Are Rewards or Penalties More Effective for Increasing Skepticism?

Bright (Yue) Hong*
School of Accountancy & MIS
DePaul University

Timothy W. Shields
Argyros College of Business and Economics
Economic Science Institute
Chapman University

3 November 2025

*Corresponding Author (yhong20@depaul.edu)

We appreciate feedback from Lori Bhaskar, Allen Blay (discussant), Eric Chan, Mandy Cheng, Willie Choi, Jane (Kennedy) Jollineau, Xiaoxing Li (discussant), Yi Luo (discussant), Nathan Mecham (discussant), Mark Peecher, Hong Qu, Dan Rimkus, Karl Schuhmacher (discussant), Ira Solomon, Jack Stecher, Ken Trotman, Richard Tubbs, Laura Wang, Michael Williamson, Dan (Yuepin) Zhou (discussant) and participants at DePaul University, the University of Alberta, University of Illinois at Urbana-Champaign, University of Iowa, University of New South Wales Sydney, University of Texas Arlington, Lehigh University, the Accounting, Behavior, and Organizations Research Conference, the Auditing Section Mid-Year Meeting, the East Coast Behavioral Accounting Workshop, the Experimental Research in Management Accounting conference, the European Audit Research Network Symposium, the Hawaii Accounting Research Conference, the Managerial Accounting Mid-Year Meeting, and the New York University Accounting and Economics Society Conference. We thank Chapman University and DePaul University for their financial support.

Are Rewards or Penalties More Effective for Increasing Skepticism?

Abstract:

Auditor incentives primarily take the form of penalties. We conduct the first what-if analysis to examine whether penalties or *economically equivalent* rewards are more effective for increasing skepticism. Taking an experimental economics approach, we incentivize participants to avoid under-testing, over-testing, and biased testing. We find that rewards versus penalties increase skepticism in risk judgments and testing decisions, and that the increased skepticism reflects presumptive doubt rather than neutral skepticism. Although audit standards require presumptive doubt only for fraud detection, we suggest that presumptive doubt can also be beneficial when material misstatements are likely and when auditors exhibit insufficient skepticism. Thus, penalties are likely sub-optimal when skepticism is needed the most. Our study integrates judgment and decision-making with experimental economics, introduces theory-based criteria to incentivize effective, efficient, and objective audits, distinguishes the nature of skepticism in judgments and decisions, and highlights the need to examine the optimal design of auditors' incentives.

Keywords: experimental economics, incentive framing, rewards, penalties, accuracy, bias, judgment and decision-making, auditing, signal detection, regulatory focus, loss aversion
JEL Classifications: C92, D82, D81, M40
Data Availability: Data are available from authors upon request.

I. INTRODUCTION

Incentive design is critical to quality control (PCAOB 2024), yet little research addresses how to design auditors' incentives effectively. We conduct the first what-if analysis comparing whether rewards or *economically equivalent* penalties better enhance skepticism. Although rewards—such as praises, recognition, and bonuses—are widely used to motivate behavior (Luft 1994), the audit setting relies heavily on penalties, including public criticism, deficiency-focused feedback, litigation risk, and fines. While some may view rewarding auditors for doing their job as unnecessary, Peecher, Solomon, and Trotman (2013) question whether penalties are optimal. Penalties can enhance productivity (Church, Libby, and Zhang 2008; Imas, Sadoff, and Samek 2017) but also suppress creativity (Kang and Piercey 2024), impair decisions through anxiety (Bhaskar 2020), and create a less collaborative environment (Garza 2023). We find that rewards induce more skepticism in risk judgments and testing decisions than economically equivalent penalties, highlighting the need to reconsider the optimal design of auditors' incentives.

A critical step in incentive design is determining the desired behavior to be incentivized. The desired behavior in auditing is to be effective and efficient at the same time (Bowlin, Hobson, and Piercey 2015; Bhaskar, Majors, and Vitalis 2023; Bol, Grabner, Haesebrouck, and Peecher 2022). Under-auditing undermines effectiveness, whereas over-auditing undermines efficiency. Thus, more auditing is not always desired—a fundamental departure from typical managerial accounting settings where more effort is always desired (Bonner and Sprinkle 2002). To incentivize the desired behavior, one option is to *reward* auditors for making effective and efficient decisions. A decision is *effective* when a material misstatement is present and auditors decide to test, and it is *efficient* when a misstatement is absent and auditors decide not to test. Another option is to *penalize* auditors for making ineffective and inefficient decisions. A

decision is *ineffective* when a material misstatement is present, but auditors decide not to test, and it is *inefficient* when a misstatement is absent, but auditors decide to test.

Both rewards and penalties should also be designed to incentivize objective decision-making, which reflects neutral skepticism where auditors do not assume the presence or absence of a material misstatement (Nelson 2009). Neutral skepticism is desired because auditing is defined as "a systematic process of objectively obtaining and evaluating evidence…" (Auditing Concepts Committee 1972, 24). Consistent with this definition, audit standards (e.g., AS 1000.11.a), researchers (Nolder and Kadous 2018; Bol et al. 2022), and accounting firms (Cohen, Dalton, and Harp 2017) generally endorse neutral skepticism over presumptive doubt. Presumptive doubt is a bias where auditors assume that a material misstatement exists, unless evidence suggests otherwise, causing auditors to assess higher risks and request more evidence (Nelson 2009). Presumptive doubt can be viewed as deviating from risk-based auditing (Kachelmeier, Majors, and Williamson 2014), which relies on accurate risk judgments (Bonner, Majors, and Ritter 2018) rather than default assumptions of a misstatement.

Scholars have called for more research on skeptical decisions to better understand how skeptical judgments convert to skeptical decisions (Hurtt, Brown-Liburd, Earley, and Krishnamoorthy 2013). We predict that rewards will increase decisions to test relative to economically equivalent penalties. Drawing on regulatory focus theory (Higgins 1998; Crowe and Higgins 1997; Shah, Higgins, and Friedman 1998), we argue that rewards versus penalties cause auditors to automatically prioritize effectiveness over efficiency—even when the incentives emphasize effectiveness and efficiency equally. We further predict that the increased decisions to test reflect presumptive doubt rather than neutral skepticism. Advancing theory that skeptical judgments must reach a *threshold* to trigger skeptical decisions (Nelson 2009), we

propose that *where* auditors set this decision threshold determines the *nature* of skepticism (see Figure 1). We propose that an unbiased decision threshold suggests neutral skepticism, and that a biased threshold favoring testing suggests presumptive doubt. This bias towards testing should increase as auditors prioritize effectiveness, which is more likely under rewards than penalties.

In an experiment, we manipulate the reward versus penalty frame between participants. We use real monetary incentives to maintain strict economic equivalence between frames by varying the fixed pay. Participants inspected 100 bags for mislabeling. Under both frames, participants are incentivized to (1) test a bag if it came from the "mislabeled" distribution (i.e., a misstatement is present), (2) not test a bag if it came from the "correctly labeled" distribution (i.e., a misstatement is absent), and (3) be objective in deciding whether to test (i.e., choose an unbiased decision threshold between testing and not testing). That is, participants are incentivized to avoid under-testing, over-testing, and biased testing. Whether one bag is mislabeled is independent of another. Participants do not know for sure whether a bag is mislabeled, but recognizing the relationship between how a bag looks like and the two distributions increases the chance of making a correct inference. Participants assessed the likelihood of mislabeling and decided whether testing is warranted for each bag.

Our experiment plays a "bridge-building" role to integrate judgment and decision-making with experimental economics (Kachelmeier and King 2002, 228), offering critical insight otherwise difficult to obtain. While prior research typically treats neutral skepticism and presumptive doubt as dispositional traits (Hurtt 2010; Quadackers, Groot, and Wright 2014), we treat them as dynamic states that can be managed through incentive design. To this end, we introduce and implement criteria (Macmillan and Creelman 2004 Chapter 2) from signal detection theory (Green and Swets 1966) to incentivize an effective, efficient, and objective audit

(see Appendix 1)—an issue central to quality control. Importantly, by grounding *two* counterfactual states in *real* distributions (mislabeled and correctly labeled), we can identify the *direction* and *magnitude* of any bias in risk judgments *and* testing decisions. As a result, we can distinguish neutral skepticism from presumptive doubt across judgements and decisions, providing important policy implications.

As predicted, participants decide to test more bags under rewards than under penalties, exhibiting presumptive doubt because their decision threshold is more biased in favor of testing. Although unpredicted, participants also assess a marginally higher likelihood of mislabeling under rewards, again exhibiting presumptive doubt because their judgments are biased towards over-estimating misstatement risks. Controlling for risk judgments does not eliminate the effect of rewards versus penalties on testing decisions. This suggests that risk judgments alone are insufficient for predicting testing decisions—the decision threshold plays an additional role. Together, heightened risk judgments and a lowered testing threshold jointly contribute to increased decisions to test under rewards. Overall, we clarify the nature of skepticism in both judgments and decisions, examine the decision threshold as a critical link converting judgments to decisions (Nelson 2009), and extend regulatory focus theory to the audit setting (Hammersley, Leiby, and Nielson 2021; Peecher, Ricci, and Zhou 2024) and to probability judgments.

Although penalties enhance productivity (Church et al. 2008), we show that rewards enhance skepticism—specifically, presumptive doubt, which audit standards (e.g., AS 2110.52; PCAOB 2014, 2017) require for fraud detection (Carpenter 2007; Nelson 2009; McAllister, Blay, and Kadous 2021). We identify two additional scenarios when presumptive doubt is beneficial: (1) when material misstatements are *likely*, such as when clients have weak internal controls (Quadackers et al. 2014) and strong reporting bias (Majors 2016), and (2) when auditors

exhibit *insufficient* skepticism for reasons such as personality traits (Hurtt 2010; Bhaskar et al. 2023) or compromised independence. Auditors often adopt clients' preferences under high client pressure (Kadous, Kennedy, and Peecher 2003), weak audit committees (Bhaskar, Hopkins, and Schroeder 2019), consulting ties (Kowaleski, Mayhew, and Tegeler 2018), or social bonds with clients (Kachelmeier and Van Landuyt 2017). Restoring objectivity is important (Zhou 2020; Bauer 2015; Ricci 2022). We propose that inducing presumptive doubt may help.

We concur with Peecher et al.'s (2013) proposal to reward auditors for fraud detection through whistle-blower eligibility. Additionally, rewards need not be monetary—cultural elements such as performance evaluations (Brazel, Leiby, and Schaefer 2022) and the tone at the top (Bol et al. 2022) can serve as rewards when framed appropriately. In this spirit, audit partners and managers have suggested inspectors to "provide positive feedback on what was done well" and "provide recommendations and/or best practices" (Tegeler, Brown, and Downey 2024, 15). Similarly, allowing auditors to perform additional work after inspection without imposing penalty can help create a collaborative environment for improving audit quality (Garza 2023). Shifting auditors' incentive frame likely requires efforts across stakeholders (e.g., supervisors, audit firms, regulators, and the media) for auditors at different levels (e.g., individuals, teams, and firms). Taking "a step-by-step approach" (Kachelmeier and King 2002, 225) to isolate the effect of incentive framing, our study likely raises more questions than answers about the optimal design of auditors' incentives—a complex but important area for future research.

II. THEORY AND HYPOTHESES

Skepticism: Decisions Versus Actions

Skepticism is key to audit quality. In his seminal work, Nelson (2009) distinguishes between skeptical judgments and skeptical actions. Hurtt et al. (2013) note that audit research has

focused more on skeptical judgments than on skeptical actions, calling for more research on skeptical actions to better understand how skeptical judgments convert to skeptical actions. Nolder and Kadous (2018) identify "intentions" as an intermediate step between skeptical judgments and skeptical actions. While actions refer to *actual* behaviors, intentions refer to *intended* behaviors. Thus, actions reflect the execution of intentions. Although more research is starting to examine skeptical actions, for understandable reasons, the actions examined are often intended, not actual, behaviors (Griffith, Hammersley, Kadous, and Young 2015; e.g., Ricci and Rimkus 2025; Brazel et al. 2022; Peecher et al. 2024). Consistent with prior research, we examine intended skeptical behaviors—auditors' decisions to test or not to test. Either choice reflects a distinct course of action.[1]

Incentive Framing: Rewards Versus Penalties

By studying auditors' testing decisions, we make a first attempt to examine whether and how rewards versus *economically equivalent* penalties affect contracting behavior in the audit setting. Economically equivalent rewards and penalties purely differ in their "verbal description", reflecting "apparently superficial differences in language" (Luft 1994, 182). Luft's (1994) seminal work on incentive framing has inspired a stream of research in managerial accounting. The outcomes examined include employees' contract choices (Luft 1994; Brink and Rankin 2013), effort choices (Hannan, Hoffman, and Moser 2005; Brink 2011; Gonzalez, Hoffman, and Moser 2020; Burke, Towry, Young, and Zureich 2023), effort-based productivity (Church et al. 2008; Imas et al. 2017; Hossain and List 2012; Van der Stede, Wu, and Wu 2020),

[1] Decisions reflect intentions to pursue a *course* of action and often involve choosing among alternatives (Bonner 1999). The alternatives can include both active and passive options. For example, deciding between sleeping and exercising entails selecting distinct behavioral paths—even if sleeping involves minimal physical movement. Likewise, deciding not to engage in certain behaviors (e.g., not smoking, not drinking, etc.) represents an intentional course of action, even though the selected options reflect inaction. Decision-makers can also choose from more than two options, such as voting for, voting against, or abstaining from voting on a proposal.

information use (Frederickson and Waller 2005), misreporting (Nichol 2019), reciprocity (Christ, Sedatole, and Towry 2012), and whistleblowing (Chen, Nichol, and Zhou 2017). We expand this stream of research for the first time to the audit setting.

This expansion is needed to advance the debate on whether rewards or penalties best enhance audit quality (Peecher et al. 2013). Additionally, it can broaden our understanding of how incentive framing influences contracting behavior in accounting, because auditing differs fundamentally from typical managerial accounting settings. For example, we do not find loss aversion (Tversky 1981) useful for predicting whether and how incentive framing affects testing decisions. Loss aversion is a common explanation for why penalties increase effort and productivity more than equivalent rewards (e.g., Church et al. 2008; Imas et al. 2017; Brink and Rankin 2013; Hannan et al. 2005). More motivated to avoid losses than to seek gains, employees under penalties exert more effort, which in turn increases productivity. This logic rests on two premises: (1) more effort *helps* employees avoid losses, and (2) employees *know* ex ante that more effort helps them avoid losses such that they will act accordingly. Researchers ensure both premises hold when designing experiments (e.g., Hannan et al. 2005).

Auditing violates both premises. First, in productivity tasks, the desired behavior that helps employees avoid losses has only *one* direction—more effort. In contrast, the desired behavior that helps auditors avoid losses has *two* directions—effectiveness (from testing) and efficiency (from not testing). That is, more testing is not always desired. Second, employees in productivity tasks *know* ex ante the direction of behavior—more effort—that helps them avoid losses. In contrast, auditors do *not* know ex ante which direction—to test or not to test—will minimize losses. Depending on whether material misstatements exist, under the penalty frame, testing can help avoid losses (if effective) or incur losses (if inefficient). Testing is effective

when material misstatements exist but inefficient when they do not. The challenge of decision-making is that auditors do not know ex ante whether material misstatements exist. Thus, even if auditors are more motivated to avoid losses than to seek gains, it is unclear whether and how their decisions (to test or not to test) would differ between penalties and rewards.

Testing Decisions: Regulatory Focus Theory

Drawing on regulatory focus theory (Higgins 1998), we predict that rewards, compared to economically equivalent penalties, will increase decisions to test by activating a strategy to prioritize audit effectiveness over audit efficiency. We argue that this strategic shift in prioritization is automatic and therefore can occur even when both rewards and penalties are designed to incentivize effectiveness and efficiency *equally*.

Specifically, rewards can make individuals see their goals as hopes and aspirations, whereas penalties can make individuals see the same goals as duties and obligations. As a result, research shows that rewards induce a promotion focus in goal-pursuit, and that penalties induce a prevention focus in goal-pursuit (Shah et al. 1998). According to regulatory focus theory (Higgins 1998), individuals can pursue the same goal with a promotion or a prevention focus, and depending on the regulatory focus they take on, individuals would pursue the same goal in different ways. Specifically, promotion-focused individuals would adopt a strategy to ensure "hits" and avoid "misses", whereas prevention-focused individuals would adopt a strategy to ensure "correct rejections" and avoid "false alarms". This focus-strategy relationship has been supported in recognition memory tasks, where participants indicate whether they remember seeing information presented earlier (Crowe and Higgins 1997).

Hits, misses, correct rejections, and false alarms are terms from signal detection theory (Green and Swets 1966). Translating these terms to the audit setting (see Table 1), we define

"hits" as deciding to test when material misstatements are present, and "misses" as deciding not to test when material misstatements are present. We argue that ensuring hits and avoiding misses reflects a strategy aimed at audit effectiveness, which emphasizes the *presence* of material misstatements. A strategy aimed at effectiveness would call for more testing and is more likely under rewards, because rewards are shown to induce a promotion focus (Shah et al. 1998).

On the other hand, we define "correct rejections" as deciding not to test when material misstatements are absent, and "false alarms" as deciding to test when material misstatements are absent. We argue that ensuring correct rejections and avoiding false alarms reflects a strategy aimed at audit efficiency, which emphasizes the *absence* of material misstatements. A strategy aimed at efficiency would call for less testing and is more likely under penalties, because penalties are shown to induce a prevention focus (Shah et al. 1998). Comparing the two strategies activated by rewards and penalties, we predict that rewards, relative to economically equivalent penalties, will increase decisions to test.

[Insert Table 1 here]

H1: *A reward frame increases decisions to test relative to a penalty frame.*

Testing Decisions: Neutral Skepticism Versus Presumptive Doubt

Increased decisions to test can reflect neutral skepticism or presumptive doubt. Adding to research that distinguishes the nature of *trait* skepticism (Hurtt 2010; Quadackers et al. 2014; Cohen et al. 2017), we develop theory to distinguish the nature of *state* skepticism in testing decisions (see Figure 1). Prior research acknowledges that skeptical judgments must reach a *threshold* to result in skeptical decisions (Shaub and Lawrence 2002; Nelson 2009). Thus, assessing higher misstatement risks does not automatically result in decisions to test. Instead, auditors decide to test only if their risk judgments surpass a subjective threshold. We posit that

where auditors set their decision threshold for testing determines the *nature* of skepticism in their testing decisions. If auditors set an *unbiased* threshold between testing and not testing, increased decisions to test reflect *neutral* skepticism, which emphasizes objective decision-making. In contrast, if auditors *lower* their threshold for deciding to test, this leniency *bias* towards testing reflects *presumptive doubt*, which assumes material misstatements exist.

[Insert Figure 1 here]

Hurtt et al. (2013) call for deeper understanding of how skeptical judgments convert to skeptical decisions. Although a decision threshold is recognized to play a critical role in this conversion (Shaub and Lawrence 2002; Nelson 2009; Kadous et al. 2003, footnote 5), this threshold remains largely unexplored in auditing, with few exceptions (Ramsay and Tubbs 2005; Sprinkle and Tubbs 1998; Blocher, Moffie, and Zmud 1986). Addressing this gap, we propose that auditors' decision threshold for testing is shaped partly by how they balance audit effectiveness and efficiency. Specifically, the more auditors prioritize effectiveness over efficiency, the lower their threshold—or bar—for initiating tests, *ceteris paribus*. This shift in the decision threshold reflects a bias in favor of testing, consistent with presumptive doubt. Accordingly, we predict that the increased decisions to test under rewards versus penalties, as hypothesized in H1, will reflect presumptive doubt rather than neutral skepticism.

H2: *Increased decisions to test under rewards versus penalties reflect presumptive doubt.*

Risk Judgments

It is difficult to predict whether and how risk judgments may differ between rewards and economically equivalent penalties. On the one hand, risk judgments may not differ because as Bonner (1999, 385) notes, "judgments reflect one's beliefs, and decisions may reflect both beliefs and preferences". Therefore, preferences may explain inconsistences between judgments

and decisions. For example, recruiters may believe a job candidate is competent but still decide to reject the candidate because they dislike the candidate. Travelers may believe it will not rain but still choose to bring a raincoat because they prefer to stay dry. Similarly, auditors may believe material misstatements are unlikely but still choose to test because they want to prioritize effectiveness over efficiency. The inconsistencies between judgments and decisions are documented in audit research (Brazel et al. 2022; Hawkins 2017; Ricci and Rimkus 2025). In developing H1, we argue that rewards versus penalties induce a strategic *preference* to prioritize effectiveness over efficiency.[2] If this preference affects only decisions, then risk judgments should not differ by frames.

On the other hand, preferences, as reflected in decisions, may shape judgments through processes such as motivated reasoning (Kunda 1990). For example, recruiters who dislike and want to reject a candidate may believe the candidate is incompetent. Travelers who prefer to stay dry and bring a raincoat may believe it will rain. Similarly, auditors who want to prioritize effectiveness over efficiency and test more may believe material misstatements are likely. In these examples, judgments and decisions are consistent with each other. Individuals can engage in motivated reasoning to construct beliefs consistent with their preferences, when the beliefs can be justified within the bounds of plausibility. For example, it would be difficult to believe it will

[2] We note that this preference for effectiveness over efficiency is conceptually distinct from individuals' risk preference (e.g., Charness, Gneezy, and Imas 2013), which refers to whether individuals dislike uncertainty (risk-averse), like uncertainty (risk-seeking), or are indifferent to uncertainty (risk-neutral). For example, when choosing between Investment A (e.g., a guaranteed $2 return) and Investment B (e.g., a 50/50 chance of receiving $1 or $3, with the same expected return of $2), a risk-neutral investor would be indifferent, while a risk-averse investor would choose the guaranteed $2. In contrast, choosing between effectiveness and efficiency involves a different kind of trade-off: choosing effectiveness risks sacrificing efficiency, while choosing efficiency risks sacrificing effectiveness. Thus, either choice entails a distinct form of risk not captured by the notion of risk preference. Finally, although it is tempting to assume that a promotion focus predicts risk-seeking choices, and that a prevention focus predicts risk-averse choices, "there is no theoretical reason why this association must always be true" (Higgins and Cornwell 2016, 61). In fact, research finds that both foci can predict risk-seeking, risk-averse, and risk-neutral choices (Scholer, Zou, Fujita, Stroessner, and Higgins 2014; Zou, Scholer, and Higgins 2014).

rain when travelers are heading to a desert. Audit research has documented evidence consistent with motivated reasoning (e.g., Kadous et al. 2003; Bhaskar et al. 2019). Therefore, if the strategic preference to prioritize effectiveness over efficiency affects beliefs about the likelihood of material misstatements, then assessed risks may be higher under rewards than under penalties.

Thus, it is possible that rewards and penalties do not alter risk judgments, and that rewards may even lead to higher assessed risks than penalties. We are not aware of conclusive evidence favoring one possibility over the other, given that the research (Crowe and Higgins 1997) that we rely on to develop H1 examines individuals' responses in recognition memory tasks without measuring the underlying beliefs. While the participants' responses are consistent with regulatory focus theory (Higgins 1998), which predicts that a promotion focus activates strategies aimed at ensuring hits and avoiding misses, and that a prevention focus activates strategies aimed at ensuring correct rejections and avoiding false alarms, it remains unclear whether these preferences in goal-pursuit strategies influence beliefs about probability. By examining judgments about misstatement risks, we aim to clarify whether goal-pursuit strategies, as predicted by regulatory focus theory, change beliefs about the likelihood of material misstatements.

RQ: *Does the incentive frame affect risk judgments?*

III. METHOD

To test our predictions, we take advantage of an experimental economics approach. First, although incentives can take many forms, we use real monetary incentives to achieve strict equivalence between rewards and penalties, consistent with prior research (Luft 1994; Church et al. 2008). Second, auditors may perceive rewards as less credible (Brazel et al. 2022) if penalties are the norm. To compare rewards and penalties fairly and cleanly, our design abstracts away

from current norms and idiosyncratic prior preferences between effectiveness and efficiency. Third, to incentivize an effective, efficient, and objective audit, we introduce and implement quantitative criteria (Macmillan and Creelman 2004 Chapter 2) from signal detection theory (Green and Swets 1966). Fourth, distinguishing neutral skepticism from presumptive doubt requires (1) two states (misstatement: present and absent) to evaluate testing decisions and (2) objectively correct answers to evaluate risk judgments, made possible by grounding *both* states in *real distributions*—an essential feature often missing in prior research. See experimental procedures in Figure 2.

[Figure 2]

Bag Inspection Task

We design a novel task that captures the essence of making risk judgments and testing decisions but does not require audit knowledge. 196 participants from Amazon Mechanical Turk assumed an inspector's role for a hypothetical company that makes heating and cooling bags for physical therapy.[3,4] We drew 100 bags from two equally likely distributions (heating and cooling: $P_H = 0.5$). Due to a mistake in production, all bags were labeled as cooling bags. All participants viewed the same 100 bags presented in a randomized order with one bag per screen. For each bag, participants assessed the likelihood of mislabeling and decided whether they would test it. Whether one bag is mislabeled is independent from another because we drew each bag independently from the distributions. As described later, participants were incentivized to (1) test

[3] To reduce the risk of collecting low quality data, we screened participants for bots and restricted participants to those, as per MTurk, who had high accuracy (i.e., greater than 95 percent approval rate), high productivity (i.e., had completed more than 1,000 other tasks), and were in the U.S. or Canada, heeding advice from others using worker platforms (e.g., Peer, Vosgerau, and Acquisti 2013; Mturk Data Blog 2018). Additionally, we restricted participants to high school graduates and above. Data collection occurred before the creation of ChatGPT.

[4] We obtained approval from a USA west coast university's Institutional Review Board (IRB). The IRB forbids researchers from deceiving participants.

a bag if it is mislabeled (state = Yes, i.e., the bag came from the heating distribution), (2) not test a bag if it is correctly labeled (state = No, i.e., the bag came from the cooling distribution), and (3) choose a neutral, unbiased testing threshold.

To facilitate participants' judgments and decisions, we design bags such that their appearance provides *imperfect* information about whether a bag is mislabeled. See Table 2 for examples of bags. Participants were told that (1) each exothermic ball in a heating bag has a 60% chance of being red and a 40% chance of being white, (2) each endothermic ball in a cooling bag has a 60% chance of being white and a 40% chance of being red, and (3) the company produced thousands of bags with the same number of heating and cooling bags. Attending to and thinking carefully about this information should help participants infer that a bag is *more likely than not* mislabeled if the number of red balls in a bag (#Red) exceeds the number of white balls in that bag (#White). Of course, if #Red > #White, there is still a chance that the bag is correctly labeled. Likewise, if #Red < #White, there is still a chance that the bag is mislabeled. However, the likelihood of mislabeling increases as #Red - #White increases. Thus, comparing #Red with #White helps but does not guarantee the inference will be correct.

[Table 2]

Participants were told that their decision to test a bag would not affect whether the next bag presented was mislabeled. Thus, participants should base their decisions solely on how each bag looks like rather than the number of bags already tested. We use seven questions to check participants' comprehension of key aspects of the task, including the features of the heating and cooling distributions, the relationship between the distributions, the independence of each bag, the concept of probability, and the relationship between a decision and payoffs. Participants could not view bags until they passed all seven comprehension checks. After inspecting all bags,

participants received feedback about their risk judgments, testing decisions, and the associated pay. Consistent with prior research (e.g., Luft 1994; Hannan, Hoffman, and Moser 2005; Christ, Sedatole, and Towry 2012), we withheld feedback until the end such that participants' judgments and decisions would only be affected by incentive framing and not by feedback on prior decisions (King and Schwartz 1999).[5]

Independent Variable

We manipulate the *incentive frame* (reward versus penalty) between participants after they start the task. In the reward frame condition, participants were informed that they would make $2 plus a $2 bonus if 2/3 or more of their testing decisions were correct (i.e., the payoff for correct rejections equals the payoff for hits). In the penalty frame condition, participants were informed that they would make $4 minus a $2 penalty if more than 1/3 of their testing decisions were incorrect (i.e., the payoff for false alarms equals the payoff for misses).[6] By varying the fixed pay between frames, we keep the incentives for the same level of performance constant between frames. Consistent with most incentive framing research (e.g., Christ et al. 2012; Hannan et al. 2005; Hossain and List 2012), we use a target-based incentive scheme, which helps keep the explicit performance expectation (target) constant between frames. Participants' expected performance and payoffs earned (mean = 3.22 US dollars) do not differ between frames (both p-values > 0.162, untabulated).[7]

[5] In the real world, the timing of feedback varies across scenarios. For example, auditors may not learn that they missed detecting a material misstatement until a whistle-blower reports it years later. Examining how feedback affects judgments and decisions is beyond the scope of our study. Readers interested in feedback can consult Kluger and DeNisi (1996) for a systematic review, which documents the mixed effects of feedback on performance and the complex mechanisms (i.e., task-motivation, task-learning, and meta-task processes) through which feedback affects performance.

[6] When recruiting participants, we describe the incentives neutrally as "you will be paid $2 or $4 depending on your performance" in the advertisement.

[7] All p-values are two sided except when otherwise indicated for directional predictions.

Under both frames, our incentives are designed to encourage an effective, efficient, and objective audit. An effective and efficient audit means that auditors should avoid under-testing and over-testing (i.e., more testing is not always better). In other words, we want to incentivize auditors to make correct decisions—test only when material misstatements are present and not test only when material misstatements are absent. An objective audit reflects neutral skepticism in deciding whether to test. That is, when auditors set their decision threshold for testing, this threshold should be unbiased between testing and not testing (see Figure 1). By incentivizing an unbiased decision threshold, any bias we observe in the threshold should be due to the framing of the incentives—our construct of interest—rather than the underlying economics of the incentives.

Next, we introduce to accounting research two criteria that incentive designs must satisfy to promote an effective, efficient, and objective audit. These criteria (Macmillan and Creelman 2004 Chapter 2) are from signal detection theory (Green and Swets 1966), an influential framework for examining decision-making under uncertainty in fields including psychology, engineering, medicine, and statistics. By applying these criteria to auditor decision-making, we extend the goal of incentive designs in accounting beyond increasing effort-based productivity, (e.g., Hannan et al. 2005; Church et al. 2008), where the desired behavior has only *one* direction (i.e., more effort is always better). We explain the two criteria provided below in more detail in Appendix 1.

Criterion 1: to encourage an effective and efficient audit, we should set $(U_C - U_F)/(U_H - U_M) = 1$. This criterion incentivizes decision-makers to choose a decision threshold that maximizes "proportion correct" (Macmillan and Creelman 2004, 37 Chapter 2),

calculated as the number of correct decisions made (i.e., hits and correct rejections) as a percentage of the number of total decisions made.

Criterion 2: to encourage an objective audit, we should set $(U_C - U_F)/(U_H - U_M) = P_H/(1 - P_H)$. With this criterion, choosing an unbiased decision threshold would maximize decision-makers' expected utility (Macmillan and Creelman 2004, 28, 38 Chapter 2).

Notation: U_i refers to auditors' utility U_i associated with the four outcomes, where i denotes hit (H), miss (M), correct rejection (C), or false alarm (F). P_H refers to the base rate of a material misstatement. Recall that a "hit" refers to deciding to test when a material misstatement is present. A "miss" refers to deciding not to test when a material misstatement is present. A "correct rejection" refers to deciding not to test when a material misstatement is absent. A "false alarm" refers to deciding to test when a material misstatement is absent.

Our design simultaneously satisfies the two criteria. Recall that we drew bags from two equally likely distributions, which makes the base rate of mislabeling $P_H = 0.5$. Additionally, our payoffs should elicit $U_C = U_H > U_F = U_M$, which then makes $(U_C - U_F)/(U_H - U_M) = 1 = P_H/(1 - P_H)$ since $P_H = 0.5$. The symmetric payoffs that we provide are critical to satisfying the two criteria without knowing each participant's distinct utility function. Specifically, we set the payoffs π_i equal between correct rejections and hits ($\pi_C = \pi_H$) and equal between false alarms and misses ($\pi_F = \pi_M$). Assume that if $\pi_C = \pi_H$, then $U(\pi_C) = U(\pi_H)$, and that if $\pi_F = \pi_M$, then $U(\pi_F) = U(\pi_M)$. With the symmetric payoffs, $\frac{U(\pi_C) - U(\pi_F)}{U(\pi_H) - U(\pi_M)} = 1$ will always hold *regardless* of the form of participants' utility function (including risk attitudes). Using

asymmetric payoffs to satisfy these criteria would require knowing each participant's specific utility function ex-ante and providing the appropriate payoffs specific to each participant.[8]

We use payoffs to elicit $U_C,\ U_H,\ U_F\ and\ U_M$, without operationalizing the *specific* and *complete* set of determinants of U_i. This design is practical and consistent with prior research. For example, Bowlin et al. (2015) use payoffs to discourage over-auditing and under-auditing without operationalizing the cost of testing, deadline pressure, reputational gains, or client satisfaction. However, for many reasons (costly testing, time pressure, client satisfaction, etc.), auditors should avoid over-testing when material misstatements are absent. This is why efficiency is desired and why in our design deciding to test results in lower payoffs than deciding not to test when bags are correctly labeled (payoffs: false alarms < correct rejections). Similarly, for many reasons (e.g., reputation, lawsuits, etc.), auditors should also avoid under-testing when material misstatements are present. This is why effectiveness is desired and why in our design deciding not to test results in lower payoffs than deciding to test when bags are mislabeled (payoffs: misses < hits).

Dependent Variables

Risk judgments

Participants *assessed the likelihood of mislabeling* for each bag. We evaluate their probabilistic judgment for each bag against a normative performance benchmark for that bag.

[8] To illustrate, assume a participant's utility function for monetary payoff π is $U(\pi) = \frac{\pi^{1-\theta}}{1-\theta}$, which exhibits constant relative risk aversion. The parameter θ captures risk attitudes, where $\theta < 0$ is risk seeking, $\theta = 0$ is risk neutral, and $\theta > 0$ is risk averse. Apart from the payoffs we provide, we assume that participants do not have prior preferences for one correct (or incorrect) decision over the other. With symmetric payoffs such as when $\pi_C = 4 = \pi_H$ and $\pi_F = 2 = \pi_M$, regardless of risk attitudes, $(U(4) - U(2))/(U(4) - U(2))$ will always be one. If payoffs were asymmetric such as when $\pi_C = 6, \pi_H = 7, \pi_F = 4$, and $\pi_M = 5$, then the ratio $(U_C - U_F)/(U_H - U_M)$ would be less than one when $\theta < 0$, one when $\theta = 0$, and greater than one when $\theta > 0$. That is, whether $(U_C - U_F)/(U_H - U_M)$ is one will depend on the specifics of the participant's utility function. Thus, to use asymmetric payoffs, the experimenter must elicit the participant's utility function first and then set payoffs *for that participant* accordingly so that $(U_C - U_F)/(U_H - U_M)$ is one for that participant. Furthermore, we believe symmetric payoffs are the simplest, most fault-tolerant, and most natural for participants to compute.

This benchmark is the Bayesian probability estimate (i.e., posterior likelihood of mislabeling), which is a function of the number of red balls minus the number of white balls in a bag (#Red – #White), as shown in Table 2. This type of benchmark is often missing in recent research on audit judgment. Without the benchmark, it is a challenge to quantify and differentiate neutral skepticism and presumptive doubt in risk judgments. We overcome this challenge with a highly precise measure, *judgment bias*.

We calculate *judgment bias* as the assessed likelihood of mislabeling for a bag minus the Bayesian estimate for that bag, consistent with measures used in prior research (Walther and Willis 2012; Duru and Reeb 2002). A bias towards over-estimating the likelihood of mislabeling suggests presumptive doubt in risk judgments. Curious readers may also wonder about judgment accuracy. As a supplemental measure, we also calculate *judgment error*, which is the absolute deviation of the assessed likelihood from the Bayesian estimate for each bag.[9] A smaller error indicates higher judgment accuracy. To compare risk judgment, bias, and error between the frames, we average each measure across 100 bags for each participant.

A strength of our design is that we use real distributions to generate the bags so we can establish a normative performance benchmark to evaluate judgments using Bayes' rule. "The technique is particularly useful in assessing…the deviations of responses from optimality" (Libby and Lewis 1977, page 254). As shown in Table 2, the more #Red relative to #White, the higher the Bayesian probability estimate. When #Red = #White, there is a 50 percent chance that the bag is mislabeled. Regulatory focus research in psychology does not examine probabilistic judgment nor use real distributions to generate the stimuli (Crowe and Higgins 1997), and thus

[9] Our measures of bias and error are consistent with those used in earnings forecasts research (e.g., Walther and Willis 2012; Duru and Reeb 2002), where forecast bias = forecasted earnings – actual earnings, and forecast error = |forecasted earnings – actual earnings|. In our task, the "actual" risk of mislabeling is the posterior likelihood that a bag is mislabeled (i.e., the Bayesian probability estimate).

cannot speak to any incentive framing effects on probabilistic judgment nor evaluate the judgment bias and accuracy against a normative benchmark.

Another benefit of establishing a normative performance benchmark is that "the technique is particularly useful in assessing the impact of information set variables on cue usage" (Libby and Lewis 1977, page 254). Note that the Bayesian estimate does not depend upon bag size (six versus twelve balls per bag). Thus, bag size is an irrelevant cue to risk judgments. By varying the bag size, we can assess whether participants systematically ignore the irrelevant cue of bag size and strictly use #Red-#White in assessing risks. As shown in the supplemental analysis, the design of two bag sizes helps us infer participants' judgment process.

Testing Decisions

We calculate *the percentage of bags tested* for each participant. To estimate where participants set their decision threshold for testing, we use the *threshold bias* measure well established in signal detection theory (Green and Swets 1966). To construct the measure, we first calculate each participant's hit rate (H) and false alarm rate (F). The hit rate is the percentage of mislabeled bags tested. The false alarm rate is the percentage of correctly labeled bags tested. Second, we estimate the *threshold bias*, calculated as $-0.5[Z(H) + Z(F)]$ (Macmillan and Creelman 2004, chapter 2).[10] A zero value suggests a neutral threshold that is unbiased between testing and not testing, consistent with neutral skepticism. A negative value suggests a lenient threshold that is biased toward testing, consistent with presumptive doubt. A positive value suggests a strict threshold that is biased towards not testing. Thus, the *threshold bias* measure

[10] Z(rate) is the inverse of the standard normal cumulative distribution, where rates are greater than zero but less than one. Z(rate) values are negative for rates less than ½, zero for a rate of ½, and positive for rates greater than ½. The specific forms denoted assume that the perceived distributions of states are normal with equal variance, and that there are sufficient observations of each state of nature so that the rates are meaningful (Macmillan and Creelman 2004; Ramsay and Tubbs 2005).

allows us to determine the *direction* and *magnitude* of any bias in testing decisions, which is key to clarifying the nature of skepticism (H2).

Although our focus is to differentiate neutral skepticism and presumptive doubt in testing decisions, curious readers may again wonder about decision accuracy. As a supplemental measure, we calculate *the percentage of correct decisions made.* Additionally, we estimate *decision accuracy* using $Z(H) - Z(F)$, another well-established measure from signal detection theory (Macmillan and Creelman 2004, chapter 1). A higher value indicates a higher sensitivity to whether a bag is mislabeled. A major contribution of signal detection theory is that its measures can separate bias from accuracy in performance evaluation (Ramsay and Tubbs 2005). For example, a decision to test when a misstatement exists could reflect the ability to recognize that a misstatement exists (i.e., accuracy) or a preference to test (i.e., bias). To separate bias from accuracy, it is critical that we establish counterfactuals (i.e., mislabeled bags *and* correctly labeled bags) in designing the task.

These measures have been used in accounting research (Blocher et al. 1986; Sprinkle and Tubbs 1998; Ramsay and Tubbs 2005). To provide some intuition about how the measures can differentiate bias from accuracy when evaluating testing decisions, suppose a participant is extremely biased in favor of testing (i.e., extreme presumptive doubt). That is, the participant sets an extremely low bar (i.e., threshold) for testing and as a result tests all bags (hit rate = false alarm rate = 100%). In this case, the decision accuracy measure will be zero, which indicates zero sensitivity to whether a bag is mislabeled. Meanwhile, the threshold bias measure will reach its minimum (negative infinity), consistent with an extremely low (lenient) bar for initiating tests.

If instead no bags are tested (hit rate = false alarm rate = 0%), the decision accuracy measure will remain at zero, again indicating zero sensitivity, but the threshold bias measure will reach its maximum (positive infinity), suggesting an extremely high (strict) bar for testing. Now, if only the mislabeled bags are tested (hit rate = 100%, false alarm rate = 0%), the decision accuracy measure will reach its maximum (positive infinity), and the threshold bias measure will become zero (unbiased). Alternatively, if only the correctly labeled bags are tested (hit rate = 0%, false alarm rate = 100%), the decision accuracy measure will drop to its minimum (negative infinity) with the threshold bias unchanged at zero (unbiased).

IV. RESULTS

Participants' Understanding of the Task

Risk Judgments

To assess whether participants understood their task, we compare participants' assessed likelihood of mislabeling for low-risk bags (#Red - #White $\leqslant$ -4, posterior likelihood of mislabeling $\leqslant$ 16.49%) versus high-risk bags (#Red - #White $\geqslant$ 4, posterior likelihood of mislabeling $\geqslant$ 83.50%). Out of the 196 participants, only five assessed a higher likelihood of mislabeling for low- versus high-risk bags. Overall, participants appeared to understand the relationship between bag compositions (#Red - #White) and the likelihood of mislabeling.

Testing Decisions

In our task, the perfect testing strategy that maximizes the expected payoffs is to test a bag when #Red > #White, not test a bag when #Red < #White, and be indifferent between testing and not testing when #Red = #White. Using perfect strategies, however, does not imply an unbiased decision threshold or maximum decision accuracy. We illustrate this point in Table 3 with three versions of hypothetical perfect strategies, which differ in decisions made only when

#Red = #White (25 bags), since any combinations of decisions made when #Red = #White (indifferent between testing and not testing) are considered as perfect. As shown in Table 3, the associated testing threshold ranges from negative (biased towards testing) to positive (biased towards not testing), even when all three strategies are perfect. Also shown in Table 3, using perfect strategies does not guarantee maximum decision accuracy (positive infinity) or correctness (100 percent correct), which are attained only when participants *know* exactly whether each bag is mislabeled ex ante. This is obviously not the case in our experiment.

We do not expect participants to test 100 percent of bags when #Red > #White and 0 percent of bags when #Red < #White. Yet, we still find that, of the 196 participants, 45 adopted the perfect testing strategy for all bags (rewards: n = 23; penalties: n = 22). Relaxing the benchmark of a perfect testing strategy, we find that among the 196 participants, 165 consistently tested an equal or higher percentage of bags as the risk of mislabeling increased from one category to the next (rewards: n = 82; penalties: n = 83). Panel A of Table 2 summarizes the eight risk categories per bag composition (#Red - #White). For each participant, we consider testing to be weakly increasing if the percentage of bags tested within a higher-risk category (e.g., #Red - #White = 4) is equal to or greater than that within the adjacent lower-risk category (e.g., #Red - #White = 2). These results suggest that most participants did not test bags randomly: their testing decisions correctly correspond to the underlying risk of mislabeling.

Unqualified Participants

Although most participants understood their task, we exclude observations from 20 participants whose risk judgments and testing decisions are internally inconsistent. Specifically, these participants tested a bag when their assessed likelihood of mislabeling for that bag was 48

percent or lower, *and* they did not test a bag when their assessed likelihood of mislabeling for that bag was 52 percent or higher.[11] Thus, we retain 176 observations in the analysis.[12]

Main Results

Test of H1

H1 predicts that a reward versus penalty frame increases decisions to test. As predicted, the percentage of bags tested is significantly higher in the reward than penalty frame condition (one-tailed p = 0.019, Table 3).[13] Untabulated panel regression analysis reveals that a reward frame significantly increased test decisions in two low-risk categories (#Red - #White = -6 and -2, posterior likelihood of mislabeling = 8.07 and 30.77 percent) and one high-risk category (#Red - #White = 4, posterior likelihood of mislabeling = 83.50 percent, all p-values < 0.05, see Figure 3).[14] As a result, the false alarm rate in the reward frame condition is significantly higher than that in the penalty frame condition (means = 39.4% versus 33.8%, p = 0.032, Table 3), with the hit rate being no different between conditions (p = 0.162, Table 3).[15]

[11] We label these participants as "flippers" (ten in each condition). To be classified as a flipper, a participant must test when the assessed likelihood of mislabeling is low *and* not test when the assessed likelihood of mislabeling is high. So, a participant who always (or never) tested would *not* be categorized as a flipper. Flippers provided internally inconsistent judgments and decisions more than 50 percent of the time (on average, for 58 bags out of 100 bags). For comparison, out of the 196 participants, 54 participants never provided internally inconsistent judgments and decisions, and 122 participants did so for no more than five out of the 100 bags. Compared to non-flippers, flippers made significantly fewer correct testing decisions, failed more task comprehension checks on their first attempt, and scored lower on the expanded cognitive reflection test (Toplak, West, and Stanovich 2014), which is indicative of analytical ability (Toplak, West, and Stanovich 2011; Welsh, Burns, and Delfabbro 2013; all p-values ≤ 0.002). These results suggest that flippers are potentially confused about and incapable of performing our task.

[12] Including all observations does not change any inferences except that it would slightly weaken the results reported for testing decisions. Specifically, the p-values reported in Table 3 would be 0.051 for the percentage of bags tested, 0.054 for the threshold bias, and 0.102 for the false alarm rate.

[13] Incentive frame does not interact with bag size in predicting any of the dependent variables (all p-values > 0.160, untabulated). Therefore, we pool data across bag sizes in tests of hypotheses.

[14] Participants identify the panels and bags identify the trials for the logit regression, which accounts for within-participants variance given multiple observations per participant. The dependent variable is Decision (1 = test, 0 = not test) at the bag level. The independent variables are Frame (rewards = 1; penalties = 0), Composition (#Red - #White), and their interaction. The regression yields a χ^2 (15) statistic of 4,827.71 and p < 0.001.

[15] Inferences do not change if we use the number of hits and the number of false alarms each participant incurred as an alternative measure, instead of the hit rate and false alarm rate per participant, as each participant saw the same 100 bags.

[Table 3 and Figure 3]

One might suspect that for participants who understood the task perfectly, the difference in testing might be greatest when it is most uncertain whether a misstatement exists. Restricting the analysis to the 45 participants who adopted the perfect testing strategy, we find that when the posterior likelihood of mislabeling is 50 percent (#Red = #White), participants in the reward frame condition decided to test 73 percent of the time, whereas participants in the penalty frame condition decided to test 60 percent of the time, directionally consistent with H1.[16] Thus, even "rational" participants decide to test more under the reward versus penalty frame when evidence suggests a 50% chance of mislabeling.

Test of H2

H2 predicts that the increased decisions to test under a reward versus penalty frame reflect presumptive doubt rather than neutral skepticism. Consistent with H2, we find that the threshold bias measure is more negative under the reward than under the penalty frame (one-tailed p = 0.024, Table 3).[17] Thus, participants in the reward frame are more biased towards deciding to test than those in the penalty frame, even though the incentives are designed to encourage objective decision-making under both frames. This result, which reflects presumptive doubt, can be explained by the increased decisions to test even when material misstatements are absent (i.e., false alarm rate, see Table 3) under rewards than under penalties.

More specifically, the threshold bias measure under the reward frame (mean = -0.16) is significantly less than zero ($t_{85} = 2.64$, p = 0.010, untabulated), suggesting that the decision threshold is biased towards testing—consistent with presumptive doubt as predicted in H2. In

[16] This 13 percent difference is not significant at conventional levels potentially due to the smaller sample size. Similarly, we observe directional support for H2 among the 45 perfect testers.

[17] Inferences do not change if we use the untransformed hit rates plus the untransformed false alarm rates ($H + F$) as an alternative measure for the testing threshold bias ($t_{174} = 2.09$, one-tailed p = 0.019, untabulated).

contrast, the threshold bias measure under the penalty frame (mean = -0.012) is not significantly different from zero ($t_{89} = 0.47$, p = 0.766, untabulated), suggesting that the threshold is unbiased (i.e., neutral skepticism). An unbiased threshold, as explained earlier, does not imply the use of perfect strategies.

Test of RQ

Our RQ asks whether the incentive frame affects risk judgments. The average assessed likelihood of mislabeling is marginally higher under a reward versus penalty frame (means = 53.4% versus 51.7%, p = 0.064, Table 3). This result suggests that preferences for goal-pursuit strategies (effectiveness or efficiency), as activated by rewards versus penalties, can change beliefs about the likelihood of material misstatements, even when both rewards and penalties are designed to incentivize *no preference* between effectiveness and efficiency. That is, the payoffs provided are symmetrical between effectiveness and efficiency.

More specifically, the average assessed risk under the reward frame significantly exceeds the Bayesian estimate of risk (i.e., the posterior likelihood of mislabeling; p = 0.012, untabulated), suggesting a judgment bias towards over-estimating the likelihood of mislabeling. Under the penalty frame, the average assessed risk does not significantly differ from the Bayesian estimate (i.e., judgment bias = 0, p = 0.607, untabulated). Thus, the reward frame induces more presumptive doubt in risk judgments than the penalty frame. This result is consistent with motivated reasoning, suggesting that a strategic preference for effectiveness over efficiency can create biased beliefs to over-estimate misstatement risks. In Figure 4, we use an asterisk to denote two risk categories (#Red - #White = -2 and 0) in which risk judgment and judgment bias are significantly higher under a reward versus penalty frame (p-values < 0.001,

untabulated).[18] Overall, these result expand the effect of regulatory focus from decisions (Crowe and Higgins 1997) to beliefs.

[Figure 4]

Process Evidence

Judgment Process: Fraction Red

Recall that bag size is an irrelevant cue to risk judgments in rational decision-making. However, participants appear to incorporate this irrelevant cue in assessing risks. Instead of strictly using #Red-#White in assessing risks, participants appear to use the percentage of red balls in a bag (i.e., fraction red, see Table 2), which is a function of bag size, as a heuristic to estimate the likelihood of mislabeling. As shown in Figure 5 Panel A, under both frames, the assessed likelihood aligns more closely with fraction red than with the posterior likelihood of mislabeling (i.e., the Bayesian estimate of risk). This pattern is more evident in twelve-ball bags, of which fraction red deviates further from the posterior likelihood, than in six-ball bags. The potential use of fraction red as a heuristic, however, should bias against observing any differences in risk judgments between frames, because all participants inspected the same set of bags, which makes fraction red constant between frames. Yet, we still observe a marginal difference in judgments by frames when testing the RQ.

[Figure 5]

The fraction red heuristic also helps explain two patterns of judgment bias in Figure 4 Panel B. The first pattern is that participants overestimate the likelihood of mislabeling relative

[18] Participants identify the panels and bags identify the trials for the regression, which accounts for within-participants variance given multiple observations per participant. The dependent variable is $\log/(1-y))$ where y represents the assessed likelihood of mislabeling. We transform the assessed likelihood to remove its lower (0%) and upper bounds (100%). When y = 0% (y = 100%), we replace its value with 0.1% (99.9%) so that the transformed value is bounded away from negative (positive) infinity. The independent variables are Frame (rewards = 1; penalties = 0), Composition (#Red - #White), and their interaction. The regression yields a χ^2 (15) statistic of 16,701 and p < 0.001.

to the posterior likelihood when the risk of misstatement is low (i.e., judgment bias > 0 when #Red - # White < 0). This pattern is understandable if participants were using fraction red as a heuristic. As shown in Figure 5 Panel A, consistent with the first pattern, fraction red is higher than the posterior likelihood for twelve-ball bags when #Red - #White < 0. The second pattern is that participants underestimate the likelihood of mislabeling relative to the posterior likelihood when the risk of mislabeling is high (i.e., judgment bias < 0 when #Red - #White > 0). Also consistent with this pattern, as shown in Figure 5 Panel A, fraction red is lower than the posterior likelihood for twelve-ball bags when #Red - #White > 0. Overall, the observed judgment bias patterns are consistent with the use of a "fraction red" heuristic in twelve-ball bags.

Process from Judgments to Decisions: A Threshold

Skeptical judgments must reach a threshold to result in skeptical decisions (Nelson 2009). We thus assume participants will decide to test a bag only if their assessed risk exceeds their decision threshold for testing. A reward versus penalty frame can increase decisions to test (H1) by lowering the threshold for deciding to test (H2), increasing the assessed risk (RQ), or both. To better understand the underlying process, we examine how a reward versus penalty frame affects testing decisions while controlling for the assessed risks. If framing continues to affect testing decisions, then the increased risk judgment (RQ) and the reduced decision threshold for initiating tests (H2) jointly explain the increased testing (H1) in the reward frame. If there are no differences in testing decisions conditional upon participants' assessed risks, then the increased risk judgments (RQ) fully explain the increased decisions to test in response to framing (H1).

Figure 5 Panel B illustrates the percentage of bags tested within each *subjective* risk category, which is constructed based on participants' assessed likelihood of mislabeling (0 – 9%, 10-24%, 25-48%, 49-51%, 52-75%, 76-90%, and 91-100%) to closely align with the objective

risk categories provided in Table 2. We find that the reward frame continues to increase testing compared to the penalty frame in two subjective risk categories (52-75% and 76-90%, p-values < 0.05, untabulated).[19] This result suggests that (1) an increased risk judgment and (2) a reduced decision threshold for initiating tests *jointly* contribute to increased testing under a reward frame. Thus, the decision threshold plays a critical intermediary role in converting skeptical judgments into skeptical decisions, consistent with Nelson's (2009) statement. More research on the decision threshold can deepen our understanding of this conversion process from judgments to decisions.

Supplemental Analysis

Accuracy of Judgments and Decisions

Framing economically equivalent incentives as rewards or penalties did not change the accuracy of risk judgments or testing decisions.[20] We find that judgment error does not differ between the two frames (Table 3, p = 0.382). As reported earlier in Panel B of Figure 4, participants in both conditions over-estimate the risk when the posterior likelihood of mislabeling is less than 50% (#Red-#White < 0), and they under-estimate the risk when the posterior likelihood of mislabeling is more than 50% (#Red-#White > 0). On average, participants in the reward frame condition did not make more or less accurate risk judgments than those in the penalty frame condition.

Similarly, although the reward frame increases decisions to test relative to the penalty frame, participants decide to test even when evidence suggests that bags are likely correctly

[19] We estimate a logit panel regression, which controls for dependencies of repeated observations of the same participant. Participants identify the panels and bags identify the trials. The dependent variable is Decision (1 = test; 0 = not test) at the bag level. The independent variables are Frame (1 = rewards; 0 = penalties), Judgment (the seven subjective risk categories), and their interaction. The regression yields a χ^2 (12) statistic of 4,901 and p < 0.001.

[20] Inferences do not change if we use the difference of the untransformed hit and false alarm rates ($H - F$) as an alternative measure for decision accuracy ($t_{174} = 0.66$, p = 0.506, untabulated).

labeled (Figure 3, #Red-#White = -6 and -2), exhibiting presumptive doubt in testing. As a result, the false alarm rate is significantly higher under the reward versus penalty frame (Table 3, $p = 0.032$). Although the hit rate is directionally higher under the reward frame, this directional increase is offset by the significantly higher false alarm rate. As a result, neither decision accuracy (Table 3, $p = 0.574$) nor the percentage of correct decisions made (Table 3, $p = 0.465$) differs by frames.

Making accurate judgments and decisions with imperfect information consumes attentional resources (Griffith, Kadous, and Young 2021). We find that although increased attention to the task increases the accuracy of judgments and decisions, the amount of attention does not appear to differ between the two frames. We use two proxies for the unobservable on-task attention: (1) participants' performance on the seven comprehension checks, (2) self-reported task effort (0 = not at all; 10 = very much). Both proxies correlate negatively with judgment error, positively with decision accuracy, and positively with the percentage of correct decisions made (all p-values < 0.002, untabulated).[21] However, neither attention proxy differs by frames (p-values > 0.794, untabulated), consistent with no difference in accuracy between the two frames. Time spent also does not differ by frames ($p = 0.742$, untabulated).

The Role of Loss Aversion

Although loss aversion is a popular explanation for incentive framing effects (Hannan et al. 2005; Imas et al. 2017), we do not find loss aversion useful for predicting whether and how incentive framing affects testing decisions. As explained in the theory section, the audit setting

[21] Results are similar if we use participants' need for cognition score (Cacioppo, Petty, and Feng Kao 1984) as another proxy for on-task attention. The need for a cognition (Cacioppo and Petty 1982) represents individuals' disposition to engage in and enjoy effortful thinking, which consumes attentional resources. Consistent with the results on comprehension checks and self-reported task effort, we find that a higher need for cognition is correlated with lower judgment error, higher decision accuracy, and more correct decisions made (Spearman, all p-values < 0.013, untabulated). Thus, making accurate judgment and decision requires effortful thinking. Inferences about all hypotheses remain unchanged if we control for participants' need for cognition.

differs from typical managerial accounting settings in at least two ways. First, the desired behavior has *two* directions: (1) testing for effectiveness, and (2) not testing for efficiency. Second, auditors do *not* know ex ante which direction helps them avoid losses. Therefore, we do not expect loss aversion to explain our results.

As robustness checks, we examine whether loss aversion explains our results. We measure loss aversion using participants' choices in six hypothetical gambles adapted from prior research (Tversky and Kahneman 1991; Kahneman 1992; Harinck, Van Dijk, Van Beest, and Mersmann 2007). We find that participants in the penalty frame condition are more loss averse than those in the reward frame condition ($t_{174} = 2.83$, $p = 0.005$, untabulated). This result is consistent with prior research that a prevention focus (induced by a penalty frame; Shah, Higgins, and Friedman 1998) is associated with more loss aversion (Polman 2012), because the prevention focus makes participants experience losses more strongly than the promotion focus (Idson, Liberman, and Higgins 2000).

As expected, loss aversion does not explain our results. Untabulated analyses reveal that controlling for loss aversion does not change any inferences about our hypotheses or research question, and that loss aversion itself does not predict any dependent variables used in our analysis (i.e., percentage of bags tested, threshold bias, risk judgment, judgment bias, judgment error, decision accuracy, and the percentage of correct decisions made, all p-values > 0.228).

Additionally, despite being more loss averse, participants in the penalty frame condition do not report feeling more motivated to obtain the maximum pay (0: not at all, 10: very much; one-tailed $p = 0.922$) or feeling more motivated to avoid making incorrect decisions (0: strongly disagree, 6: strongly agree; one-tailed $p = 0.254$) than those in the reward frame condition. Despite being more loss averse, participants in the penalty frame condition do not spend more

time or exhibit increased on-task attention or superior accuracy in judgments and decisions. Overall, loss aversion does not appear to have a psychological or behavioral impact on our participants. By examining a theory that is not loss aversion, we broaden the understanding of how incentive framing affects contracting behavior in accounting.

V. CONCLUSION

In this study, we introduce and implement two theory-based criteria (Macmillan and Creelman 2004 Chapter 2) to incentivize an effective, efficient, and objective audit. Taking an experimental economics approach, we find that rewards induce more skepticism in risk judgments and testing decisions than economically equivalent penalties without sacrificing accuracy. We find that the increased skepticism reflects presumptive doubt rather than neutral skepticism. Presumptive doubt is required for fraud detection (Nelson 2009) and can also be beneficial when auditors exhibit insufficient skepticism and when material misstatements are likely. Our result highlights the need to reconsider the appropriate framing of auditors' incentives, supporting Peecher's (2013) proposal.

Advancing accounting theory on skepticism (Nelson 2009; Nolder and Kadous 2018), we demonstrate that the decision threshold plays a critical role in converting skeptical judgements to skeptical decisions. We connect the decision threshold with neutral skepticism and presumptive doubt (Figure 1), providing a way to distinguish the nature of state skepticism in testing decisions. Observing evidence consistent with regulatory focus theory (Higgins 1998) rather than loss aversion, we broaden our understanding of how incentive framing affects contracting behavior in accounting. Our results on risk judgments expand the effect of regulatory focus theory from decisions to beliefs.

The incentives we provide are based on outcomes obtained (misses, false alarms, correct rejections, hits). Outcome-based incentives are common in practice (Brazel, Jackson, Schaefer, and Stewart 2016; Xu and Kalelkar 2020). Future research can examine alternative ways of providing incentives, such as contracting on the decision process instead of the outcomes obtained (Peecher et al. 2013) or incentivizing presumptive doubt instead of neutral skepticism (Brazel et al. 2022). Continued research can help optimize the design of auditor incentives, a key determinant of audit quality.

REFERENCES

Auditing Concepts Committee. 1972. Report of the Committee on Basic Auditing Concepts. *The Accounting Review* 47: 15–74.

Bauer, T. D. 2015. The Effects of Client Identity Strength and Professional Identity Salience on Auditor Judgments. *The Accounting Review* 90 (1): 95–114.

Bhaskar, L. S. 2020. How do Risk-Based Inspections Impact Auditor Behavior? Experimental Evidence on the PCAOB's Process. *The Accounting Review* 95 (4): 103–126.

Bhaskar, L. S., P. E. Hopkins, and J. H. Schroeder. 2019. An Investigation of Auditors' Judgments When Companies Release Earnings Before Audit Completion. *Journal of Accounting Research* 57 (2): 355–390.

Bhaskar, L. S., T. M. Majors, and A. Vitalis. 2023. How does depletion interact with auditors' skeptical dispositions to affect auditors' challenging of managers in negotiations? *Contemporary Accounting Research* 40 (4): 2288–2313.

Blocher, E., R. P. Moffie, and R. W. Zmud. 1986. Report format and task complexity: Interaction in risk judgments. *Accounting, Organizations and Society* 11 (6): 457–470.

Bol, J., I. Grabner, K. Haesebrouck, and M. E. Peecher. 2022. Well-Calibrated Professional Skepticism: Its Benefits on Auditor Responsiveness to the Risk of Material Misstatement and Its Roots in Culture Controls and Auditor Values. Foundation for Audit Research Working Paper.

Bonner, S. E. 1999. Judgment and Decision-Making Research in Accounting. *Accounting Horizons* 13 (4): 385–398.

Bonner, S. E., and G. B. Sprinkle. 2002. The effects of monetary incentives on effort and task performance: theories, evidence, and a framework for research. *Accounting, organizations and society* 27 (4–5): 303–345.

Bonner, S., T. Majors, and S. Ritter. 2018. Prepopulating Audit Workpapers with Prior Year Assessments: Default Option Effects on Risk Rating Accuracy. *Journal of Accounting Research* 56 (5): 1453–1481.

Bowlin, K. O., J. L. Hobson, and M. D. Piercey. 2015. The Effects of Auditor Rotation, Professional Skepticism, and Interactions with Managers on Audit Quality. *The Accounting Review* 90 (4): 1363–1393.

Brazel, J. F., S. B. Jackson, T. J. Schaefer, and B. W. Stewart. 2016. The Outcome Effect and Professional Skepticism. *The Accounting Review* 91 (6): 1577–1599.

Brazel, J. F., J. Leiby, and T. J. Schaefer. 2022. Do Rewards Encourage Professional Skepticism? It Depends. *The Accounting Review* 97 (4): 131–154.

Brink, A. G. 2011. The effect of contract frame on the perceived fairness and planned effort under economically equivalent bonus, penalty, and combination contracts. *The Journal of Theoretical Accounting Research* 6 (2): 145.

Brink, A. G., and F. W. Rankin. 2013. The effects of risk preference and loss aversion on individual behavior under bonus, penalty, and combined contract frames. *Behavioral Research in Accounting* 25 (2): 145–170.

Burke, J., K. L. Towry, D. Young, and J. Zureich. 2023. Ambiguous Sticks and Carrots: The Effect of Contract Framing and Payoff Ambiguity on Employee Effort. *The Accounting Review* 98 (1): 139–162.

Cacioppo, J. T., and R. E. Petty. 1982. The need for cognition. *Journal of Personality and Social Psychology* 42 (1): 116–131.

Cacioppo, J. T., R. E. Petty, and C. Feng Kao. 1984. The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment* 48 (3): 306–307.
Carpenter, T. D. 2007. Audit Team Brainstorming, Fraud Risk Identification, and Fraud Risk Assessment: Implications of SAS No. 99. *The Accounting Review* 82 (5): 1119–1140.
Charness, G., U. Gneezy, and A. Imas. 2013. Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization* 87: 43–51.
Chen, C. X., J. E. Nichol, and F. H. Zhou. 2017. The Effect of Incentive Framing and Descriptive Norms on Internal Whistleblowing. *Contemporary Accounting Research* 34 (4): 1757–1778.
Christ, M. H., K. L. Sedatole, and K. L. Towry. 2012. Sticks and Carrots: The Effect of Contract Frame on Effort in Incomplete Contracts. *The Accounting Review* 87 (6): 1913–1938.
Church, B. K., T. Libby, and P. Zhang. 2008. Contracting Frame and Individual Behavior: Experimental Evidence. *Journal of Management Accounting Research* 20 (1): 153–168.
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.
Crowe, E., and E. T. Higgins. 1997. Regulatory Focus and Strategic Inclinations: Promotion and Prevention in Decision-Making. *Organizational Behavior and Human Decision Processes* 69 (2): 117–132.
Duru, A., and D. M. Reeb. 2002. International Diversification and Analysts' Forecast Accuracy and Bias. *The Accounting Review* 77 (2): 415–433.
Frederickson, J. R., and W. Waller. 2005. Carrot or Stick? Contract Frame and Use of Decision-Influencing Information in a Principal-Agent Setting. *Journal of Accounting Research* 43 (5): 709–733.
Garza, B. A. 2023. Inspectors' Incentive Perceptions and Assessment Timing: Inspectors' Requests and Auditors' Responses. *The Accounting Review* 98 (6): 197–221.
Gonzalez, G. C., V. B. Hoffman, and D. V. Moser. 2020. Do Effort Differences between Bonus and Penalty Contracts Persist in Labor Markets? *Accounting Review* 95 (3): 205–222.
Green, D. M., and J. A. Swets. 1966. *Signal Detection Theory and Psychophysics*. New York: Wiley.
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.
Griffith, E. E., K. Kadous, and D. Young. 2021. Improving Complex Audit Judgments: A Framework and Evidence. *Contemporary Accounting Research* 38 (3): 2071–2104.
Hammersley, J. S., J. Leiby, and C. Nielson. 2021. Improving Auditors' Review of Inconsistent Audit Evidence. *SSRN Electronic Journal*.
Hannan, R. L., V. B. Hoffman, and D. V. Moser. 2005. Bonus versus Penalty: Does Contract Frame Affect Employee Effort? *Experimental Business Research*: 151–169.
Harinck, F., E. Van Dijk, I. Van Beest, and P. Mersmann. 2007. When Gains Loom Larger Than Losses. *Psychological Science* 18 (12): 1099–1105.
Hawkins, E. M. 2017. When auditors' skeptical judgments do not lead to skeptical actions. PhD Thesis, University of South Carolina.
Higgins, E. T. 1998. Promotion and Prevention: Regulatory Focus as A Motivational Principle. *Advances in Experimental Social Psychology*: 1–46.

Higgins, E. T., and J. F. Cornwell. 2016. Securing foundations and advancing frontiers: Prevention and promotion effects on judgment & decision making. *Organizational Behavior and Human Decision Processes* 136: 56–67.
Hossain, T., and J. A. List. 2012. The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations. *Management Science* 58 (12): 2151–2167.
Hurtt, R. K. 2010. Development of a Scale to Measure Professional Skepticism. *AUDITING: A Journal of Practice & Theory* 29 (1): 149–171.
Hurtt, R. K., H. Brown-Liburd, C. E. Earley, and G. Krishnamoorthy. 2013. Research on Auditor Professional Skepticism: Literature Synthesis and Opportunities for Future Research. *Auditing: A Journal of Practice & Theory* 32: 45–97.
Idson, L. C., N. Liberman, and E. T. Higgins. 2000. Distinguishing Gains from Nonlosses and Losses from Nongains: A Regulatory Focus Perspective on Hedonic Intensity. *Journal of Experimental Social Psychology* 36 (3): 252–274.
Imas, A., S. Sadoff, and A. Samek. 2017. Do People Anticipate Loss Aversion? *Management Science* 63 (5): 1271–1284.
Kachelmeier, S. J., and R. R. King. 2002. Using Laboratory Experiments to Evaluate Accounting Policy Issues. *Accounting Horizons* 16 (3): 219–232.
Kachelmeier, S. J., T. Majors, and M. G. Williamson. 2014. Does Intent Modify Risk-Based Auditing? *The Accounting Review* 89 (6): 2181–2201.
Kachelmeier, S. J., and B. W. Van Landuyt. 2017. Prompting the Benefit of the Doubt: The Joint Effect of Auditor-Client Social Bonds and Measurement Uncertainty on Audit Adjustments. *Journal of Accounting Research* 55 (4): 963–994.
Kadous, K., S. J. 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): 759–778.
Kahneman, D. 1992. Reference points, anchors, norms, and mixed feelings. *Organizational Behavior and Human Decision Processes* 51 (2): 296–312.
Kang, Y. J., and D. Piercey. 2024. Would a Balanced PCAOB Inspection Approach Increase Auditors' Use of Innovative Audit Procedures? Working Paper. University of Massachusetts Amherst.
King, R. R., and R. Schwartz. 1999. Legal Penalties and Audit Quality: An Experimental Investigation*. *Contemporary Accounting Research* 16 (4): 685–710.
Kluger, A. N., and A. DeNisi. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin* 119 (2): 254–284.
Kowaleski, Z. T., B. W. Mayhew, and A. C. Tegeler. 2018. The Impact of Consulting Services on Audit Quality: An Experimental Approach. *Journal of Accounting Research* 56 (2): 673–711.
Kunda, Z. 1990. The case for motivated reasoning. *Psychological bulletin* 108 (3): 480.
Libby, R., and B. L. Lewis. 1977. Human information processing research in accounting: The state of the art. *Accounting, Organizations and Society* 2 (3): 245–268.
Luft, J. 1994. Bonus and penalty incentives contract choice by employees. *Journal of Accounting and Economics* 18 (2): 181–206.
Macmillan, N. A., and C. D. Creelman. 2004. *Detection Theory: A User's Guide*. Psychology Press.

Majors, T. M. 2016. The interaction of communicating measurement uncertainty and the dark triad on managers' reporting decisions. *The Accounting Review* 91 (3): 973–992.
McAllister, M., A. D. Blay, and K. Kadous. 2021. Fraud Brainstorming Group Composition in Auditing: The Persuasive Power of a Skeptical Minority. *Accounting Review* 96 (3): 431–448.
Mturk Data Blog. 2018. Tips For Requesters On Mechanical Turk : The BOT problem on Mturk. http://turkrequesters.blogspot.com/2018/08/the-bot-problem-on-mturk.html.
Nelson, M. W. 2009. A Model and Literature Review of Professional Skepticism in Auditing. *AUDITING: A Journal of Practice & Theory* 28 (2): 1–34.
Nichol, J. E. 2019. The Effects of Contract Framing on Misconduct and Entitlement. *The Accounting Review* 94 (3): 329–344.
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.
PCAOB. 2014. Staff Audit Practice Alert No. 8 Matters Related to Auditing Revenue in An Audit of Financial Statements (12).
PCAOB. 2017. Staff Audit Practice Alert No. 15 Matters Related to Auditing Revenue from Contracts with Customers (15).
PCAOB. 2024. A FIRM'S SYSTEM OF QUALITY CONTROL AND OTHER AMENDMENTS TO PCAOB STANDARDS, RULES, AND FORMS (2024).
Peecher, M. E., M. A. Ricci, and Y. (Daniel) Zhou. 2024. Promoting proactive auditing behaviors. *Contemporary Accounting Research* 41 (1): 620–644.
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.
Peer, E., J. Vosgerau, and A. Acquisti. 2013. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* 46 (4): 1023–1031.
Polman, E. 2012. Self–other decision making and loss aversion. *Organizational Behavior and Human Decision Processes* 119 (2): 141–150.
Quadackers, L., T. Groot, and A. Wright. 2014. Auditors' Professional Skepticism: Neutrality versus Presumptive Doubt. *Contemporary Accounting Research* 31 (3): 639–657.
Ramsay, R. J., and R. M. Tubbs. 2005. Analysis of Diagnostic Tasks in Accounting Research Using Signal Detection Theory. *Behavioral Research in Accounting* 17 (1): 149–173.
Ricci, M. A. 2022. How better client service performance affects auditors' willingness to challenge management's preferred accounting. *Accounting, Organizations and Society* 103: 101377.
Ricci, M. A., and D. Rimkus. 2025. Inconsistent responses to uncooperative client manager behavior: When auditors' judgments and actions diverge. *Accounting, Organizations and Society* 114: 101593.
Scholer, A. A., X. Zou, K. Fujita, S. J. Stroessner, and E. T. Higgins. 2014. When risk seeking becomes a motivational necessity. *Motivation Science* 1 (S): 91–115.
Shah, J., T. Higgins, and R. S. Friedman. 1998. Performance incentives and means: How regulatory focus influences goal attainment. *Journal of Personality and Social Psychology* 74 (2): 285–293.
Shaub, M. K., and J. E. Lawrence. 2002. A taxonomy of auditors' professional skepticism. *Research on Accounting Ethics* 8 (167–194).

Sprinkle, G. B., and R. M. Tubbs. 1998. The Effects of Audit Risk and Information Importance on Auditor Memory During Working Paper Review. *The Accounting Review* 73 (4): 475–502.

Tegeler, A. C., V. L. Brown, and D. H. Downey. 2024. Auditor Perceptions, Reactions, and Responses to PCAOB Inspection Feedback. *The Accounting Review*: 1–28.

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 & Cognition* 39 (7): 1275–1289.

Toplak, M. E., R. F. West, and K. E. Stanovich. 2014. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*.

Tversky, A., and D. Kahneman. 1991. Loss Aversion in Riskless Choice: A Reference-Dependent Model. *The Quarterly Journal of Economics* 106 (4): 1039–1061.

Van der Stede, W. A., A. Wu, and S. Y.-C. Wu. 2020. An Empirical Analysis of Employee Responses to Bonuses and Penalties. *The Accounting Review* 95 (6): 395–412.

Walther, B. R., and R. H. Willis. 2012. Do investor expectations affect sell-side analysts' forecast bias and forecast accuracy? *Review of Accounting Studies* 18 (1): 207–227.

Welsh, M., N. R. Burns, and P. Delfabbro. 2013. The Cognitive Reflection Test: how much more than Numerical Ability? *Proceedings of the Annual Meeting of the Cognitive Science Society* 35: 1587–1592.

Xu, Q., and R. Kalelkar. 2020. Consequences of Going-Concern Opinion Inaccuracy at the Audit Office Level. *Auditing: A Journal of Practice & Theory* 39 (3): 185–208.

Zhou, Y. (Daniel). 2020. Mitigating the Influence of Motivated Reasoning on Auditor Judgment. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network.

Zou, X., A. A. Scholer, and E. T. Higgins. 2014. In pursuit of progress: Promotion motivation and risk preference in the domain of gains. *Journal of Personality and Social Psychology* 106 (2): 183–201.

Table 1: Terminology

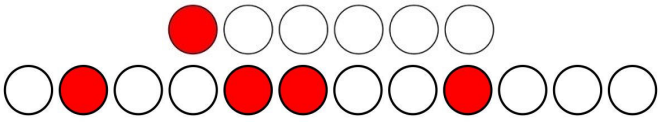
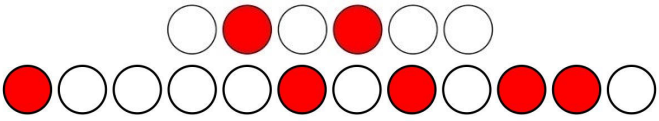
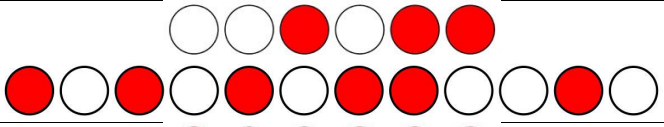
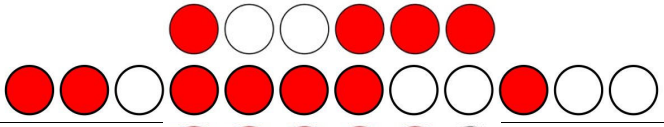
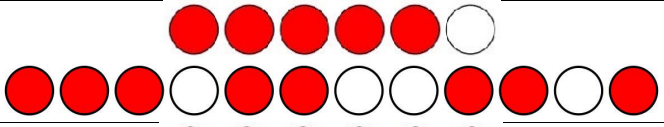
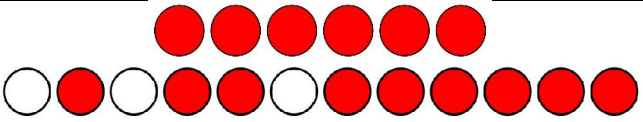
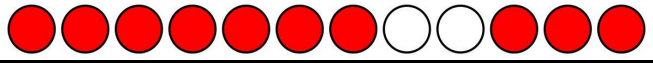
	Misstatement Absent	Misstatement Present
Decide Not to Test	Correct Rejection	Miss (Type II Error)
Decide to Test	False Alarm (Type I Error)	Hit

Table 2: Bags

Panel A: Composition of the 100 bags seen by each participant.

Risk of Mislabeling	Bag composition (#Red - #White)	Number of six balls bags	Number of twelve ball bags
Lowest	-6		4
	-4	6	3
	-2	12	11
	=	14	11
	+2	12	8
	+4	5	10
	+6	1	2
Highest	+8		1
		Total: 50	Total: 50

Panel B: Bag examples and the posterior likelihood of mislabeling.

Bag composition	Posterior likelihood	Fraction Red	Six and twelve ball bag examples
-6	8.1%	25.0%	
-4	16.5%	16.7% 33.3%	
-2	30.8%	33.3% 41.7%	
=	50.0%	50.0% 50.0%	
+2	69.2%	66.7% 58.3%	
+4	83.5%	83.3% 66.7%	
+6	91.9%	100.0% 75.0%	
+8	96.2%	83.3%	

Participants saw the same set of 100 bags presented in a randomized order. Bag composition is the difference between red and white balls. Fraction red is the number of red balls divided by the bag size.

Because each bag was independently drawn, and our IRB forbids researchers from deceiving participants, the distribution of bags was not symmetric by bag composition (e.g., the number of bags where #Red - #White = 4 is not the same as the number of bags where #Red - #White = -4).

The posterior likelihood is the probability that a bag is mislabeled, which increases as the number of red balls minus the number of white balls (#Red - #White) increases per bag, as illustrated below. Specifically, the conditional probability density function of a bag of N balls having X red balls, where $X \in \{0, 1, 2, \ldots, N\}$, is given by the probability mass function of the binomial distribution. The conditional probability that a heating bag has X red balls and N-X white balls is $f(X|\text{Heating}) = .4^{N-X}.6^{X}\binom{N}{X}$, and the conditional probability that a cooling bag has X red balls and N-X white balls is $f(X|\text{Cooling}) = .6^{N-X}.4^{X}\binom{N}{X}$. Using Bayes Rule, given X red balls observed, the posterior likelihood that the bag is a heating bag (mislabeled) is:

$$\frac{P_H\, f(X|\text{Heating})}{P_H\, f(X|\text{Heating}) + (1 - P_H)\, f(X|\text{Cooling})}$$

P_H is the prior probability (base rate) that the bag came from the heating distribution, which is 50%. The equation above reduces to $1/(1 + (2/3)^k)$, where k is the number of red balls minus the number of white balls observed. When the difference is zero (#Red - #White = 0), the posterior likelihood that the bag is a heating bag (mislabeled) equals the prior probability (50%). With more (fewer) white balls than red, the probability decreases (increases) from the prior.

Table 3: Tests of Hypotheses and Research Question

Descriptive statistics: mean (standard deviation)			**Testing Decisions**				**Risk Judgments**		
Frame	% Tested	Threshold Bias	Hit Rate	False Alarm Rate	Decision Accuracy	% Correct	Judgment Error	Assessed Likelihood	Judgment bias
Penalties N = 90	50.1% (13.5%)	-0.013 (0.408)	67.0% (18.3%)	33.8% (14.6%)	0.925 (0.528)	66.6% (9.4%)	11.1% (5.9%)	51.7% (4.7%)	0.3% (4.7%)
Rewards N = 86	54.7% (15.8%)	-0.160 (0.561)	70.7% (16.7%)	39.4% (19.3%)	0.881 (0.496)	65.6% (8.8%)	12.0% (6.4%)	53.4% (7.0%)	1.9% (7.0%)
Hypothetical Perfect Strategies:									
Version 1	39%	0.334	63.3%	15.7%	1.346	74%			
Version 2	64%	-0.447	85.7%	43.1%	1.240	71%			
Version 3	52%	-0.067	74.9%	28.6%	1.214	72%			
Prediction: Penalties – Rewards									
	H1: < 0	H2: > 0	No prediction	No prediction	No prediction	No prediction	No prediction	RQ	RQ
t (df = 174)	-2.10	1.99	-1.41	-2.12	0.56	0.73	-0.88	-1.86	-1.86
p-value	.019†	.024†	.162	.032	0.574	.465	0.382	0.064	0.064

†P-values are one-tailed, given directional predictions.

The perfect strategy refers to testing when #Red > #White, not testing when #Red < #White, and being indifferent between testing and not testing when #Red = #White. Being indifferent when #Red = #White means *any* combinations of decisions made are considered as perfect. For example, for three bags with #Red = #white, combinations such as (test, test, not test), (test, test, test), and (not test, not test, not test) are all considered as perfect. In this table, the three versions of perfect strategies are hypothetical and for illustration purposes only. The versions differ only in decisions made for the 25 bags with #Red = #White out of the 100 bags.
Perfect strategy version 1 refers to not testing any bags with #Red = #White.
Perfect strategy version 2 refers to testing all bags with #Red = #White.
Perfect strategy version 3 refers to randomly testing when #Red = #White. We report the mean value of 176 iterations of this strategy.

Dependent variables:
% Tested represents the percentage of bags tested out of the 100 bags per participant.

Threshold Bias = -0.5[$Z(H) + Z(F)$] calculated across 100 bags per participant. A zero value indicates an unbiased threshold that is neutral between testing and not testing. The more negative the value, the lower the testing threshold (relative to neutral) and the higher bias toward testing. The more positive the value, the higher the testing threshold (relative to neutral) and the higher bias towards not testing.

- *Z (H)* = the inverse of the standard normal cumulative distribution of the hit rate.
- *Z (F)* = the inverse of the standard normal cumulative distribution of the false alarm rate.
- Z (Rate) = negative infinity if Rate = 0%; Z (Rate) = 0 if Rate = 50%; Z (Rate) = positive infinity if Rate = 100%.

Hit rate is the percentage of mislabeled bags that are tested.

False alarm rate is the percentage of correctly labeled bags that are tested.

Decision accuracy = $Z(H) - Z(F)$ calculated across 100 bags per participant. A higher value indicates higher accuracy or sensitivity to whether a bag is mislabeled.

% Correct = (#hits + #correct rejections)/100 bags per participant. It is the percentage of correct decisions made based on the actual bag type.

Judgment Error = |assessed likelihood for a bag – the posterior likelihood of mislabeling for that bag|, averaged across 100 bags per participant. A higher value represents lower accuracy. See the posterior likelihood of mislabeling in Table 2.

Assessed likelihood is participants' assessed likelihood of mislabeling averaged across 100 bags.

Judgment bias = assessed likelihood for a bag – the posterior likelihood of mislabeling for that bag, averaged across 100 bags per participant. See the posterior likelihood of mislabeling in Table 2.

Independent variables:

We manipulate the *incentive frame* (penalty versus reward) between participants. Participants were informed that they would make $4 for inspecting 100 bags in the *penalty frame* condition. Additionally, they could pay a $2 penalty if more than 1/3 of their inspection decisions were incorrect. Participants would not pay the penalty if 1/3 or fewer of their decisions were incorrect. Participants were informed that they would make $2 for inspecting 100 bags in the *reward frame* condition. Additionally, they could earn a $2 bonus if 2/3 or more of their inspection decisions were correct. Participants would not earn the bonus if fewer than 2/3 of their decisions were correct. Incorrect decisions refer to decisions that result in misses or false alarms. Correct decisions refer to decisions that result in hits or correct rejections. See Table 1 for the four outcomes.

Figure 1: Distinguishing the Nature of State Skepticism in Testing Decisions

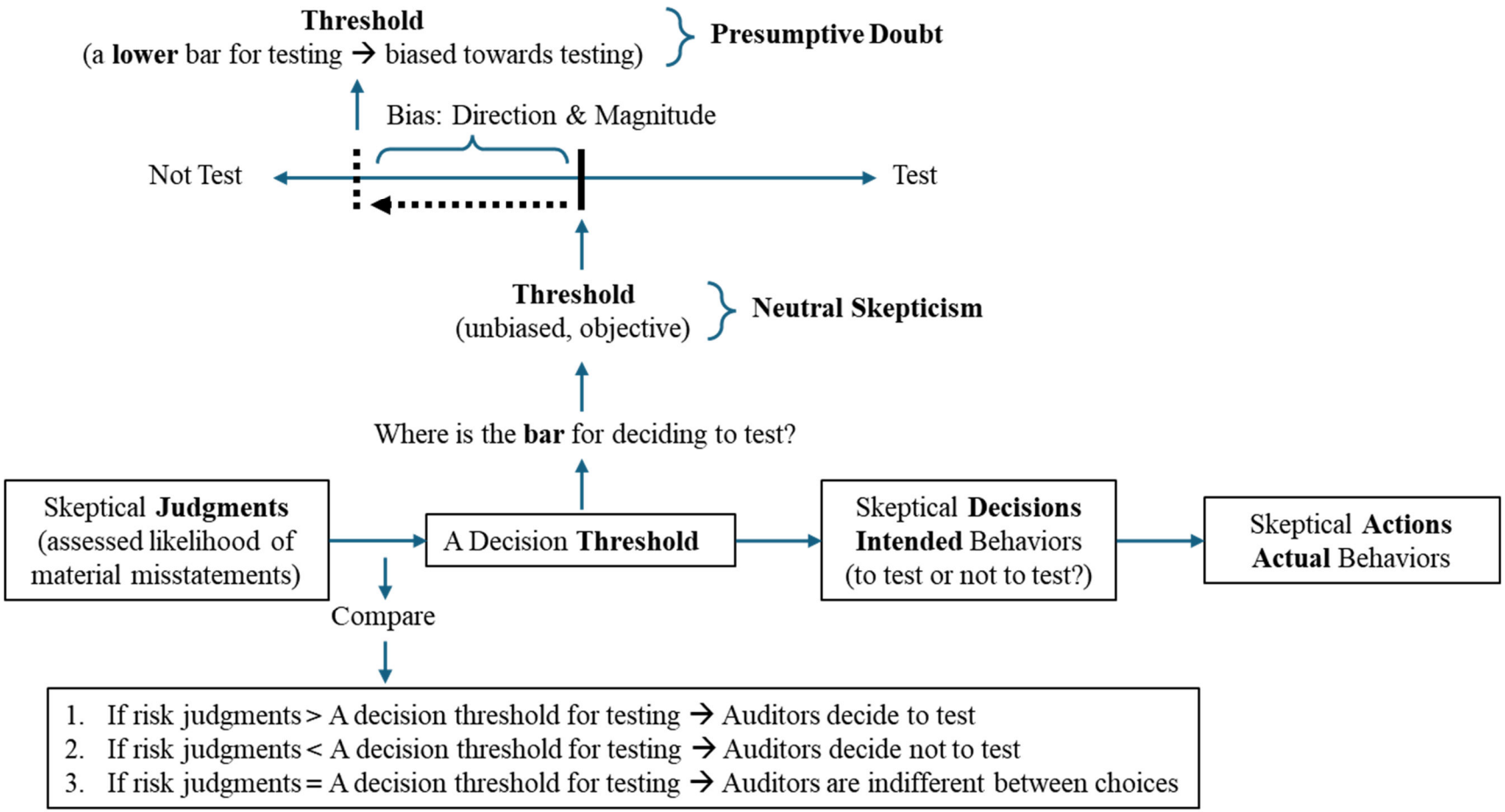


Figure 2: Summary of Experimental Procedures

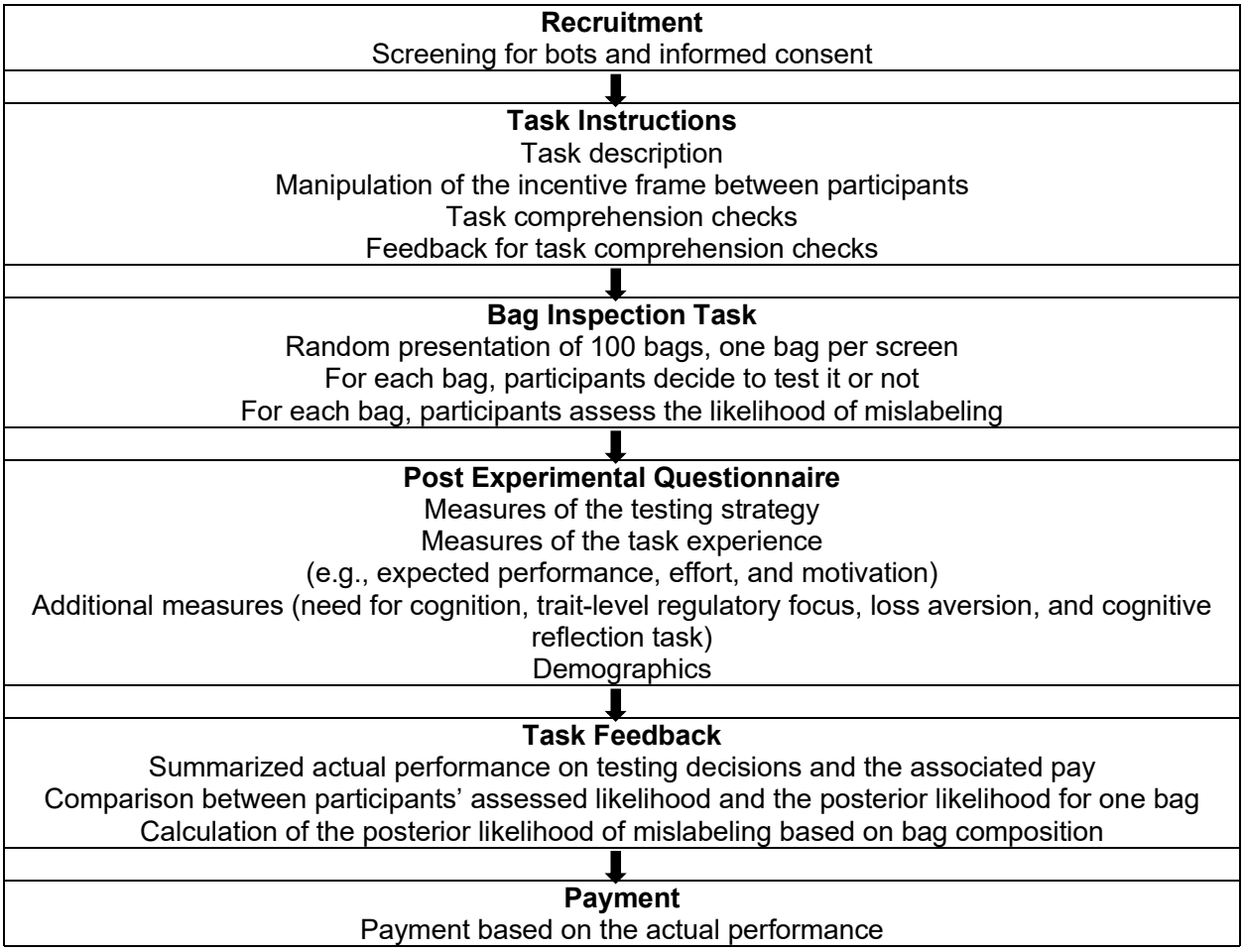
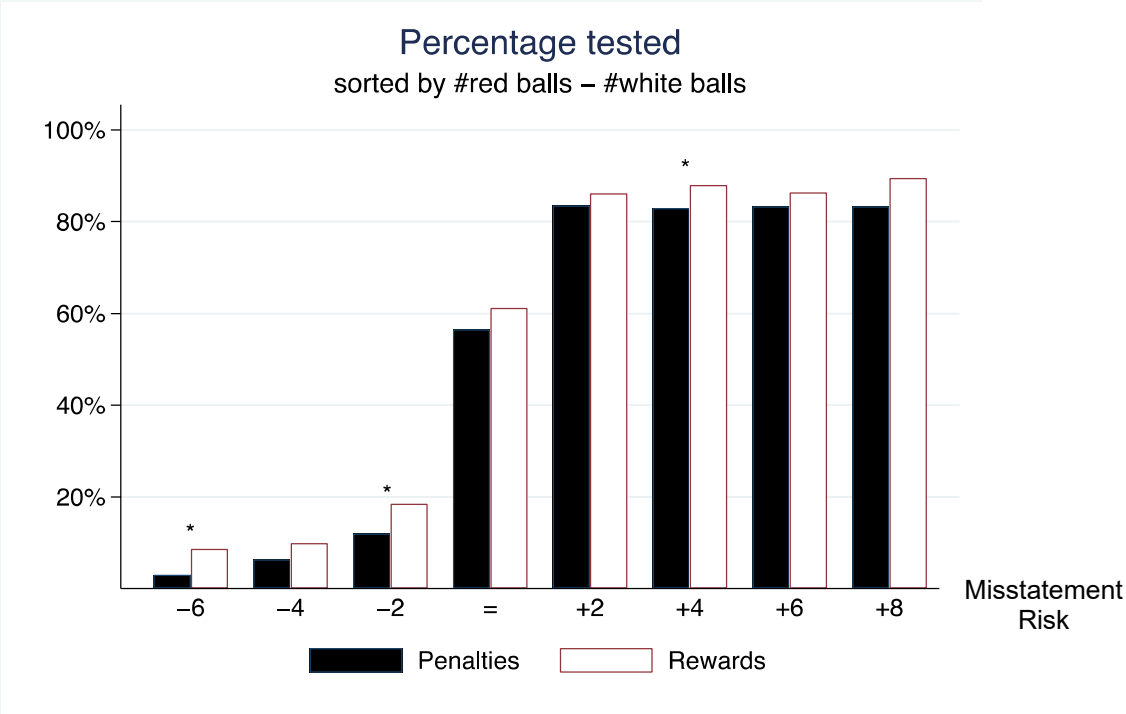


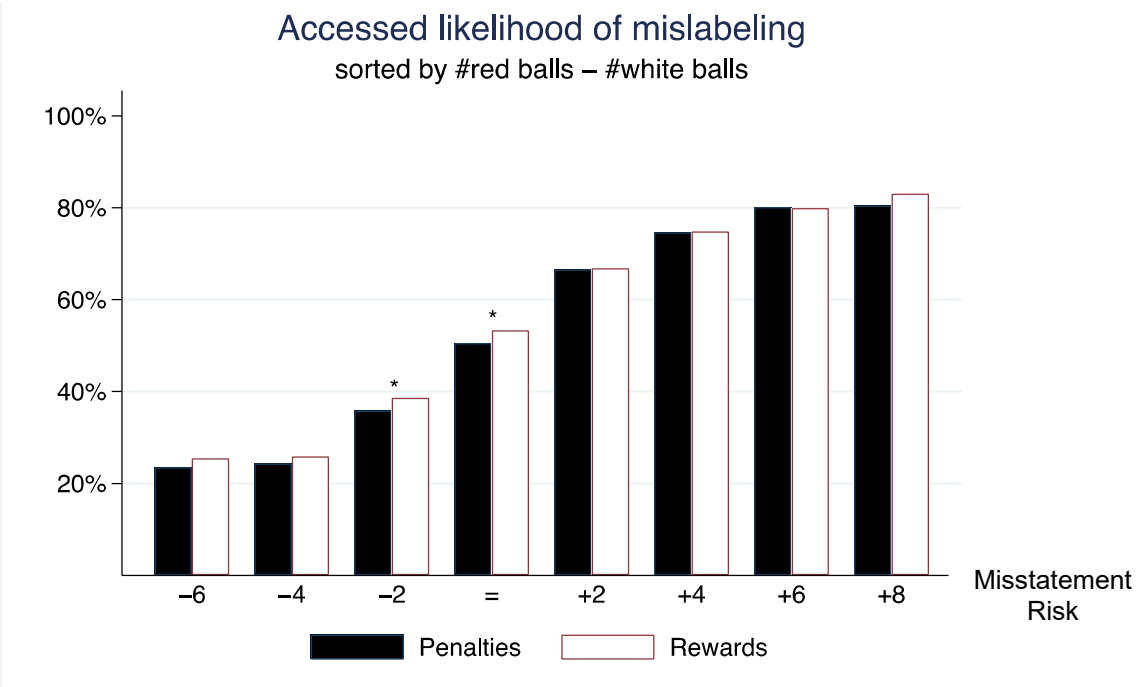
Figure 3: The Percentage of Bags Tested by Bag Composition (#Red - #White)



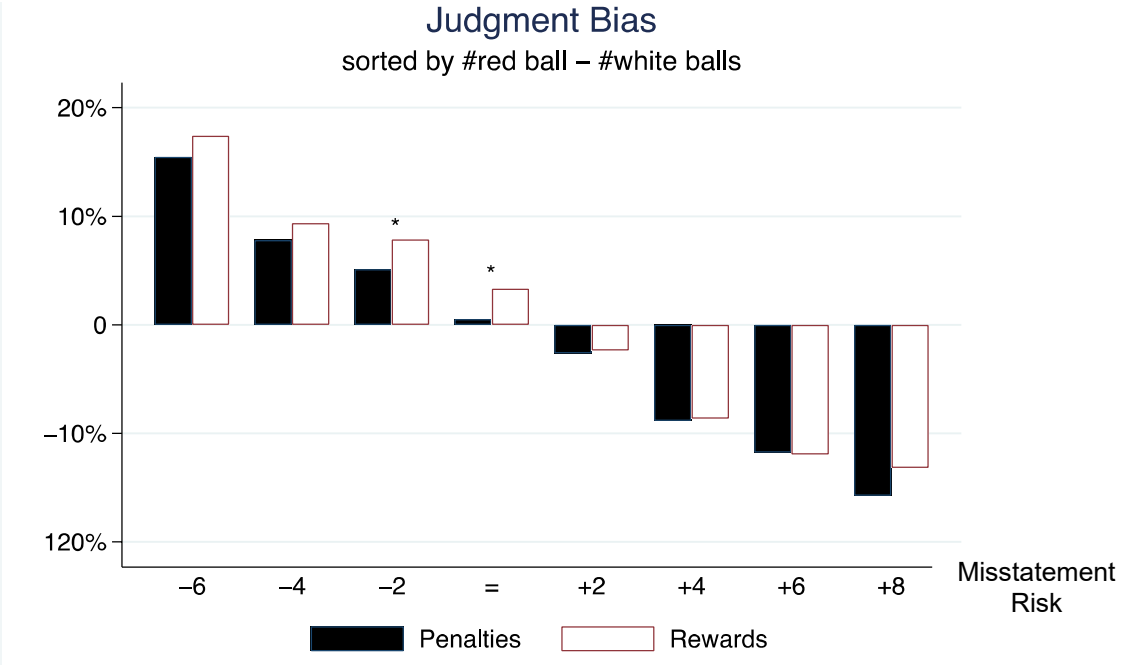
Note: * Different by incentive frame at a 5% level. See variable definitions in Table 2 and Table 3.

Figure 4: Risk Judgment and Judgment Bias by Bag Composition (#Red - #White)

Panel A: The assessed likelihood of mislabeling by bag composition



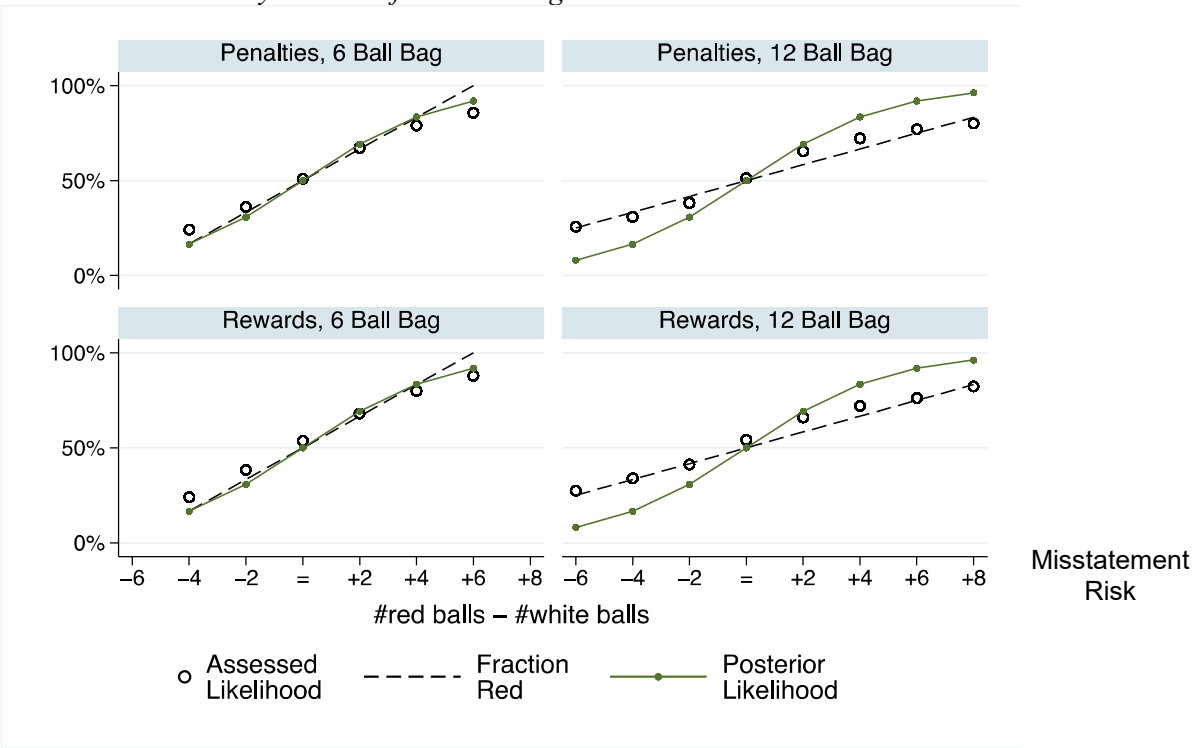
Panel B: Judgment bias by bag composition



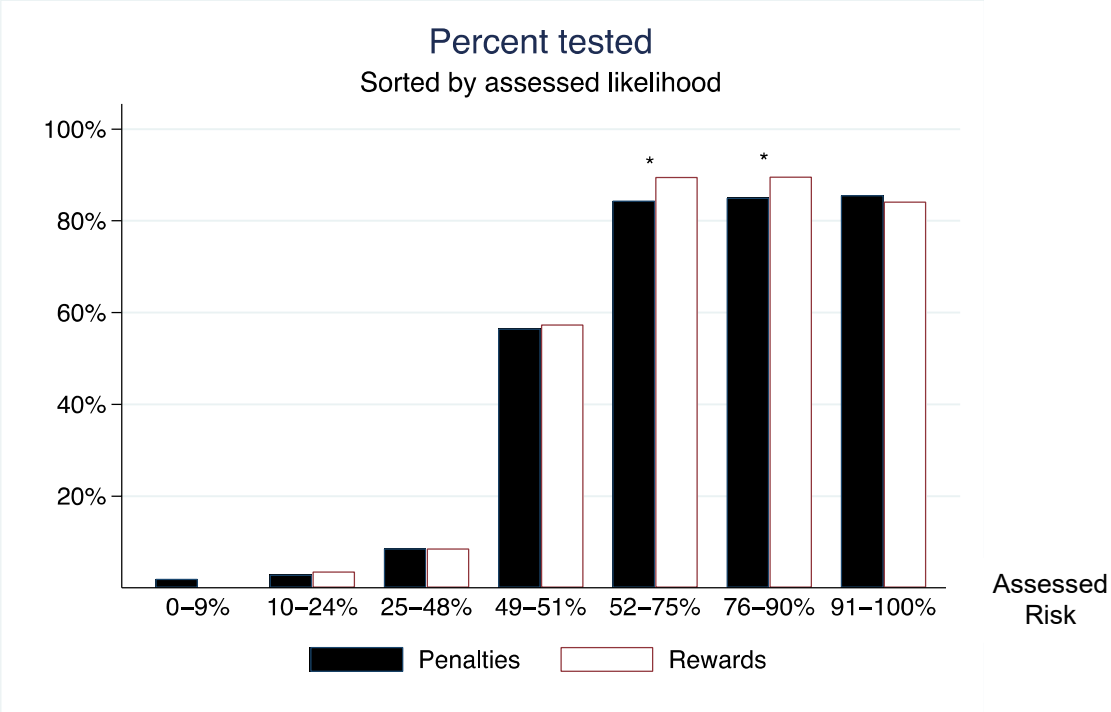
Note: * Different by incentive frame at a 1% level. All judgment bias is significantly different from zero at 1% level except for the #red = #white composition in the penalty frame condition. See variable definitions in Table 2 and Table 3.

Figure 5: Supplemental Analysis on the Judgment and Decision Process

Panel A: Assessed likelihood by incentive frame and bag size



Panel B: Percent tested conditioned on assessed likelihood



Note: *Different by incentive frame at a 5% level. See variable definitions at Table 2 and Table 3.

Appendix 1: Criteria for Incentivizing an Effective, Efficient, and Objective Audit

Notation

	State: No	State: Yes
Response: no	Outcome: Correct Rejection U_C	Outcome: Miss U_M
Response: yes	Outcome: False Alarm U_F	Outcome: Hit U_H

1. The Yes/No state refers to the presence/absence of a material misstatement.
2. The yes/no response refers to the decision to test/not to test.
3. U_i is auditors' utility for each possible outcome, where i denotes hit (H), miss (M), correct rejection (C), or false alarm (F).
    a. A miss is also called a type II error or a false negative.
    b. A false alarm is also called a type I error or a false positive.
    c. The shaded cells represent correct decisions based on the outcomes obtained.
4. P_H is the base rate of a material misstatement.

The Optimal Decision Threshold for Initiating Tests

Skeptical judgments must reach a threshold to result in skeptical decisions (Nelson 2009). Thus, we assume that auditors will decide to test only if their risk judgment exceeds their decision threshold for initiating tests. Equation (1) illustrates how rational auditors will choose the optimal decision threshold X_k that maximizes their expected utility. We will interpret and derive equation (1) later.

$$\frac{f(X_k|Yes)}{f(X_k|No)} = \frac{(1-P_H)}{P_H}\frac{(U_C-U_F)}{(U_H-U_M)} \quad (1)$$

Criterion 1: Set $(U_C - U_F)/(U_H - U_M) = 1$ to Incentivize an Effective and Efficient Audit

According to signal detection theory (Macmillan and Creelman 2004, 37 Chapter 2), the decision threshold maximizes "proportion correct" when the right-hand side of equation (1) equals $(1-P_H)/P_H$. Proportion correct refers to the number of correct decisions made (i.e., hits and correct rejections) as a percentage of the number of total decisions made. Thus, to maximize proportion correct, we should set $\frac{(1-P_H)}{P_H}\frac{(U_C-U_F)}{(U_H-U_M)} = (1-P_H)/P_H$, which reduces to $(U_C - U_F)/(U_H - U_M) = 1$.

Criterion 2: Set $(U_C - U_F)/(U_H - U_M) = P_H/(1-P_H)$ to Incentivize an Objective Audit

In signal detection theory (Chapter 2, Macmillan and Creelman 2004, 28; Ramsay and Tubbs 2005), an unbiased decision threshold is located at X_k where the two distributions (state =

Yes; state = No) intersect, which makes $\frac{f(X_k|Yes)}{f(X_k|No)} = 1 = \frac{(1-P_H)}{P_H}\frac{(U_C-U_F)}{(U_H-U_M)}$ in equation (1). Thus, to incentivize objective decision-making, we should set $(U_C - U_F)/(U_H - U_M) = P_H/(1 - P_H)$.

Interpreting Equation (1)

The optimal decision threshold X_k is a function of the *base rate* of a material misstatement P_H and auditors' utility U_i associated with the *four outcomes*, as summarized by the right hand-side of equation (1). U_i is a function of the benefits and costs associated with the four outcomes. Thus, auditors' incentives (pecuniary or nonpecuniary) affect U_i, which in turn affects X_k. Consistent with auditors should avoid over-testing and under-testing (Bowlin et al. 2015), we assume $\min\{U_H, U_C\} > \max\{U_M, U_F\}$ such that auditors derive greater utility from making correct than incorrect decisions based on the outcomes.

The optimal decision threshold X_k decreases (i.e., auditors set a lower bar for initiating tests) as the right hand-side of equation (1) decreases in value. For example, increased benefits of testing should lower the bar for testing for the decision threshold X_k to be optimal. Specifically, increased benefits of testing should increase auditors' utility from incurring hits U_H and/or false alarms U_F. Keeping P_H, U_C, and U_M constant, $(U_H - U_M)$ increases and $(U_C - U_F)$ decreases, which reduces the right-hand side of equation (1) and thereby lowers the optimal testing threshold X_k on the left-hand side. Thus, rational decision-making calls for a biased (more lenient) testing threshold in response to increased benefits of testing.

On the left hand-side of equation (1), $f(X|Yes)$ is the conditional probability density function of observing imperfect information X given the distribution of misstatement present (state = Yes). Likewise, $f(X|No)$ is the conditional probability density function of observing imperfect information X given the distribution of misstatement absent (state = No). The optimal testing threshold X_k is the value of imperfect information X at which rational auditors are indifferent between testing and not testing. Thus, auditors decide to test if $X > X_k$ and not test if $X < X_k$.

Deriving Equation (1)

The modeled auditor makes a risk judgment before deciding to test or not to test. Yes and No refer to two distributions: misstatement present (state = Yes) and misstatement absent (state = No). When making the risk judgment, the auditor assesses the posterior likelihood of the Yes state (i.e., misstatement present) given imperfect information X. This assessment considers the prior probability (base rate) P_H that the state is Yes, the conditional probability density functions of observing X given that the state is Yes $f(X|Yes)$, and the conditional probability density function of observing X given that the state is No $f(X|No)$.

Assume the auditor considers the consequences of testing versus not testing after seeing the private information X. $U_C,\ U_H,\ U_F\ and\ U_M$ are a function of the benefits and costs associated with the four outcomes. For example, U_H is a function of the reputation gain and cost of testing, and U_F is a function of the efficiency loss and cost of testing. We abstract away from modelling the specific and complete set of determinants of $U_C,\ U_H,\ U_F\ and\ U_M$, consistent with

prior research (Ramsay and Tubbs 2005; Macmillan and Creelman 2004). Also consistent with prior research that auditors should avoid over-testing and under-testing (Bowlin et al. 2015), we assume $\min\{U_H, U_C\} > \max\{U_M, U_F\}$ such that auditors derive greater utility from making correct than incorrect decisions based on the outcomes.

When the auditor decides to test, the outcomes can only be a hit or a false alarm, and the expected utility is:

$$\underbrace{\frac{P_H f(X|Yes)}{P_H f(X|Yes) + (1-P_H) f(X|No)}}_{posterior\ likelihood\ of\ Yes\ given\ X} U_H + \underbrace{\frac{(1-P_H) f(X|No)}{P_H f(X|Yes) + (1-P_H) f(X|No)}}_{posterior\ likelihood\ of\ No\ given\ X} U_F \tag{1a}$$

When the auditor decides not to test, the outcomes can only be a miss or a correct rejection, and the expected utility is:

$$\underbrace{\frac{P_H f(X|Yes)}{P_H f(X|Yes) + (1-P_H) f(X|No)}}_{posterior\ likelihood\ of\ Yes\ given\ X} U_M + \underbrace{\frac{(1-P_H) f(X|No)}{P_H f(X|Yes) + (1-P_H) f(X|No)}}_{posterior\ likelihood\ of\ No\ given\ X} U_C \tag{1b}$$

Thus, conditional upon the private information X, the auditor decides to test if equation (1a) is greater than equation (1b), not to test if equation (1a) is less than equation (1b), and is indifferent between testing and not testing if the two equations are equal in value. Denote X_k as the value of X where the auditor is indifferent between testing and not testing. To find X_k set equation (1a) equal to equation (1b), which then simplifies to equation (1). Solving equation (1) yields the optimal threshold X_k, which maximizes the auditor's expected utility.

The auditor decides to test if $X > X_k$, not to test if $X < X_k$, and is indifferent if $X = X_k$.

$$P_H\, f(X_k|Yes)\, (U_H - U_M) = (1-P_H)\, f(X_k|No)\, (U_C - U_F)$$

$$\Rightarrow \frac{f(X_k|Yes)}{f(X_k|No)} = \frac{(1-P_H)}{P_H} \frac{(U_C - U_F)}{(U_H - U_M)} \tag{1}$$