532.65 | 42 | 501.25 |        |         |

Post Hoc Simple Main Effect Tests

| More vs. Less Indicative Evidence (E) Given Doubtful Skepticism Prompt | Difference | Fisher’s LSD p-value<sup>a</sup> | Tukey HSD p-value<sup>b</sup> |
|------------------------------------------------------------------------|------------|---------------------------------|-------------------------------|
| More vs. Less Indicative Evidence (E) Given Neutral Skepticism Prompt | -12.75     | 0.183                           | 0.535                         |
| Doubtful vs. Neutral Skepticism Prompt (P) Given Less Indicative Evidence | 28.15      | 0.007                           | 0.036                         |
| Doubtful vs. Neutral Skepticism Prompt (P) Given More Indicative Evidence | 11.96      | 0.180                           | 0.530                         |

Notes: All p-values are two-tailed equivalents. For variable definitions, see Table 1. <sup>a</sup>Fisher’s Least-Significant Difference reflects t-tests for all pairwise comparisons, with no adjustments to the observed significance levels for multiple comparisons. <sup>b</sup>Tukey Honestly-Significant Difference test uses the largest value of the difference between all pairwise comparisons and the Studentized Range Distribution to control for experiment-wise alpha = 0.05. It is a conservative test of the significance of comparisons, which is useful for exploratory analyses.
| Experimental condition(s) | Revenue Projection Reasonableness Assessments mean (std. error) |
|---------------------------|-------------------------------------------------------------|
| Weaker WTD–More Indicative Evidence | 5.00 (0.45) |
| n = 19 | |
| Weaker WTD–Less Indicative Evidence | 5.89 (0.45) |
| n = 18 | |
| Stronger WTD–More Indicative Evidence | 5.27 (0.49) |
| n = 15 | |
| Stronger WTD–Less Indicative Evidence | 5.42 (0.49) |
| n = 19 | |

### 2-way ANOVA on Revenue Projection Reasonableness Assessments

| Source               | SS  | df | MS   | F-stat | p-value |
|----------------------|-----|----|------|--------|---------|
| Evidence (E)         | 4.78| 1  | 4.78 | 1.23   | 0.272   |
| Wise-Thinking Disposition (WTD) | 0.18| 1  | 0.18 | 0.05   | 0.832   |
| E x WTD              | 2.37| 1  | 2.37 | 0.61   | 0.438   |
| Total between-cells variance | 7.33 |     |      |        |         |
| Error                | 261.34| 67 | 3.90 |        |         |

### Contrast ANOVA and Residual Between-Cell Variance Test

| Planned Contrast [0, +1, 0, -1] | SS  | df | MS   | F-stat | p-value |
|---------------------------------|-----|----|------|--------|---------|
| Residual between-cells variance | 7.13| 2  | 3.71 | 0.95   | 0.392   |
| Total between-cells variance    | 7.33|    |      |        |         |
| Error                           | 261.34| 67 | 3.90 |        |         |

### Contrast Variance Residual, q²

| q² | 97.2% |

### Simple Effects for Revenue Projection Reasonableness Assessments

| Effect                                                        | t   | p-value |
|--------------------------------------------------------------|-----|---------|
| More vs. less indicative Evidence given auditors with stronger WTD | 0.23| 0.822   |
| More vs. less indicative Evidence given auditors with weaker WTD | 1.37| 0.176   |
| Stronger vs. weaker WTD given more indicative Evidence         | 0.39| 0.697   |
| Stronger vs. weaker WTD given less indicative Evidence          | 0.72| 0.474   |
TABLE 5 (continued)
Experiment 2: Inferential statistics

**Panel B: Next Action to address reasonableness of revenue projections**

| Experimental condition(s)                                      | Percentage of appropriate Next Action follow-up (std. error) |
|----------------------------------------------------------------|-------------------------------------------------------------|
| Weaker WTD–More Indicative Evidence                           | 63.16                                                       |
| n = 19                                                        | (11.37)                                                     |
| Weaker WTD–Less Indicative Evidence                           | 77.78                                                       |
| n = 18                                                        | (10.08)                                                     |
| Stronger WTD–More Indicative Evidence                         | 93.33                                                       |
| n = 15                                                        | (6.67)                                                      |
| Stronger WTD–Less Indicative Evidence                         | 63.16                                                       |
| n = 19                                                        | (11.37)                                                     |

**Logistic Regression for Next Action**

|                                | Z    | p-value |
|--------------------------------|------|---------|
| Intercept                      | 1.13 | 0.257   |
| Evidence (E)                   | 0.96 | 0.335   |
| Wise-thinking Disposition (WTD)| 1.84 | 0.065   |
| E × WTD                        | 2.07 | 0.038   |

**Contrast and Simple Effects for Next Action**

|                                | Z    | p-value |
|--------------------------------|------|---------|
| Planned Contrast: Auditors will follow up with their senior associate more often when Evidence is more vs. less indicative that revenue projections are overly aggressive, but only when they have stronger WTDs. (Contrast weights: 0 more indicative Evidence, weaker WTD; +1 more indicative Evidence, stronger WTD; 0 less indicative Evidence, weaker WTD; and -1 less indicative Evidence, stronger WTD) | 1.84 | 0.033*   |

**Simple Effects**

- More vs. less indicative Evidence given auditors with stronger WTD | 2.06 | 0.020*   |
- More vs. less indicative Evidence given auditors with weaker WTD  | -0.97| 0.331   |
- Stronger vs. weaker WTD given more indicative Evidence             | 2.06 | 0.020*   |
- Stronger vs. weaker WTD given less indicative Evidence             | -0.97| 0.331   |

**Notes:** All p-values are two-tailed except for Hypothesis 1 contrast and indicated (*) simple effects for Next Action, which are one-tailed. Revenue Projection Reasonableness Assessments: auditors’ assessment of the likelihood that the five-year revenue projections used in management’s goodwill impairment assessment analysis are reasonable. Specifically, participants were asked, “Based on your evaluation, how likely is it that the five-year revenue projections are reasonable?” on an 11-point scale. The scale had the following labels: 0 “Not at all Likely”, 5 “As Likely as not” and 10 “Extremely likely” with contrast weights of [0, +1, 0, -1]. Next Action: auditors’ choice of
whether or not to initiate discussion with their senior associate regarding any concerns about the reasonableness of the revenue projections prepared by management. See Figure 2 for the alternative responses. The contrast weights of $[0, +1, 0, -1]$ correspond to $[A, B, C, D]$ in Figure 2. See Table 1 for definitions of independent factors.