5-13-2022

# How Do Reward Versus Penalty Framed Incentives Affect Diagnostic Performance in Auditing?

Bright (Yue) Hong

Timothy W. Shields

# How Do Reward Versus Penalty Framed Incentives Affect Diagnostic Performance in Auditing?

## Comments

ESI Working Paper 22-06

# How Do Reward Versus Penalty Framed Incentives Affect Diagnostic Performance in Auditing?

Bright (Yue) Hong
(Corresponding Author)
School of Accountancy & MIS
Driehaus College of Business
DePaul University
yhong20@depaul.edu

Timothy W. Shields
Argyros School of Business and Management
Economic Science Institute
Chapman University
shields@chapman.edu

May 13, 2022

**Abstract:** Prior research examines how rewards versus economically equivalent penalties affect effort. However, accountants perform various diagnostic analyses that involve more than exerting effort. For example, auditors often need to identify whether a material misstatement is the underlying cause of a phenomenon among the possible causes. Testing helps identify the cause, but testing is costly. When participants are incentivized to test accurately (rather than test more) and objectively (unbiased between testing and not testing), we find that framing the incentives as rewards versus equivalent penalties increases testing by lowering the subjective testing criterion *and* by increasing the assessed risk of material misstatement. However, testing increases primarily when a misstatement is absent, causing more false alarms under a reward frame with no improvement in misstatement detection. Penalties are pervasive in auditing. Our study suggests that rewards are more effective for increasing testing, and that increasing testing blindly can impair audit efficiency.

Keywords: Frame, rewards, penalties, objectivity, accuracy, judgment, diagnostic tasks, experiment, auditing
JEL Classifications: C92, D82, D81, M40
Data Availability: Data are available from authors upon request.

We appreciate the feedback from Scott Asay, Allen Blay, Joseph Brazel, Mandy Cheng, Deni Cikurel, Kirsten Fanning, Gary Hecht, Wenqian Hu, Alex Johanns, Jane Jollineau, Robert Knechel, Amy Kristof-Brown, Marlys Lipe, Yi Luo, Elaine Mauldin, Molly Mercer, Mary Mindak, Nandu Nagarajan, Mark Peecher, Mark Penno, Steven Salterio, Donna Schmitt, Lisa Sedor, Ira Solomon, Bin Srinidhi, Jack Stecher, Matthew Stern, Ken Trotman, Ramgopal Venkataraman, Laura Wang, Michele Williams, Michael Williamson, Dan Zhou, workshop participants at Chapman University, DePaul University, Emory University, the University of Alberta, University of Illinois at Urbana-Champaign, University of Iowa, and University of New South Wales Sydney, University of Texas Arlington, and conference participants at the 2022 Auditing Section Mid-Year Meeting, the 2022 Hawaii Accounting Research Conference, the East Coast Behavioral Accounting Workshop, and the 2021 Accounting, Behavior and Organizations Research Conference. We thank Chapman University and DePaul University for their financial support.

# I. INTRODUCTION

Incentives are *anything* that motivates behavior. Neo-classical economic theory predicts that framing equivalent incentives as rewards versus penalties should not change agents' behavior. However, a stream of research finds that imposing penalties versus economically equivalent rewards increases agents' effort and effort-based productivity (e.g., Hannan, Hoffman, and Moser 2005; Church, Libby, and Zhang 2008; Brink 2011; Hossain and List 2012). This stream of research, however, has not examined diagnostic performance that is also a part of many accountants' work, such as identifying the cause of labor efficiency variances (Brown 1987), the cause of missing earnings targets, the cause of differences between tax and book income, and the cause of cash flow shortage. Effort is not the only determinant of diagnostic performance. Therefore, we know little about whether and how incentive framing affects diagnostic performance.

Examining this question is particularly important and relevant to auditing. First, auditing is diagnostic in nature. Auditors must obtain reasonable assurance about whether a material misstatement is the cause of a phenomenon among the possible causes (AS 1001.02; AS 1101.03). Second, the design of auditor incentives affects the supply of audit quality (Watts and Zimmerman 1981; DeFond and Zhang 2014). Although penalties are less preferred by agents and less commonly used than rewards in designing incentives (Luft 1994; Brink and Rankin 2013), auditors face penalties in various forms and from various sources, such as criticism from internal and external inspectors, reduced performance ratings from supervisors, scrutiny from the media, lawsuits from investors, fines from regulators, and license suspension from professional oversight bodies (Knechel, Salterio, Ballou 2007; Peecher, Solomon, and Trotman 2013). Third, prior research examines the provision (providing versus not providing rewards) rather than the

framing (providing rewards versus economically equivalent penalties) of auditor incentives (Barr-Pulliam, Brazel, McCallen, and Walker 2020; Brazel, Leiby, and Schaefer 2021). We take the first step to examine the framing of auditor incentives regardless of their forms or sources.

Diagnostic performance is challenging because different causes can lead to the same phenomenon. Of interest to auditors is whether a material misstatement is the underlying cause. For example, if auditors observe an increase in retail sales but a decrease in the number of retail stores, a possible cause is that sales are overstated, and another possible cause is that the client is shifting to online retailing. Testing increases the chance of identifying the cause. However, testing consumes the audit budget and delays completion. Auditors have a limited budget and generally fewer than 60 days to complete an annual audit (SEC 2002). Therefore, it may not be possible to test every phenomenon. Auditing standards prescribe risk-based testing (AS 1101.10, 11). More testing is needed if the risk of material misstatement is high than when it is low. We examine risk judgment (i.e., the assessed risk of material misstatement) and testing action (i.e., test or not) in diagnostic performance.

In a binary diagnostic task where a misstatement is either present or absent, and where auditors decide to either test or not test, there are four possible outcomes (see Table 1): a false alarm (misstatement absent/test), a hit (misstatement present/test), a correct rejection (misstatement absent/not test), and a miss (misstatement present/not test). Drawing on prior research (e.g., Blocher, Moffie, and Zmud 1986; Brown 1981; Sprinkle and Tubbs 1998; Ramsay and Tubbs 2005; Macmillan and Creelman 2004), we assume that auditors decide to test or not by comparing their risk judgment to a subjective testing criterion (i.e., threshold) that could vary across auditors. Auditors are more likely to test if the assessed risk of misstatement exceeds the subjective threshold and less likely to test if otherwise.

Objectivity and accuracy are desired qualities of diagnostic performance. Auditing standards generally prescribe objective judgments and actions. For example, AS 1015.08 requires "objective evaluation of evidence", and AS 1015.09 states that "the auditor neither assumes that management is dishonest nor assumes unquestioned honesty" when evaluating and requesting evidence. Additionally, auditing standards prescribe risk-based testing rather than more testing regardless of the risk of material misstatement. The former values the accuracy of judgments and actions, whereas the latter does not. Under the risk-based testing approach, it is critical that auditors make accurate risk judgments such that testing is likely to result in hits rather than false alarms, and that not testing is likely to result in correct rejections rather than misses.

When auditors are incentivized to be objective and accurate, we predict that framing the incentives as rewards versus equivalent penalties changes performance objectivity by lowering the subjective testing criterion. Given the same risk judgment, the lower the testing criterion, the more likely auditors will test. Therefore, we predict that a reward frame increases testing. The subjective testing criterion is determined, in part, by how auditors weigh the four outcomes in binary diagnostic tasks. If auditors weigh ensuring hits and ensuring correct rejections equally (i.e., weigh avoiding misses and avoiding false alarms equally), then auditors should choose an objective criterion and be unbiased between testing and not testing (also see Nelson's 2009 discussion on the "neutral" view of professional skepticism). Supervisor preference, testing cost, and the design of incentives can affect how auditors weigh the outcomes.

Crowe and Higgins (1997) find that a reward versus penalty frame increases the weights on ensuring hits and avoiding misses in memory recognition. Unlike *remembering* the previous occurrence of a phenomenon as in memory-recognition tasks, diagnosing the cause of a

phenomenon among the possible causes requires *reasoning*. Nonetheless, if a reward frame also increases the focus on ensuring hits and avoiding misses in diagnostic tasks, a reward frame should lower the subjective testing criterion and increase testing. Given the same testing criterion, the higher the assessed risk, the more likely auditors will test. We further explore whether framing also affects the assessed risk to better understand whether the predicted increase in testing results from a change in the subjective criterion, the risk judgment, or both. Crowe and Higgins (1997) do not examine the underlying beliefs of memory performance and thus do not allow us to predict whether framing affects risk judgment.

We further predict that framing equivalent incentives as penalties versus rewards increases the accuracy of risk judgments and testing actions when auditors are incentivized to be objective and accurate. Prior research finds that penalties versus rewards of the same magnitude increase on-task attention (Yechiam and Hochman 2013, 2014), and that increased attention improves the accuracy of signal detection (Broadbent and Gregory 1963). Accurately diagnosing the cause of a phenomenon requires reasoning, which consumes attentional resources. Therefore, increased on-task attention induced by a penalty frame should improve diagnostic accuracy. If we interpret on-task attention as effort (Kahneman 1973), our reasoning is consistent with prior findings that imposing a penalty frame increases agents' effort and effort-based productivity (e.g., Brink 2011; Church et al. 2008; Hannan et al. 2005). In our case, cognitive effort increases diagnostic accuracy.

We test our predictions in a between-participants experiment, holding participants' incentives equivalent between the reward and penalty frames. Participants are provided with 100 bags and asked to identify whether each bag is mislabeled (i.e., the cause) based on its appearance (i.e., the phenomenon). An incorrect label represents a material misstatement. Key to

diagnostic decision-making, a mislabeled and correctly labeled bag can look the same. Attending to relevant information and engaging in effortful reasoning should help participants infer whether a bag is mislabeled. Participants are asked to assess the likelihood of mislabeling for each bag and decide whether they will test the bag. Testing reveals whether a misstatement exists, but testing is costly. Participants are incentivized to test objectively and accurately under both frames, consistent with what auditing standards prescribe. By using real distributions to generate the bags, we establish normative benchmarks to evaluate performance objectivity and accuracy.

We find that incentive framing changes performance objectivity rather than accuracy. As predicted, a reward frame lowers the subjective testing criterion and increases testing. Interestingly, participants are unaware of the corresponding change in how they weigh the hits (versus correct rejections) and misses (versus false alarms), which suggests that framing affects the subjective testing criterion automatically and unconsciously. Although unpredicted, the assessed risk of mislabeling is marginally higher under a reward than a penalty frame. Controlling for the assessed risk reduces but does not eliminate the effect of framing on testing. Therefore, a reward frame increases testing by increasing the assessed risk *and* by lowering the testing criterion. Also unpredicted, participants facing the reward frame increase testing mostly when a bag is correctly labeled rather than mislabeled, causing more false alarms under a reward versus penalty frame with no improvement in misstatement detection. Therefore, increasing testing blindly can reduce audit efficiency.

Our study makes several contributions. First, we examine the effect of incentive framing on diagnostic performance. Prior incentive framing research focuses on effort-based productivity such as translating symbols (e.g., Church et al. 2008) and completing sliders (e.g., Imas, Sadoff, and Samek 2017), but effort is not the only determinant of diagnostic performance. As reported

later, our proxies for effort (i.e., on-task attention), increase diagnostic accuracy but do not predict testing. In fact, we observe that framing affects performance objectivity with no effect on our proxies for effort. By examining the key audit constructs of risk judgments and testing actions, we provide new insight into how framing affects performance objectivity beyond memory recognition. We find that a reward frame increases testing by lowering the subjective testing criterion *and* by increasing the assessed risk. Although we focus on the audit setting, our findings should apply to other accounting settings of diagnostic performance.

Second, by examining the objectivity and accuracy of diagnostic performance, we provide a more precise understanding of how auditor incentives can be designed to motivate the desired behavior. If heightened risk judgments and more testing are desired (i.e., professional skepticism under "presumptive doubt" about management assertions, see Nelson 2009), then framing auditor incentives as rewards is more effective than penalties. Encouragingly, regulators recently started to acknowledge the good practices of accounting firms in inspection reports, which used to disclose audit deficiencies only (PCAOB 2019, 2020). Conversely, if accurate risk judgments and testing actions are desired, framing auditor incentives as rewards is not more effective than penalties. In fact, increasing testing blindly can reduce audit efficiency. Accounting firms who are concerned with audit costs may consider framing auditor incentives as rewards on high-risk audits where increased testing is necessary.

Finally, we contribute to the research methodology by designing a task that is effective and efficient for examining diagnostic performance. Our task captures the nature of diagnostic decision-making by allowing different causes for the same phenomenon and by requiring effortful reasoning for identifying the cause. Our task utilizes real distributions to establish normative performance benchmarks for evaluating objectivity and accuracy, which are important

qualities of judgments and actions. Our task features parameters (e.g., incentives) that can be easily changed for future research. To understand whether and how incentive framing affects diagnostic performance, we incentivize participants to be objective and accurate, following what auditing standards prescribe, but this need not be the case for future research. Our task employs an abstract setting that does not require specialized knowledge but general knowledge that novice participants possess. We hope our task will be helpful to the broader field of experimental accounting.

## II. THEORY AND HYPOTHESES

### Objectivity and Accuracy in Diagnostic Performance

Objectivity and accuracy are desired qualities of diagnostic performance. Auditing standards generally prescribe objective evaluation of evidence, requiring auditors to neither assume that a misstatement exists nor assume that a misstatement does not exist in evaluating and gathering evidence (AS 1015.08, 09). Additionally, auditing standards prescribe a risk-based testing approach. More testing is needed if the risk of material misstatement is high, and less testing is needed if the risk of material misstatement is low (AS 1101.10, 11). This approach balances audit effectiveness with audit efficiency, recognizing that testing is costly to meeting tight budgets and deadlines. Under this approach, it is critical that auditors make accurate risk judgments such that testing targets higher risk areas where a material misstatement is more likely to exist.

Distinguishing objectivity from accuracy is important for evaluating auditor performance (Ramsay and Tubbs 2005). For example, if the risk of material misstatement is high, and if an auditor increases testing, it is unclear whether the increased testing is based on a biased assumption or an accurate inference that a misstatement is likely to exist. If this auditor also

increases testing when the risk of material misstatement is low, then it is clear that this auditor cannot accurately distinguish whether a misstatement exists, and that testing is based on a biased assumption that a misstatement is likely to exist regardless of the changes in the phenomenon. On the other hand, if this auditor decreases testing when the risk of material misstatement is low, then it is clear that this auditor can accurately distinguish the presence or absence of a misstatement, and that testing is based on an objective evaluation of the phenomenon.

**Objectivity**

Drawing on prior research (e.g., Blocher et al. 1986; Macmillan and Creelman 2004), we assume that an auditor decides to test a phenomenon by comparing the assessed risk with a subjective testing criterion. An auditor is more likely to test if the assessed risk exceeds the criterion. Given the same risk judgment, the higher the criterion, the less likely the auditor will test. Although auditing standards generally prescribe objective judgments and actions, several factors can cause auditors to choose a testing criterion that deviates from a neutral standpoint (Sprinkle and Tubbs 1998; Ramsay and Tubbs 2005). Among these factors are the frequency of a material misstatement that an auditor encounters in prior experience and the relative weight that an auditor places on the four outcomes (false alarms, hits, correct rejections, and misses) in binary diagnostic tasks.

For example, if an auditor has only worked on audits of restated financial statements, this prior experience can cause the auditor to presume a higher than warranted base rate of misstatements, lower the testing threshold, and be biased towards testing in future audits, *ceteris paribus*. Additionally, if an auditor's supervisor punishes the auditor for incurring false alarms, as commonly observed in practice and research (Brazel, Jackson, Schaefer and Stewart 2016; Brazel, Gimbar, Maksymov, and Schaefer 2019), the anticipation of supervisor punishment can

cause the auditor to weigh avoiding false alarms more than avoiding misses. As a result, the auditor will increase the testing threshold and be biased towards not testing in future audits, *ceteris paribus*.

We predict that a reward, as opposed to a penalty, frame causes auditors to lower their subjective testing criterion and increase testing (see Appendix 2 for arguments based on expected utility). Research on memory recognition finds that, when participants were asked to indicate whether a picture or a word currently presented was provided in a previous list, a reward versus penalty frame causes participants to ensure hits and avoid misses (Crowe and Higgins 1997; Levine, Higgins, and Choi 2000; Bowen, Marchesi, and Kensinger 2020). Obviously, in diagnostic tasks auditors are not asked to indicate whether they remember the previous occurrence of a phenomenon but are asked to infer whether a material misstatement is the cause of the phenomenon among the possible causes. Although diagnostic performance requires reasoning whereas memory performance does not, if a reward frame also causes auditors to be more concerned with ensuring hits and avoiding misses, auditors should lower their subjective testing criterion and be more inclined to test under a reward frame.

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

**H2:** *A reward frame lowers the subjective testing criterion relative to a penalty frame.*

We explore whether auditors' risk judgments also contribute to the predicted increase in testing under a reward versus penalty frame. Recall that auditors are assumed to test a phenomenon if their assessed risk exceeds their subjective testing criterion, and not to test if otherwise. Therefore, increased testing (H1) could result from a decrease in the subjective testing criterion (H2), an increase in the assessed risk (i.e., the risk judgment underlying the testing action), or both. Prior research on memory recognition does not examine participants' underlying

beliefs (Crowe and Higgins 1997; Levine et al. 2000; Bowen et al. 2020), and therefore cannot speak to whether framing affects the assessed risk. Therefore, we pose a research question.

**RQ:** *Does the incentive frame affect the assessed risk of material misstatement?*

**Accuracy**

Recall that auditing standards require risk-based testing rather than more testing regardless of risk, and that accurate risk judgments are critical in risk-based testing. Auditor expertise increases the accuracy of risk judgments by facilitating the effective selection and representation of informational cues from a phenomenon (Bonner 1990; Hammersley 2006). The quality of information available for risk judgments and the amount of attention that auditors invest in risk judgments can also affect the accuracy of risk judgments and the risk-based testing. When auditors are incentivized to be accurate, we predict that framing the same incentives as penalties versus rewards increases the accuracy of risk judgments and testing actions by increasing the cognitive effort (i.e., attention) that auditors invest in diagnostic tasks.

Accounting research finds that imposing penalties as opposed to economically equivalent rewards motivates participants to increase productivity by increasing effort (e.g., Hannan et al. 2005; Church et al. 2008; Brink 2011; Hossain and List, 2012). A common explanation is that the disutility of losing one dollar is greater than the utility of gaining one dollar (i.e., loss aversion, see Kahneman and Tversky 1979). Therefore, participants facing penalties versus rewards (i.e., potential losses versus gains) are more motivated to increase effort or productivity when doing so helps them avoid losses. When increased accuracy helps auditors avoid losses, a penalty frame should motivate more cognitive effort and improve diagnostic accuracy.

Additionally, attention research shows that imposing penalties versus rewards of the same magnitude (e.g., losing one dollar versus gaining one dollar) improves feedback-based learning

when participants' attention is depleted (Yechiam and Hochman 2014), suggesting that penalties increase on-task attention relative to rewards (Yechiam and Hochman 2013). Attention has the same meaning as effort (Kahneman 1973). In this sense, attention research and accounting research are consistent on the implications of incentive framing. Heightened attention increases the accuracy of detecting sounds (Broadbent and Gregory 1963). If a penalty frame increases on-task attention, heightened attention should also increase the accuracy of detecting material misstatements.

**H3:** *A penalty frame increases the accuracy of risk judgments and testing actions relative to a reward frame.*

Note that H3 does not imply that the answer to our research question is yes. In other words, increased judgment accuracy under a penalty frame (H3) does not allow us to predict that incentive framing affects the level of the assessed risk (RQ). For example, if the actual risk (i.e., likelihood) of material misstatement is 50 percent, assessing a 45 percent likelihood is more accurate than assessing a 40 or 60 percent likelihood. However, the level of a more accurate risk judgment (45 percent) can be either below (60 percent) or above (40 percent) the level of a less accurate risk judgment. Therefore, increased judgment accuracy does not predict the level of the assessed risk. Also, note that H1 does not imply H3. That is, testing more does not imply that testing is less accurate. Additionally, H2 does not imply H3, meaning that auditors can increase or decrease their subjective testing criterion independent of their judgment accuracy or testing accuracy.

# III. METHOD

## Participants

We recruit 196 participants from Amazon Mechanical Turk for our incentivized experiment.[1,2] We argue that using MTurk participants is efficient and effective for testing our theory: imposing rewards decreases the subjective testing criterion and diagnostic accuracy relative to imposing economically equivalent penalties. To evaluate the testing criterion and diagnostic accuracy precisely, we must establish normative benchmarks for what the "correct" risk judgments and testing actions are. We do so by designing an abstract task with real distributions underlying the risk of material misstatement. Our task does not require specialized auditor knowledge but general knowledge that MTurk participants possess. Thus, using auditor participants for our task would be unnecessary. Using auditor participants when unnecessary may reduce other researchers' access to these valuable participants (Libby, Bloomfield, and Nelson 2002) and create a tragedy of the commons (Garrett 1968).

We argue that our findings are likely applicable to auditors for three reasons. First, prior research suggests that novice participants are reasonable proxies of professionals when tasks do not require specialized knowledge (Elliott, Hodge, Kennedy, and Pronk 2007; Bolton, Ockenfels, and Thonemann 2012). Indeed, Hong (2022) replicates the effects observed among auditors on complex judgment with novice participants in a judgment task that requires only general knowledge. Second, as we discuss later, framing affects testing (H1) and the subjective testing criterion (H2) automatically and unconsciously. Like other experimental participants, auditors

[1] To reduce the risk of collecting low quality data, we 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 2014; Mturkdata 2018). Additionally, we restricted participants to high school graduates and above.
[2] We obtained approval from the Institutional Review Board of a west coast university where the experiment took place.

are susceptible to automatic, unconscious effects, which are difficult to eliminate even with targeted interventions (Brazel et al. 2019).[3] Third, compared to novice participants, auditors are likely to allocate more cognitive effort to tasks that are relevant to their work (i.e., initially more engaged, see Hong 2022). Higher effort increases diagnostic accuracy, as we argue in developing H3. If auditors are more accurate than novice participants, incentive framing is unlikely to have an incremental effect on accuracy for auditors. We find that framing does not affect accuracy even with novice participants. Therefore, inferences about H3 are unlikely to change with auditor participants.

**Bag Inspection Task**

Participants performed a diagnostic task in which they assessed the risk of misstatement and decided whether they will test any potential misstatement, following the procedures in Figure 1. Specifically, participants assumed an auditor's role for a hypothetical company that makes heating and cooling bags for physical therapy. Due to a mistake in production, all bags were labeled as cooling bags. The company would not change a bag label unless testing revealed that the bag was a heating bag. A mistake in the label represents a material misstatement in the financial statements, which will remain uncorrected unless testing reveals that the misstatement exists. Recall that under the risk-based testing approach, more testing is needed if the risk that a material misstatement exists is high, and less testing is needed if the risk is low. Therefore, as we

[3] Specifically, Hong (2022) observes the same automatic, unconscious effects of priming a promotion versus prevention focus (i.e., regulatory focus) among auditors *and* among novice participants (see footnote 3 in Hong 2022). Imposing a reward versus penalty frame is a way of priming a promotion versus prevention focus (Crowe and Higgins 1997), which is the theoretical basis for H1 and H2. A promotion focus represents the motivational orientation for pursuing hopes and aspirations, whereas a prevention focus represents the motivational orientation for pursuing duties and obligations (Higgins 1998). A reward frame causes decision-makers to view the incentivized goal as hopes and aspirations, whereas a penalty frame causes decision-makers to view the same goal as duties and obligations (Shah, Higgins, and Friedman 1998). The evidence from Hong (2022) suggests that the results of H1 and H2 should generalize to auditors.

describe later, participants were incentivized to test a bag if the label is misstated and not test a bag if the label is correctly stated.

Key to diagnostic decision-making, different causes can lead to the same phenomenon. By looking at the phenomenon, auditors do not know for sure whether a material misstatement is the underlying cause. Similarly in our task, a mislabeled bag and a correctly labeled bag can have the same appearance. However, attending to relevant information and engaging in effortful reasoning should help participants infer whether mislabeling is the underlying cause based on a bag's appearance. Specifically, participants were told that (i) each thermal ball in a *heating* bag has a 60 percent chance of being *red* and a 40 percent chance of being white, (ii) each thermal ball in a *cooling* bag has a 60 percent chance of being *white* and a 40 percent chance of being red, and (iii) the company produced thousands of bags with the same number of heating and cooling bags. Attending to this information and engaging in effortful reasoning should help participants infer that a bag is likely mislabeled and requires testing if the number of red balls in a bag exceeds the number of white balls in that bag (see examples of bags in Table 2). As we explain later, the bags come in two sizes.

Participants reviewed 100 bags drawn from two equally likely distributions (heating and cooling). Unknown to participants, 49 bags were heating, and 51 bags were cooling. All participants saw the same 100 bags in random order. For each bag, participants (i) assessed the likelihood that the bag was mislabeled, and (ii) decided whether they would test the bag. Testing would reveal the bag type and allow the company to relabel the bag if necessary. Not testing means the bag retains the "cooling" label. 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 risk judgments and testing actions on the appearance of each bag rather than the number of

bags already tested. We checked participants' comprehension of these key aspects of the task. Participants could not proceed to reviewing the bags until they answered all seven comprehension checks correctly. We did not reveal the "correct" risk judgment or the bag type immediately after participants reviewed each bag. Instead, we provided feedback after all bags were inspected. Future research can examine whether providing interim feedback affects diagnostic performance.

**Independent Variable**

Following the spirit of the auditing standards (AS 1101.10, 11; AS 1015.09), we incentivize participants to test *accurately* based on the risk of material misstatement and *objectively* with no bias between testing and not testing. We manipulate the *incentive frame* (reward versus penalty) between participants, keeping their payoffs constant for a given level of accuracy between frames. In the reward frame condition, participants were informed that they would make two dollars plus a two-dollar bonus if 2/3 or more of their testing decisions turned out to be correct. Correct decisions refer to testing a mislabeled bag (i.e., a hit) or not testing a correctly labeled bag (i.e., a correct rejection). In the penalty frame condition, participants were informed that they would make four dollars minus a two-dollar penalty if more than 1/3 of their testing decisions turned out to be incorrect. Incorrect decisions refer to testing a correctly labeled bag (i.e., a false alarm) or not testing a mislabeled bag (i.e., a miss). To promote unbiased testing, in both conditions, participants' payoffs are designed to be the same for incurring hits or correct rejections, and the same for incurring false alarms or misses.

Consistent with real-world audits, testing is costly. We embed the testing cost in participants' payoffs. In both conditions, testing correctly labeled bags is designed to reduce participants' payoffs. However, incurring the testing cost is worthwhile when bags are

mislabeled, consistent with real-world situations where the benefit of detecting a material misstatement that *exists* outweighs the testing cost incurred to detect that misstatement. Thus, testing mislabeled bags is designed to increase participants' payoffs. In our task, the company also desires accurate and unbiased testing, consistent with participants' incentives (see details at Appendix 1 for the entire instrument). Participants learned their performance and payoffs at the end of the study (mean = 3.22 dollars). The payoffs do not differ between conditions ($t_{194} = 1.40$, $p = 0.163$).[4]

**Dependent Variables**

Participants provided their risk judgments and decided their testing actions on the same screen for each bag (one bag per screen). The heating and cooling distributions from which we drew the bags provide normative benchmarks for analyzing judgments and actions. The dependent variables related to risk judgments are judgment bias and judgment inaccuracy. The dependent variables related to testing actions are the percentage of bags tested, the location of the testing criterion, and testing accuracy. We describe each variable below.

***Risk judgments***

*Risk judgments* are the assessed likelihood of mislabeling averaged across 100 bags. To evaluate risk judgments, we use the posterior likelihood of mislabeling as a normative benchmark. Accordingly, *judgment bias* is the assessed likelihood of mislabeling for a bag minus the posterior likelihood for that bag averaged across 100 bags. *Judgment inaccuracy* is the absolute deviation of the assessed likelihood for a bag from the posterior likelihood for that bag averaged across 100 bags.[5] The posterior likelihood is a function of the number of red balls

[4] All p-values are two-tailed except when otherwise noted for directional predictions.
[5] Our measures of bias and inaccuracy are consistent with proxies used in earnings forecasts research (e.g., Walther and Willis 2013; Duru and Reeb 2002), where forecast bias is the signed forecast error (i.e., forecasted earnings –

minus the number of white balls in a bag (i.e., #Red - #White), given the properties and prior probabilities of the heating and cooling distributions (see Table 2). For example, if #Red - #White is zero, the posterior likelihood of mislabeling is 50 percent regardless of the bag size (six or twelve balls per bag). The higher the value of #Red - #White, the higher the posterior likelihood of mislabeling. Note that in Table 2 different bag compositions can produce the same #Red - #White. For example, a six-ball bag with two red balls has the same posterior likelihood of mislabeling as a twelve-ball bag with five red balls. By varying the bag size, we can verify if participants strictly use #Red-#White in assessing risks.

***Testing Actions***

We calculate *the percentage of bags tested* based on participants' testing action for each bag. To evaluate participants' testing criterion and testing accuracy, we adopt measures from signal detection theory (Macmillan and Creelman 2004; Ramsay and Tubbs 2005). Specifically, we calculate the hit rate and false alarm rate of each participant. The hit rate is the percentage of *mislabeled* bags that are tested, and the false alarm rate is the percentage of *correctly* labeled bags that are tested. Recall that drawing the bags from the two equally likely distributions (heating and cooling) resulted in 49 mislabeled bags, and 51 correctly labeled bags. Therefore, the actual bag types, along with participants' testing action for each bag, allow us to calculate the hit rate and false alarm rate of each participant.

The *location* of *the subjective testing criterion* is an additive function of a participant's hit rate (H) and false alarm rate (F): $-0.5\,[Z(H) + Z(F)]$.[6] A zero value suggests that a

actual earnings), and forecast inaccuracy is the absolute value of the forecast error. In our task, the actual risk of material misstatement is the posterior likelihood that a bag is mislabeled.

[6] 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 are normally distributed 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).

participant is unbiased between testing and not testing, and that the participant chooses a neutral criterion for testing. In our task, the neutral criterion is located at which the number of red balls equals the number of white balls in a bag (i.e., #Red - #White = 0). Recall that the prior probability of mislabeling is 50 percent (i.e., the heating and cooling distributions are equally likely), and that participants' payoffs are the same for incurring hits or correct rejections and the same for incurring misses or false alarms.[7] Therefore, our task is designed to promote the choice of a neutral criterion. Choosing the neutral criterion means that participants will test a bag if the observed #Red - #White is greater than zero (i.e., exceeds the neutral criterion) and that participants will not test a bag if the observed #Red - #White is less than zero (i.e., falls short of the neutral criterion).

The criterion location takes a negative value when a participant lowers the subjective testing threshold from the neutral criterion. In this case, the cutoff point for testing is located at which #Red - #White is negative, suggesting an action bias towards testing. Testing all bags regardless of the risk of mislabeling suggests an extremely low testing threshold. In this case, the hit rate and false alarm rate are both 100 percent. Accordingly, the criterion location takes the value of negative infinity, suggesting an extremely low testing threshold. On the other hand, the criterion location takes a positive value when a participant increases the subjective testing threshold from the neutral criterion. In this case, the cutoff point for testing is located at which

[7] These parameters serve as a starting point for understanding whether and how incentive framing affects diagnostic performance. We recognize that the parameters in practice may differ from our design in idiosyncratic ways. For example, working on restatement audits may cause auditors to perceive the prior probability of a misstatement to be greater than 50%; working with supervisors who punish costly skepticism may create stronger incentives to ensure correct rejections (versus hits) and avoid false alarms (versus misses); auditing clients who face stricter regulatory oversight (e.g., brokers-dealers and banks) may create stronger incentives to ensure hits (versus correct rejections) and avoid misses (versus false alarms). In addition to promoting objective testing, another benefit of our parameters is that they are simple, which allow novice participants with general knowledge to accurately infer whether a bag is mislabeled based on its appearance with reasonable effort. Future research can alter the parameters to suit the research question of interest.

#Red - #White is positive, suggesting an action bias towards not testing. Not testing any bags regardless of the risk of mislabeling suggests an extremely high testing threshold. In this case, the hit rate and false alarm rate are both zero percent. Accordingly, the criterion location takes the value of positive infinity, suggesting an extremely high testing threshold.

*Testing accuracy* is a function of the difference between the hit rate and false alarm rate: $Z(H) - Z(F)$. A higher value indicates higher accuracy in differentiating the mislabeled bags from the correctly labeled bags. If a participant tests all mislabeled bags *and* does not test any correctly labeled bags, the hit rate will be 100 percent, and the false alarm rate will be zero percent. In this case, testing accuracy is positive infinity, indicating perfect accuracy. If a participant tests all bags regardless of the risk of mislabeling, both the hit and false alarm rate will be 100 percent. In this case, testing accuracy is zero, suggesting failure to differentiate the mislabeled bags from the correctly labeled bags. Similarly, if a participant does not test any bags, regardless of the risk of mislabeling, both the hit and false alarm rate will be zero percent. In this case, testing accuracy is again zero, suggesting failure to differentiate. We also use the percentage of testing decisions that turn out to be correct as an additional proxy for testing accuracy. Recall that correct decisions refer to testing a mislabeled bag or not testing a correctly labeled bag.

## IV. RESULTS

### Participants' Understanding of the Auditor Role

#### *Risk Judgments*

To assess whether participants understood their auditor role, we evaluate whether participants' risk judgments and testing actions correspond to the risk of material misstatement. Recall that the greater the number of red balls relative to white balls in a bag (#Red - #White),

the higher the posterior likelihood of mislabeling (see Table 2). 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 participants assessed a higher likelihood of mislabeling for low- versus high-risk bags. Overall, participants appeared able to understand the relationship between bag compositions and the likelihood of mislabeling.

***Testing Actions***

Recall that participants were incentivized to test accurately and objectively. Attending to the relevant information and engaging in effortful reasoning should help participants arrive at a perfect testing strategy that maximizes their payoffs. Under both frames, the perfect testing strategy is to test a bag when the posterior likelihood of mislabeling is greater than 50% (#Red > #White), to not test a bag when the posterior likelihood is lower than 50% (#Red < #White), and to be indifferent in testing when the posterior likelihood is equal to 50% (#Red = #White). Out of the 196 participants, 45 participants adopted the perfect testing strategy for all bags (rewards: n = 23; penalties: n = 22).

Relaxing the benchmark of a perfect testing strategy, we examine whether testing weakly increases as the risk of misstatement increases. Panel A of Table 2 summarizes the eight risk categories per bag composition (i.e., #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). Among the 196 participants, 165 participants tested an equal or higher percentage of bags as the risk of misstatement increases from one category to the next across all eight

categories (rewards: n = 82; penalties: n = 83). Therefore, most participants increased testing as the risk of misstatement increased.

### *Unqualified Participants*

The results above suggest that, overall, participants understood their role as auditors. However, we exclude observations from 20 participants whose risk judgments and testing actions are internally inconsistent. Specifically, these participants tended to test 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.[8] Thus, we retain 176 observations in the tests of hypotheses.

## Tests of Hypotheses

### *Test of H1*

H1 predicts that a reward versus penalty frame increases testing. Recall that we kept participants' payoffs economically equivalent between frames. If participants' testing actions only depend on payoffs, then framing should not affect testing. However, as predicted, participants in the reward frame condition test a higher percentage of bags than those in the penalty frame condition (means = 54.7% versus 50.1%, one-tailed p = 0.019, Table 3).[9] Figure 2

[8] 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 misstatement is low *and* not test when the assessed likelihood of misstatement is high. So, a participant who always (or never) tested would not be categorized as a flipper. Flippers provided internally inconsistent judgments and actions 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 inconsistent judgments and actions, 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 in their first attempt, and scored lower on the expanded cognitive reflection test (Toplak, West, and Stanovich 2014) that is indicative of analytical ability (Toplak, West, and Stanovich 2011; Welsh, Burns, and Delfabbro 2013; all p-values $\leq$ 0.002). These results suggest that flippers are potentially confused about and incapable of performing our task. Including all observations does not change any inferences except that it would slightly weaken the results reported for testing actions. Specifically, the p-values reported in Table 3 would be 0.051 for the percentage of bags tested, 0.054 for the criterion location, and 0.102 for the false alarm rate.

[9] 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.

shows that testing increases under a reward versus penalty frame for all eight risk categories (i.e., bag compositions in Table 2). To explore which risk categories drive the main result of increased testing, we estimate a panel regression in equation (3). Participants identify the panels and bags identify the trials. The dependent variable is *Choice*, measured at the bag level. For each bag, *Choice* equals 1 if that bag was tested and 0 if not. *Treatment* is the incentive frame (rewards = 1; penalties = 0). *Composition* is an indicator variable denoting the eight risk categories.

$$DV = \alpha_0 + \alpha_1 Treatment + \sum \alpha_2^k \, Composition + \sum \alpha_3^k \, Treatment \; X \; Composition \quad (3)$$

A reward frame significantly increases the likelihood of testing than a penalty frame in three risk categories. We denote the three categories with an asterisk in Figure 2 ($\alpha_1 + \alpha_3^k > 0$ for the ($k$) categories at a 5% level, untabulated). Two of the categories have a lower risk of misstatement, with the posterior likelihood of mislabeling being 8.07 percent for #Red - #White = -6 and 30.77% for #Red - #White = -2. The third category (#Red - #White = 4) has a higher risk of misstatement, with the posterior likelihood of mislabeling being 83.50 percent.[10]

***Test of H2***

H2 predicts that a reward versus penalty frame lowers the subjective testing criterion. Recall that a zero value in the criterion location indicates a neutral standpoint, meaning that participants are unbiased between testing and not testing. A negative value indicates a testing criterion that is lower than the neutral standpoint, suggesting an action bias towards testing. A positive value indicates an increased testing criterion from the neutral standpoint, suggesting a bias towards not testing. We find that the value of the criterion location is lower in the reward

[10] 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 = 0), participants in the reward frame condition tested 73 percent of the time, whereas participants in the penalty frame condition tested 60 percent of the time. While the direction is consistent with H1, the difference is not significant at conventional levels.

versus penalty frame condition (one-tailed p = 0.024, Table 3). Therefore, H2 is supported.[11,12] More specifically, the criterion location under a reward frame is on average -0.16 and significantly different from zero ($t_{85} = 2.64$, p = 0.010), suggesting a testing bias. In contrast, the criterion location under a penalty frame is on average -0.012 and not significantly different from zero ($t_{89} = 0.47$, p = 0.766), suggesting unbiased testing.

Recall that the criterion location is a function of the transformed hit rate and false alarm rate. We further explore how the incentive frame affects the hit rate and false alarm rate. The hit rate is the percentage of mislabeled bags that are tested, and the false alarm rate is the percentage of correctly labeled bags that are tested. In our study, the hit rate does not differ by incentive frames (p = 0.162, Table 3). However, 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), due to increased testing in the two lower-risk categories (#Red - #White = -6 and -2, Figure 2). Therefore, increasing testing blindly can increase audit cost with no improvement in misstatement detection.[13]

Interestingly, our results suggest that framing affects the subjective testing criterion automatically and unconsciously. In developing H2, we argue that a reward frame increases the weights on ensuring hits and avoiding misses compared to a penalty frame. Participants' responses to post-task questions suggest that they are *unaware* of the changes in how they weigh hits (versus correct rejections) and misses (versus false alarms) in response to the difference in

[11] 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 criterion location ($t_{174} = 2.09$, one-tailed p = 0.019, untabulated).
[12] Trait-level regulatory focus (Higgins et al. 2001) does not differ by conditions ($t_{174} = 0.93$, p = 0.352). Inferences about H1 and H2 do not change if we control for trait-level regulatory focus. Specifically, trait-level regulatory focus does not by itself or interact with the incentive frame in predicting testing or the criterion location (all p-values > 0.478). This result suggests that the observed framing effects should apply to those who self-select into accounting because of their trait-level regulatory focus.
[13] Inferences do not change if we use the number of hits and the number of false alarms each participant incurred as alternative measures, instead of the hit rate and false alarm rate per participant.

framing, even though their subjective testing criterion changes in response to the difference in framing. Specifically, participants rated their agreement with four statements on seven-point Likert scales in the post-experimental questionnaire: 1) "I was concerned about mistakenly testing a bag that is correctly labeled", 2) "I was concerned about mistakenly NOT testing a bag that is mislabeled", 3) "I wanted to make sure that all bags that are mislabeled get tested", and 4) "I wanted to make sure that all bags that are correctly labeled do NOT get tested". We calculate the relative weights on hits versus correct rejections by subtracting the rating for statement (4) from that for statement (3). Similarly, we calculate the relative weights on misses versus false alarms by subtracting the rating for statement (1) from that for statement (2). Neither of the self-reported relative weights differ by frames (p-values > 0.421).

***Tests of H3***

H3 predicts that a penalty frame increases the accuracy of risk judgments and testing actions relative to a reward frame. We fail to find evidence consistent with H3. Specifically, judgment inaccuracy, measured as the absolute error in the assessed likelihood of mislabeling averaged across 100 bags, does not differ by incentive frames (one-tailed p = 0.191, Table 3). Testing accuracy, measured as $Z(H) - Z(F)$, also does not differ by incentive frames (one-tailed p = 0.287, Table 3).[14] Inferences do not change when we use the number of correct testing decisions made as an alternative measure for testing accuracy (one-tailed p = 0.233, Table 3).

Recall that in developing H3, we argue that increased on-task attention (i.e., cognitive effort) increases the accuracy of risk judgments and testing actions. In our task, attending to case information and investing effort in the task suggest increased on-task attention. Therefore, we use participant's performance on the seven comprehension checks and their self-reported effort

[14] Inferences do not change if we use the difference of the untransformed hit and false alarm rates ($H - F$) as an alternative measure for testing accuracy ($t_{174} = 0.66$, one-tailed p = 0.253, untabulated).

invested in the task (0: not at all; 10: very much) as proxies for the unobservable on-task attention. Consistent with our reasoning, both proxies are significantly positively correlated with judgment accuracy and testing accuracy (all p-values < 0.002). However, neither proxy differs by incentive frames (p-values > 0.794). Additionally, neither proxy is significantly correlated with the percentage of bags tested or the testing criterion location (all p-values > 0.282), suggesting that the decision to test or not depends on more than attention and effort.

To summarize, we find that a reward versus penalty frame increases testing by lowering the subjective testing criterion. However, framing fails to induce a difference in effort (i.e., on-task attention), which is observed to increase diagnostic accuracy. Prior research finds that imposing penalties versus economically equivalent rewards increases effort (e.g., Hannan et al. 2005), which further increases productivity (e.g., Hossain and List 2012; Imas et al. 2017). We do not observe such an effect potentially because the cognitive effort required in diagnostic decision-making, such as attending to relevant information and engaging in logical reasoning, is less responsive to incentive framing compared to the effort examined in prior research, such as choosing effort levels (e.g., Hannan et al. 2005; Brink 2011; Christ, Sedatole, and Towry 2012; Gonzalez, Hoffman, and Moser 2020), completing sliders (e.g., Imas et al. 2017), translating symbols (e.g., Church et al. 2008), and producing electronic parts (e.g., Hossain and List 2012; Van der Stede, Wu, and Wu 2020).

***Examining Alternative Explanations***

In this section, we consider alternative explanations for our results. We find that participants in the penalty frame condition are marginally higher in the need for cognition ($t_{174} = 1.92$, $p = 0.056$) and they are more loss averse ($t_{174} = 2.83$, $p = 0.005$) than those in the reward frame condition. The need for cognition represents individuals' disposition to engage in and

enjoy effortful thinking (Cacioppo and Petty 1982). We measure the need for cognition using the scale developed by Cacioppo, Petty, and Feng Kao (1984). Loss aversion represents the extent to which the decrease in utility from losing one dollar exceeds the increase in utility from gaining one dollar (Kahneman and Tversky 1979). We measure loss aversion using participants' choices in six hypothetical gambles adapted from prior research (Kahneman 1992; Tversky and Kahneman 1991; Harinck, Van Dijk, Van Beest, and Mersmann 2007). We next examine whether the differences in participants' need for cognition and loss aversion drive the effects of incentive framing.

We find that the differences in need for cognition do not explain our test results for H1 and H2. Specifically, the need for cognition score is not significantly correlated with the percentage of bags tested or the criterion location (Spearman, both p-values > 0.635, untabulated). The need for cognition score is also not significantly correlated with the hit rate (Spearman, p = 0.187, untabulated). However, a higher need for cognition is marginally associated with a lower false alarm rate (Spearman, r = -0.138, p = 0.068), raising the concern that the false alarm rate is higher under a reward than penalty frame because participants in the reward frame condition are lower in need for cognition. Including need for cognition as a covariate does not eliminate the effect of the incentive frame; the false alarm rate remains marginally higher under the reward versus the penalty frame ($t_{173} = 1.87$, p = 0.064).

We also find that the differences in need for cognition do not drive our test results for H3. Higher diagnostic accuracy requires attention and effortful thinking. Consistent with our proxies for on-task attention (i.e., performance on comprehension checks and self-reported task effort), a higher need for cognition is associated with higher testing accuracy, a higher number of correct testing decisions made, and higher judgment accuracy (all p-values < 0.013, untabulated). Recall

that participants in the penalty frame condition score marginally higher on need for cognition compared to those in the reward frame condition. Therefore, testing accuracy, the number of correct decisions made, and judgment accuracy should be higher under a penalty frame than a reward frame, implicitly supporting H3. Yet we still do not observe support for H3. Therefore, H3 remains unsupported.

Finally, the differences in loss aversion cannot explain any of our results for two reasons. First, loss aversion does not correlate with any of our dependent variables reported in Table 3 (all p-values > 0.218). Second, prior research frequently cites loss aversion as the reason why imposing penalties versus economically equivalent rewards increases effort and effort-based productivity (Hannan et al. 2005; Brink 2011; Hossain and List 2012; Imas et al. 2017), assuming that agents have the same level of loss aversion between frames. In our study, the level of loss aversion is higher under the penalty versus reward frame. If the cognitive effort required in our diagnostic task is akin to the effort examined in the aforementioned research, then the difference in loss aversion should motivate more cognitive effort in the penalty frame condition in addition to the difference in framing, biasing in favor of finding support for H3. Yet, we fail to find support that participants' performance accuracy (H3), self-reported motivation to maximize payoffs (0: not at all; 10: very much), and on-task attention (proxied as above) are higher under the penalty versus reward frame (all one-tailed p-values > 0.190).

**Test of Research Question**

Auditors are assumed to test a phenomenon if their assessed risk exceeds the subjective testing criterion. A reward versus penalty frame can increase testing (H1) by lowering the subjective testing criterion (H2), increasing the assessed risks (RQ), or both. We take two steps to better understand the mechanism through which framing affects testing. First, we examine

whether the incentive frame affects the assessed risk of material misstatement (RQ). Second, we examine the effect of framing on testing while controlling for the assessed risk. Any remaining effect of framing should be attributed to changes in the testing criterion. Although the test of H2 suggests a lowered testing criterion under a reward versus penalty frame, the criterion location measure assumes that the assessed risk is accurate under both frames, which may not be true in our experiment.

Regarding our RQ, the assessed likelihood is marginally higher under a reward versus penalty frame (means = 53.4% versus 51.7%, p = 0.064, Table 3), suggesting that a reward frame induces a judgment bias towards assuming that a material misstatement exists (p = 0.064, Table 3). More specifically, the assessed likelihood of mislabeling significantly exceeds the posterior likelihood of mislabeling in the reward frame condition (i.e., judgment bias > 0, p = 0.012, untabulated), and the assessed likelihood does not significantly differ from the posterior likelihood in the penalty frame condition (i.e., judgment bias = 0, p = 0.607, untabulated).

To explore which risk category drives the overall effect of framing on risk judgments, we estimate equation (3) using a panel regression. Participants identify the panels and bags identify the trials. The dependent variable is $\log\left( {}^{y}/_{1-y} \right)$ 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. We find that the transformed assessed likelihood is significantly higher under a reward frame than under a penalty frame in two risk categories. We denote the two categories (#Red - #White = -2 and 0) with an

asterisk in Panel A of Figure 3 ($\alpha_1 + \alpha_3^k > 0$, for the ($k$) categories at a 1% level, untabulated).[15] Also in these two categories, we find that participants' judgment bias (i.e., assessed – posterior likelihood) is more positive under a reward versus penalty frame using equation (3) (see Panel B of Figure 3).

Next, we test the effect of incentive frame on testing actions controlling for participants' risk judgments. We estimate equation (4) using logit regression and the same panel that we used in equation (3). The dependent variable is *Choice* (1 = test; 0 = not test) at the bag level. The independent variables are *treatment* (1 = rewards; 0 = penalties) and *belief* (participants' assessed risk categories). Specifically, *belief* has seven categories based on participants' assessed likelihood of mislabeling: 0 – 9%, 10-24%, 25-48%, 49-51%, 52-75%, 76-90%, and 91-100%.

$$Choice = \alpha_0 + \alpha_1 Treatment + \sum \alpha_2^k Belief + \sum \alpha_3^k Treatment\ X\ Belief \quad (4)$$

Figure 4 illustrates the percentage of bags tested within each subjective risk category (belief), constructed based on participants' assessed likelihood of mislabeling. After conditioning choice to test by assessed likelihood, the effect of the incentive frame remains significant in two subjective risk categories ($p < 0.05$, untabulated). We denote these subjective categories (52-75% and 76-90%) with an asterisk. Therefore, controlling for conscious risk judgments does not eliminate the effect of incentive frame on testing actions, suggesting that increased risk judgments and reduced testing criteria *jointly* contribute to increased testing under a reward frame.

**Potential Heuristics in Risk Judgments**

[15] This result is robust to using panel Tobit regression and fitting the same equation with the assessed likelihood as the dependent variable.

As illustrated in Panel B of Figure 3, participants overestimate the likelihood of mislabeling when the risk of misstatement is low (posterior likelihood < 50%), and they underestimate the likelihood of mislabeling when the risk of misstatement is high (posterior likelihood > 50%; all p-values < 0.01). This pattern appears to be the result of participants using heuristics in risk judgments. Specifically, participants appear to use the percentage of red balls in a bag (i.e., fraction red, see Table 2) to estimate the likelihood of mislabeling (Figure 4, Panel B). Regressing the assessed likelihood against predictor *fraction red* shows a better fit than predictor *posterior likelihood*, as evidenced in lower Akaike's information criterion and Bayesian information criterion (untabulated).[16] Using fraction red to estimate the likelihood of mislabeling should reduce any differences in risk judgments between incentive frames. Yet, we still observe a marginal difference in judgments by frames in the test of RQ.

Interestingly, Panel B of Figure 4 suggests that, as bag size increases, using fraction red to estimate the likelihood of mislabeling appears to amplify the judgment bias pattern observed in Panel B of Figure 3. In other words, as bag size increases from six balls to twelve balls per bag, participants overestimate the risk of misstatement to a greater exent when the risk is low (i.e., #Red - #White < 0), and they underestimate the risk of misstatement to a greater extent when the risk is high (i.e., #Red - #White > 0). Table 2 explains why: the posterior likelihood of mislabeling is a function of #Red - #White regardless of bag size, whereas fraction red is a function of bag size rather than #Red - #White. When the bag size is small (i.e., six balls per bag), fraction red approximates the posterior likelihood of mislabeling in value. However, when the bag size is large (i.e., twelve balls per bag), fraction red further deviates from the posterior

[16] Akaike and Bayesian Information Criterion are two ways of scoring competing models based on their log-likelihood and complexity. In other words, each criterion deals with both the risk of overfitting and the risk of underfitting, albeit with different weights.

likelihood, exceeding the posterior when the risk is low (#Red - #White < 0) and falling short of the posterior when the risk is high (#Red - #White > 0).

## V. CONCLUSION AND DISCUSSION

Diagnostic performance is integral to many accountants' jobs, such as identifying the cause of budget overruns, the cause of declining sales, the cause of changes in working capital, and the cause of employee turnover among the possible causes. We examine diagnostic performance in the audit setting. Auditors must provide reasonable assurance about whether a material misstatement is the underlying cause of an observed phenomenon. Given a phenomenon, auditors must assess the risk of misstatement and decide whether they will perform additional testing to help identify the cause. Auditing standards value accuracy and objectivity in auditors' judgments and actions. When auditors are provided with incentives to test accurately and objectively, we examine whether and how framing equivalent incentives as rewards versus penalties affects auditors' risk judgments and testing actions. Economic theory predicts that framing should not affect performance. However, in an experiment, we find that a reward versus penalty frame increases testing by lowering participants' subjective testing criterion *and* by increasing their assessed risk. Framing affects the testing criterion automatically and unconsciously. Controlling for the assessed risk does not eliminate the effect of framing on testing.

In our experiment, increased testing under a reward versus penalty frame did not improve the identification of misstatement but resulted in more false alarms when financial statements were fairly presented. For accounting firms that might be concerned with controlling audit costs, our result highlights the importance of improving the accuracy of auditors' risk judgments and testing actions. On the other hand, increased testing, even if it causes more false alarms with no

improvement in misstatement detection, can be beneficial in several ways. First, at the early stage of auditor tenure, increased testing regardless of the misstatement risk can help deter clients from manipulating financial statements in the future. Second, in repeated audits, clients learn auditors' testing approach overtime. Increased testing in low-risk areas helps prevent clients from hiding misstatement in low-risk areas that auditors normally would not test. Third, increased testing helps auditors gather more evidence and thus develop more precise beliefs about their clients, even when increased testing results in false alarms.

Our study provides a *baseline* understanding of whether and how the incentive frame affects diagnostic performance. Participants were incentivized to test accurately without bias, as desired by the auditing standards. In practice, an auditor may have already developed a strong bias towards (or away from) testing, for example, from interacting with a supervisor who consistently rewards (or punishes) testing that results in false alarms (Brazel et al. 2016, 2021). It is then unclear whether a reward versus penalty frame can further increase testing. Our results suggest that conscious risk judgment does not solely rationalize increased testing, and that framing affects the subjective testing criterion automatically and unconsciously. Auditors are susceptible to automatic, unconscious effects, which even targeted interventions fail to eliminate (e.g., Brazel et al. 2019). Therefore, our results should generalize to situations where auditors have already developed a bias towards or away from testing. Consistent with our reasoning, the effect of framing on testing did not disappear when participants assessed a greater than 50 percent likelihood of misstatement, for which the obvious action is to test (Figure 4 Panel A). Future research can test the generalizability of our findings.

Finally, auditors are responsible for providing "reasonable assurance about whether the financial statements are free of material misstatement" (AS1001). Regulators, investors, and the

media have generally held auditors accountable for financial statement outcomes (misstated or not), particularly when client bankruptcies or restatements occur after auditors issue an unqualified opinion, even though the opinion can be justified by auditors' judgment processes (Peecher et al. 2013). In our experiment, participants are evaluated and paid based on financial statement outcomes rather than their judgment processes. However, applying the "correct" judgment process (i.e., adopting the perfect testing strategy), also increases the *likelihood* of being accurate about the outcomes (i.e., testing accuracy). We acknowledge that auditors who apply the "correct" judgment process can still issue an "incorrect" opinion, while at the same time auditors who apply the "incorrect" judgment process can also issue a "correct" opinion, about financial statement outcomes. Given this discrepancy, future research can examine whether designing auditor incentives based on financial statement outcomes versus judgment processes is more effective for improving audit quality.

## REFERENCES


Barr-Pulliam, D., J. F. Brazel, J. McCallen, and K. Walker. 2020. *Data Analytics and Skeptical Actions: The Countervailing Effects of False Positives and Consistent Rewards for Skepticism* (October 19, 2020). Available at https://ssrn.com/abstract=3537180

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.

Bonner, S. E. 1990. Experience effects in auditing: The role of task-specific knowledge. *The Accounting Review* 65 (1): 72-92.

Brazel, J. F., C. Gimbar, E. M. Maksymov, and T. J. Schaefer. 2019. The outcome effect and professional skepticism: A replication and a failed attempt at mitigation. *Behavioral Research in Accounting* 31(2): 135-143.

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. R. Schaefer. 2021. Do Rewards Encourage Professional Skepticism? It Depends. *The Accounting Review* (forthcoming).

Brink, AG. 2011. The Effect of Contract Frame on the Perceived Fairness and Planned Effort under Economically Equivalent Bonus, Penalty, and Combination Contracts. *Journal of Theoretical Accounting Research* (1): 145-153

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

Broadbent, D. E., and M. Gregory. 1963. Division of attention and the decision theory of signal detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences* 158 (971): 222-231.

Brown, C. 1987. Diagnostic inference in performance evaluation: Effects of cause and event covariation and similarity. *Contemporary Accounting Research* 4 (1): 111-126.

Brown, C. 1981. Human information processing for decisions to investigate cost variances. *Journal of Accounting Research*: 62-85.

Bolton, G., A. Ockenfels, and U. Thonemann. 2012. Managers and Students as Newsvendors. *Management Science* 58 (12): 2225-2233.

Bowen, H. J., M. L. Marchesi, and E. A. Kensinger. 2020. Reward motivation influences response bias on a recognition memory task. *Cognition* 202: 104337.

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.

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.

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.

DeFond, M. and J. Zhang. 2014. A review of archival auditing research. *Journal of Accounting and Economics* (58) 2–3: 275-326.

Duru, A., and D. M. Reeb. 2002. International diversification and analysts' forecast accuracy and bias. *The Accounting Review* 77 (2): 415-433.

Elliott, W. B., F. D. Hodge, J. J. Kennedy, and M. Pronk. 2007. Are MBA students a good proxy for nonprofessional investors? *The Accounting Review* 82 (1): 139-168.

Garrett, H. 1968. The tragedy of the commons. *Science* 162(3859): 1243-1248.

Gonzalez, G. C., V. B. Hoffman, and D. V. Moser. 2020. Do Effort Differences between Bonus and Penalty Contracts Persist in Labor Markets? *The Accounting Review* 95 (3): 205-222.

Hammersley, J. S. 2006. Pattern identification and industry-specialist auditors. *The Accounting Review* 81 (2): 309-336.

Hannan, R. L., V. B. Hoffman, and D. V. Moser. 2005. Bonus versus penalty: does contract frame affect employee effort? In *Experimental Business Research: Economic and Managerial Perspectives*: pp. 151-169. Holland: Springer.

Harinck, F., E. Van Dijk, I. Van Beest, and P. Mersmann. 2007. When gains loom larger than losses: Reversed loss aversion for small amounts of money. *Psychological Science* 18 (12): 1099-1105.

Higgins, E. T. 1998. Promotion and prevention: Regulatory focus as a motivational principle. In *Advances in Experimental Social Psychology* 30: 1–46. Elsevier.

Higgins, E. T., R. S. Friedman, R. E. Harlow, L. C. Idson, O. N. Ayduk, and A. Taylor. 2001. Achievement orientations from subjective histories of success: Promotion pride versus prevention pride. *European Journal of Social Psychology* 31 (1): 3-23.

Hong, Y. B. 2022. Initial task engagement: Unlocking the value of fit and non-fit to improve audit judgments. *The Accounting Review* (forthcoming).

Hossain, T., and J. A. List. 2012. The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*: 58 (12), 2151-2167.

Imas. A., S. Sadoff, and A. Samek. 2017. Do People Anticipate Loss Aversion? *Management Science* 63 (5):1271-1284.

Kahneman, D. 1973. *Attention and Effort* (Vol. 1063, pp. 218-226). Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D. 1992. Reference points, anchors, norms, and mixed feelings. *Organizational Behavior and Human Decision Processes* 51 (2): 296-312.

Kahneman, D., and A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* 47 (2): 263–292.

Knechel, R.W., S. Salterio, and B. Ballou. 2007. *Auditing: Assurance and Risk*. South-Western College Pub.

Levine, J. M., E. T. Higgins, and H. S. Choi. 2000. Development of strategic norms in groups. *Organizational Behavior and Human Decision Processes* 82 (1): 88-101.

Libby, R., R. Bloomfield, and M. W. Nelson. 2002. Experimental research in financial accounting. *Accounting, Organizations and Society* 27 (8): 775-810.

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*. Taylor & Francis Group.

Mturkdata. 2018. *The BOT problem on Mturk*. Available at: http://turkrequesters.blogspot.com/2018/08/the-bot-problem-on-mturk.html Accessed June 2020.

Nelson, M. W. 2009. A model and literature review of professional skepticism in auditing. *Auditing: A Journal of Practice & Theory* 28 (2): 1-34.

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. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46(4): 1023-1031.

Public Company Accounting Oversight Board (PCAOB). 2019. *Staff Preview of 2018 Inspection Observations*. Available at https://pcaobus.org/Inspections/Documents/Staff-Preview-2018-Inspection-Observations.pdf

Public Company Accounting Oversight Board (PCAOB). 2020. *Staff Update and Preview of 2019 Inspection Observations*. Available at https://pcaobus.org/Inspections/Documents/Staff-Preview-2019-Inspection-Observations-Spotlight.pdf

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.

Securities and Exchange Commission (SEC). 2002. *Acceleration of Periodic Report Filing Dates and Disclosure Concerning Website Access to Reports*. *Release Nos. 33-8128; 34-46464.* Washington, DC: SEC.

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.

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*: 475-502.

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* 20 (2): 147-168.

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. Wu (2020) An Empirical Analysis of Employee Responses to Bonuses and Penalties. *The Accounting Review* 95 (6): 395–4.

Walther, B. R., and R. H. Willis. 2013. Do investor expectations affect sell-side analysts' forecast bias and forecast accuracy? *Review of Accounting Studies* 18 (1): 207-227.

Watts, R., and Zimmerman, J. 1981. *Auditors and the determination of accounting standards*. Working Paper. University of Rochester.

Welsh, M., N. Burns, and P. Delfabbro. 2013. The cognitive reflection test: How much more than numerical ability? *In Proceedings of the Annual Meeting of the Cognitive Science society* Vol. 35, No. 35.

Yechiam, E., and G. Hochman. 2013. Losses as modulators of attention: review and analysis of the unique effects of losses over gains. *Psychological Bulletin* 139 (2): 497.

Yechiam, E., and G. Hochman. 2014. Loss attention in a dual-task setting. *Psychological Science* 25 (2): 494-502.

## Appendix 1: Research Instrument

https://drive.google.com/file/d/1Jl-6Dwh3fiCr99iiS8SB8jwWTu0t7Qk_/view?usp=sharing

## Appendix 2: Expected Utility Representation of H1 and H2

In binary diagnostic tasks, an auditor's decision to test a phenomenon or not results in four possible outcomes shown in Table 1. Each outcome results in different utility for the auditor. We denote the utility functions for the four outcomes as $U_1$ for a false alarm, $U_2$ for a hit, $U_3$ for a correct rejection, and $U_4$ for a miss. We assume that the worst favorable outcome results in greater utility than the best unfavorable outcome (i.e., $min\{U_2, U_3\} > max\{U_4, U_1\}$), that testing increases the certainty as to whether a misstatement is present, and that testing consumes resources. Therefore, if a misstatement is present, the auditor prefers to test, and if a misstatement is absent, the auditor prefers to avoid costly testing. Our assumptions are consistent with the auditing standards that require a risk-based testing approach, which values diagnostic accuracy. We do not define any specific form of the auditor's utility functions or any risk preferences.

A testing strategy considers the posterior likelihood of misstatement after observing the phenomenon $X$ and the outcome utilities when deciding whether to test or not. If the auditor decides to test, then the outcome will be a false alarm or hit, and if the auditor decides not to test, the outcome will be a correct rejection or miss. If the auditor tests, then the expected utility is:

$$\underbrace{P_H \frac{\Pr(X|Yes)}{P_H \Pr(X|Yes) + (1-P_H)\Pr(X|No)}}_{posterior\ likelihood\ of\ misstatement} U_2 + \underbrace{(1-P_H)\frac{\Pr(X|No)}{P_H \Pr(X|Yes) + (1-P_H)\Pr(X|No)}}_{posterior\ likelihood\ of\ no\ mistatement} U_1 \quad (1)$$

$P_H$ is the prior probability of misstatement. $Pr(X|state)$ is the probability of observing the phenomenon $X$, given the state that a misstatement is present (Yes) or absent (No). If the auditor does not test, then the expected utility is:

$$\underbrace{P_H \frac{\Pr(X|Yes)}{P_H \Pr(X|Yes) + (1-P_H)\Pr(X|No)}}_{posterior\ likelihood\ of\ misstatement} U_4 + \underbrace{(1-P_H)\frac{\Pr(X|No)}{P_H \Pr(X|Yes) + (1-P_H)\Pr(X|No)}}_{posterior\ likelihood\ of\ no\ mistatement} U_3 \quad (2)$$

A criterion $X_k$ is the point where the auditor is indifferent between testing or not. To find the criterion, we set equation (1) equal to equation (2), which characterizes the criterion value in terms of prior probabilities and outcome utilities:

$$P_H\ Pr(X_k \mid Yes)\ (U_2 - U_4) = (1-P_H)\ Pr(X_k \mid No)\ (U_3 - U_1)$$
$$\Longrightarrow \frac{Pr(X_k \mid Yes)}{Pr(X_k \mid No)} = \frac{(1-P_H)}{P_H}\frac{(U_3 - U_1)}{(U_2 - U_4)} \quad (3)$$

Solving equation (3) yields the criterion $X_k$ (i.e., threshold). To satisfy the equation, the higher the right-hand side value of equation (3), the higher a $X_k$ is needed. For example, *ceteris paribus*, the criterion increases when the prior probability of misstatement ($P_H$) decreases (i.e., the right-hand side value increases). That is, the fewer misstatements that an auditor observes in past audits, the more likely the auditor will increase the subjective testing criterion and be biased towards not testing. As another example, *ceteris paribus,* the criterion decreases when the utility of a hit ($U_2$) increases (i.e., the right-hand side value decreases). That is, the more an auditor cares about ensuring hits, the more likely the auditor will lower the subjective testing criterion and be biased towards testing.

A testing strategy depends on the relationship of the subjective testing criterion $X_k$ to the posterior likelihood of misstatement $p$. For any observed phenomenon, the updated posterior likelihood of misstatement is evaluated against the criterion:

$$p \gtreqless X_k \quad (4)$$

The auditor tests when the left-hand side of equation (4) is greater than the right-hand side, does not test when the left-hand side is less than the right-hand side, and is indifferent when the two sides are equal. The right-hand side $X_k$ serves as the testing threshold. When $U_2 = U_3$, $U_1 = U_4$,

and $P_H = (1 - P_H)$, then the criterion is 50%. In this case, auditors should test a phenomenon if the assessed likelihood of material misstatement is greater than the threshold (i.e., $p > 50\%$), and auditors should not test a phenomenon if the assessed likelihood of misstatement is less than the threshold (i.e., $p < 50\%$).[17]

We predict that a reward versus penalty frame increases testing (H1) by lowering the subjective testing criterion (H2). Recall that a reward versus penalty frame increases participants' tendency to ensure hits and avoid misses in memory recognition (Crowe and Higgins 1997; Levine et al. 2000; Bowen et al. 2020). Assuming this finding generalizes to binary diagnostic tasks, we capture the increased focus on hits and misses in equation (5). Compared to equation (3), equation (5) has an additional term $\alpha > 1$, which represents the increased focus on pursuing hits $U_2$ and avoiding misses $U_4$. As a result, $\alpha$ lowers the right-hand side value of equation (5) compared to equation (3), thereby lowering the testing criterion $X_k$ needed to satisfy equation (5) under a reward versus penalty frame (H2). If an auditor was indifferent to testing using a criterion when framing is absent, then adding the term ($\alpha > 1$) decreases the criterion, so now the auditor tests as we predict in H1. Also note our predictions are true regardless of auditors' risk-preferences (e.g., degree of loss aversion).

$$\frac{Pr(X_k \mid Yes)}{Pr(X_k \mid No)} = \frac{1}{\alpha}\frac{(1 - P_H)}{P_H}\frac{(U_3 - U_1)}{(U_2 - U_4)} \tag{5}$$

[17] A decision based on the equation can result in ex-post undesired outcomes (i.e., a miss or false alarm), just as a decision that ignores the equation can result in ex-post desired outcomes (i.e., a hit or correct rejection). This is consistent with the reality that even a well-planned audit can fail to reveal a misstatement.

# Tables and Figures

## Table 1: Diagnostic Audit Task

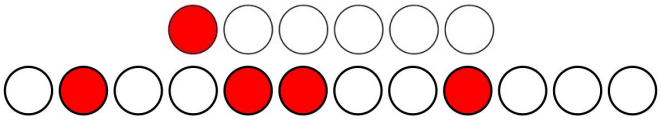
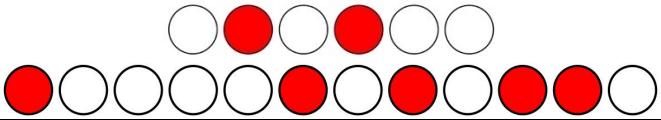
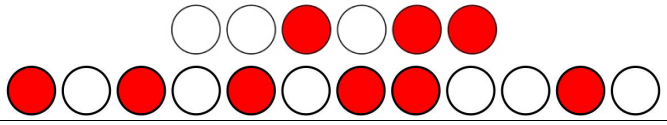
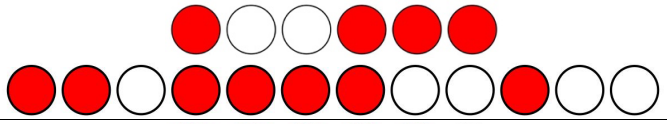
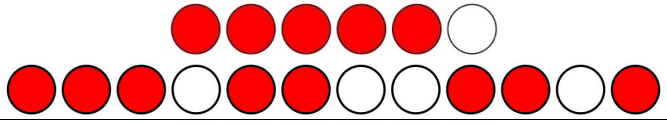
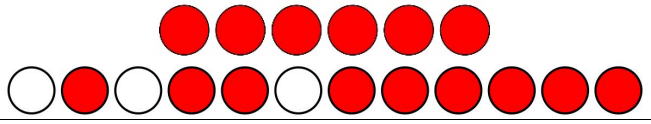
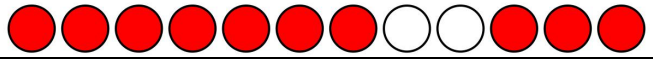
| | Misstatement Absent | Misstatement Present |
|---|---|---|
| Test | False Alarm | Hit |
| Do not test | Correct rejection | Miss |

**Table 2: Bags**

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

| The Risk of Misstatement | 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%<br>33.3% |  |
| -2 | 30.8% | 33.3%<br>41.7% |  |
| = | 50.0% | 50.0%<br>50.0% |  |
| +2 | 69.2% | 66.7%<br>58.3% |  |
| +4 | 83.5% | 83.3%<br>66.7% |  |
| +6 | 91.9% | 100.0%<br>75.0% |  |
| +8 | 96.2% | 83.3% |  |

Participants saw the same set of 100 bags presented in random order. Bag composition is the difference between red and white balls. Fraction red is the number of red balls divided by the bag size.
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 probability of a bag of $N$ balls having $X$ red balls, where $X$ is 0 up to $N$, is given by the probability mass function of the binomial distribution. The probability that a heating bag has $X$ red balls and $N$-$X$ white balls is $\Pr(X|Heating) = .4^{N-X}.6^{X}\binom{N}{X}$, and the probability that a cooling bag has $X$ red

balls and *N*-*X* white balls is $\Pr(X|Cooling) = .6^{N-X}.4^{X}\binom{N}{X}$. Using Bayes Rule, given *X* red balls observed, the posterior likelihood that the bag is a heating bag (mislabeled) is

$$\frac{P_H \Pr(X|Heating)}{P_H \Pr(X|Heating) + (1 - P_H) \Pr(X|Cooling)}$$

$P_H$ is the prior probability 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**

| **Descriptive statistics: mean (standard deviation)** | **Testing Actions** | | | | | | **Risk Judgments** | | |
|---|---|---|---|---|---|---|---|---|---|
| Treatment | % Tested | Criterion Location | Hit Rate | False Alarm Rate | Testing Accuracy | % Correct | Judgment Inaccuracy | 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%) |
| **Prediction: Penalties - Rewards** | | | | | | | | | |
| | H1: < 0 | H2: > 0 | No prediction | No prediction | H3: > 0 | H3: > 0 | H3: < 0 | RQ | RQ |
| t (df = 174) | -2.10 | 1.99 | -141 | -2.12 | 0.56 | 0.73 | -0.88 | -1.86 | -1.86 |
| p-value | .019† | .024† | .162 | .032 | 0.287† | .233† | 0.191† | 0.064 | 0.064 |

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

Dependent variables:

*% Tested* represents the percentage of bags tested averaged across 100 bags.

*Criterion location* = $-0.5[Z(H) + Z(F)]$ where H represents the hit rate and F represents the false alarm rate. A zero value indicates a neutral testing criterion with no bias between testing and not testing. A negative value indicates a reduced testing threshold from the neutral criterion, suggesting a bias towards testing. A positive value indicates an increased testing threshold from the neutral criterion, suggesting a bias towards not testing. Hit rate is the percentage of mislabeled bags that are tested. False alarm rate is the percentage of correctly labeled bags that are tested.

*Testing accuracy* = $Z(H) - Z(F)$. A higher value indicates higher accuracy in testing across 100 bags.

*% Correct* is the percentage of testing decisions correctly made (i.e., hits and correct rejections) based on the actual bag type.

*Judgment Inaccuracy* = |assessed likelihood – posterior likelihood of mislabeling|, averaged across 100 bags. A higher value represents lower accuracy. See posterior likelihood of mislabeling at Table 2.

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

*Judgment bias* = assessed likelihood – posterior likelihood of mislabeling, averaged across 100 bags. See posterior likelihood of mislabeling at Table 2.

Independent variables:

We manipulate the *incentive frame* between participants. In the *penalty frame* condition, participants were informed that they would make $4 for inspecting 100 bags. Additionally, they could pay a $2 penalty if more than 1/3 of their inspection decisions turned out to be incorrect. Participants would not pay the penalty if 1/3 or fewer of their decisions turned out to be incorrect. In the *reward frame* condition, participants were informed that they would make $2 for inspecting 100 bags. Additionally, they could earn a $2 bonus if 2/3 or more of their inspection decisions turned out to be correct. Participants would not earn the bonus if fewer than 2/3 of their decisions turned out to be 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 diagnostic outcomes.

**Figure 1: Summary of Experimental Procedures**

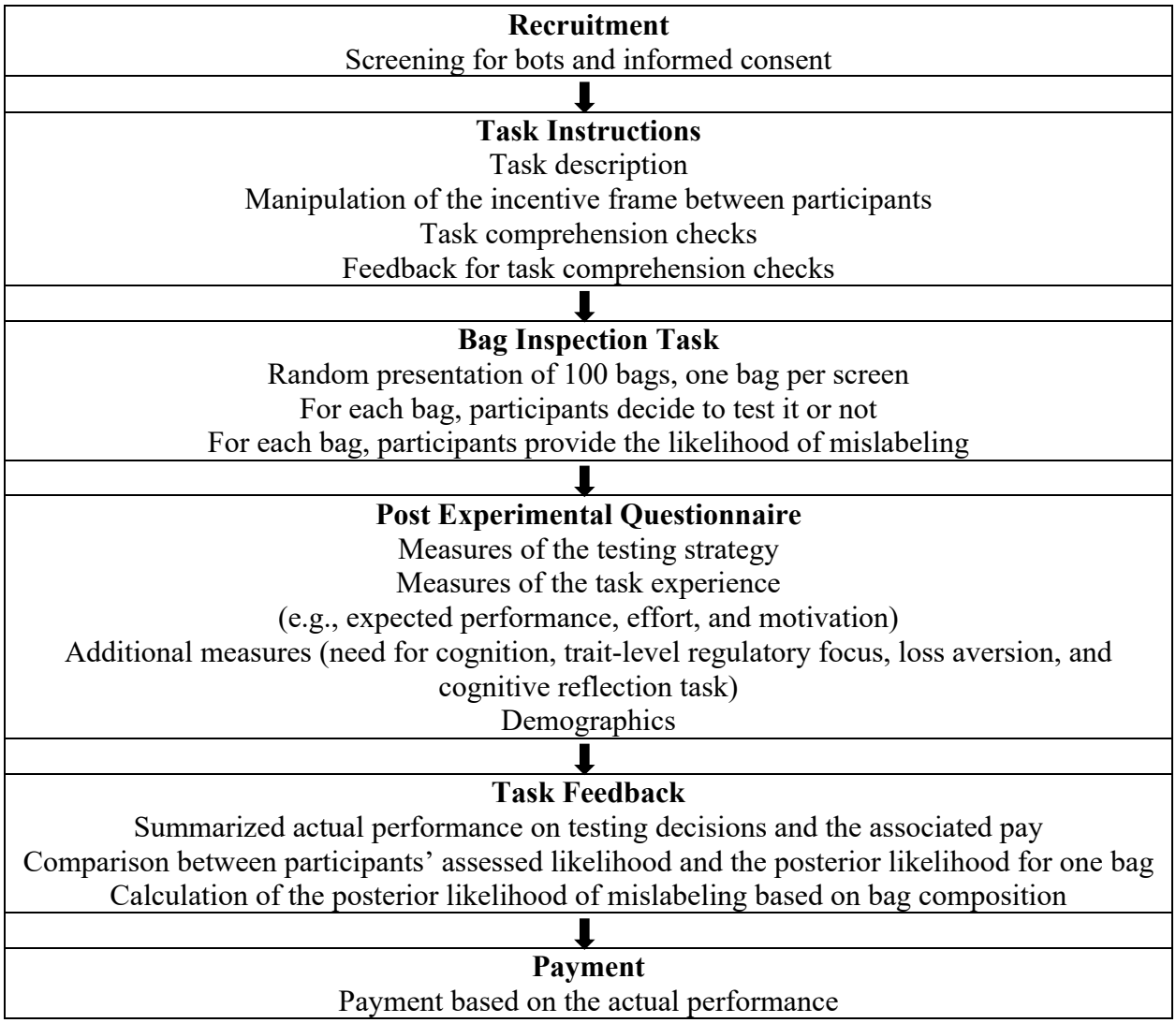
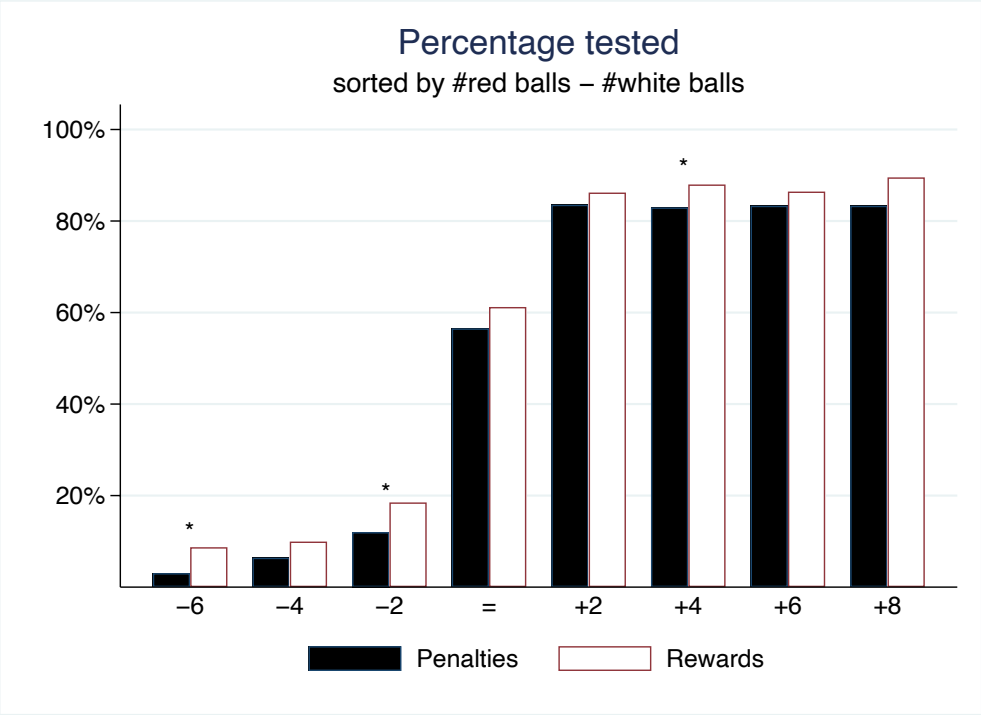
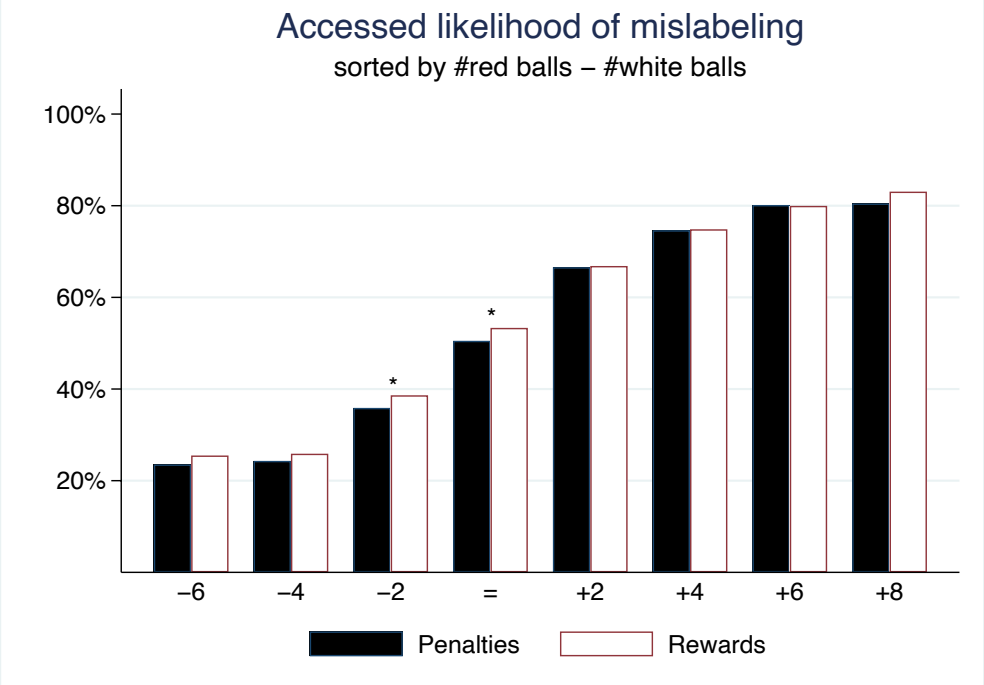


**Figure 2: The Percentage of Bags Tested by Bag Composition (#Red - #White)**

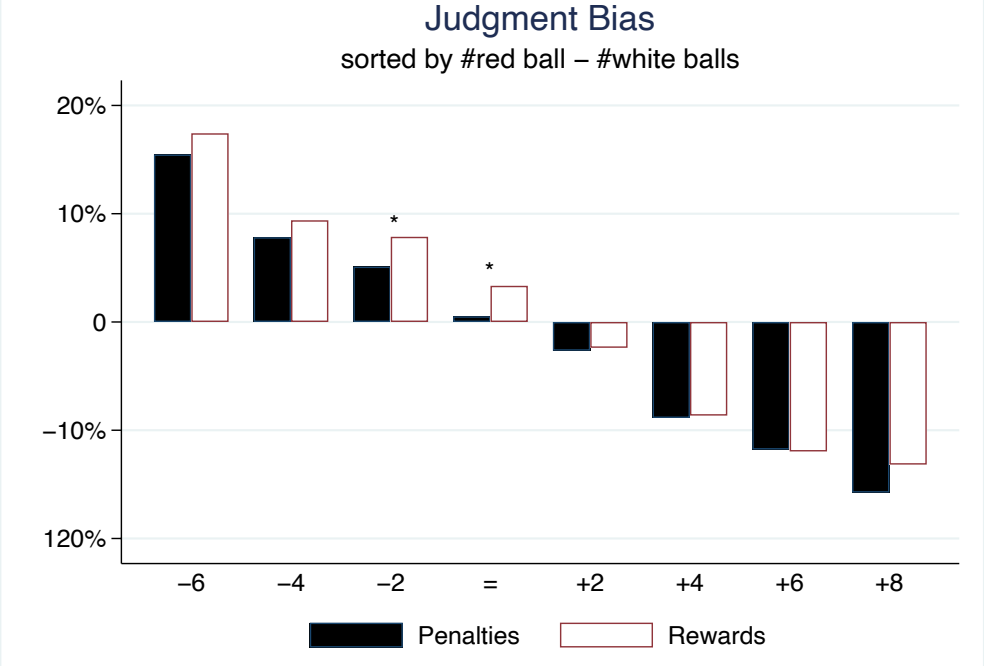



Note: * Different by incentive frame at a 5% level. See variable definitions at Table 2 and Table 3.

**Figure 3: Risk Judgment and Judgment Bias by Bag Composition (#Red - #White)**



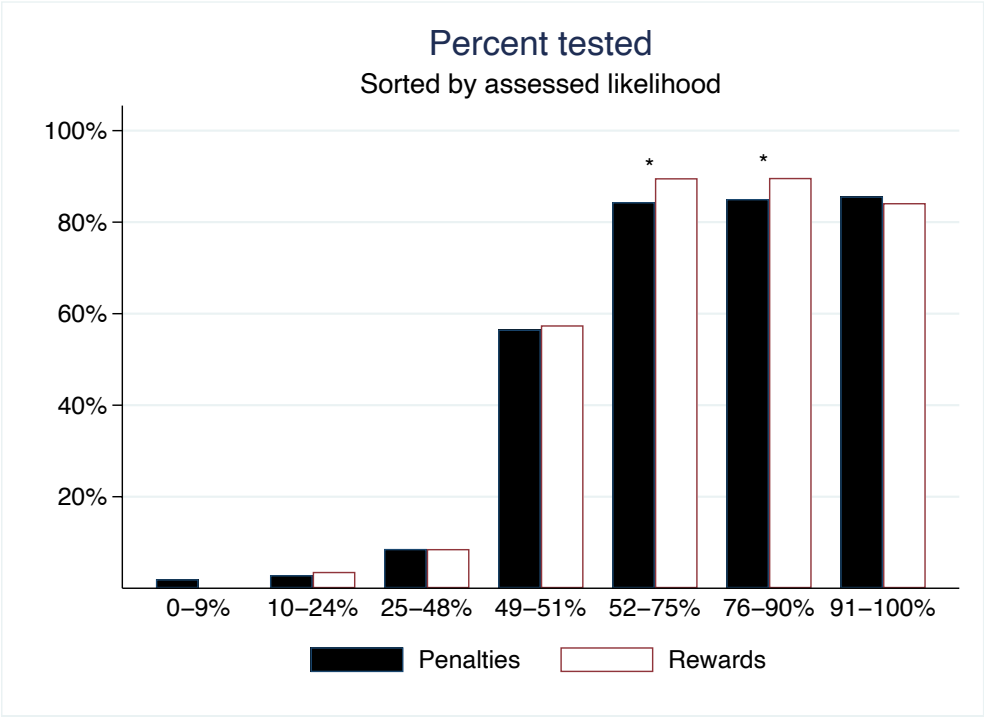

*Panel A: Assessed likelihood of mislabeling by composition*



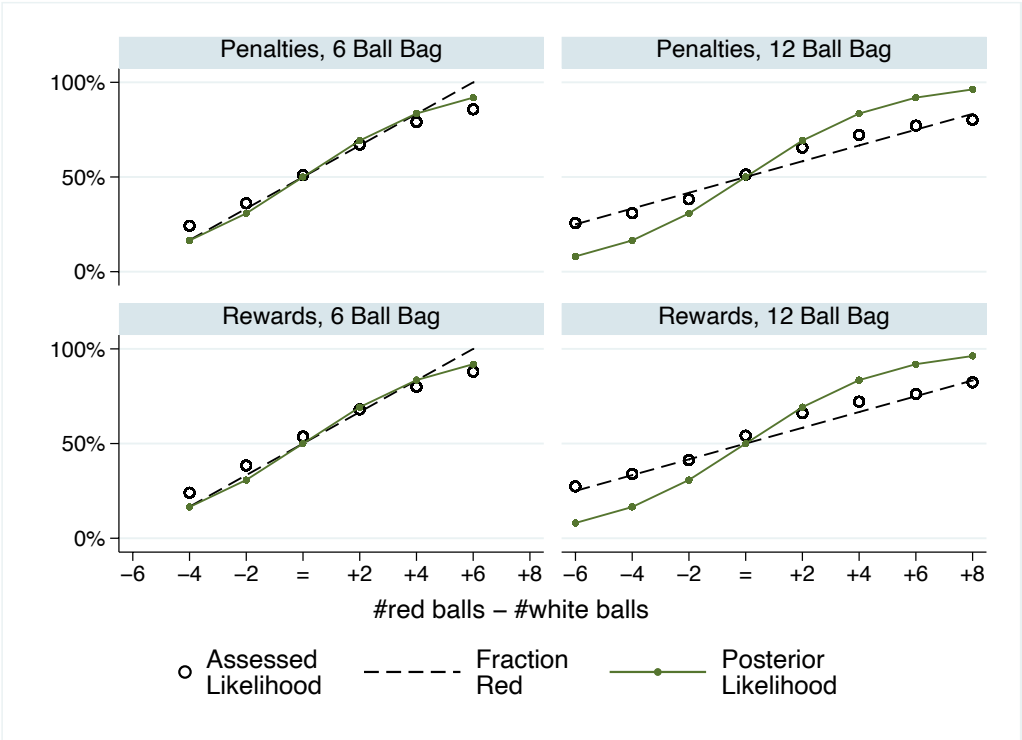

*Panel B: Judgment bias by composition*

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 at Table 2 and Table 3.

**Figure 4: Testing Action Conditioning on Risk Judgment and Risk Judgment by Bag Size**




*Panel A: Percent tested by assessed likelihood*




*Panel B: Assessed likelihood by incentive frame and bag size*

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