**Evaluation of Internal Control Combinations: Biased Judgments or Critical Thinking?**

**Ralph E. Viator**
**Professor**

**Penelope Lee Bagley**
**Assistant Professor**

**Area of Accounting**
**Rawls College of Business**
**Texas Tech University**
**Lubbock, TX 79409**

**We appreciate the constructive feedback provided by Michael Bamber, Jacqueline Hammersley, Derek Dalton, Marc Ortegren, Nancy Harp, and Beau Barnes. Funding for this study was provided by a research grant from the Rawls College of Business.**

**Evaluation of Internal Control Combinations: Biased Judgments or Critical Thinking?**


**SUMMARY**

PCAOB Auditing Standard No. 5 requires auditors to consider that multiple control deficiencies***, in combination***, may constitute a significant deficiency or material weakness. Unfortunately, prior accounting research indicates that auditors exhibit a conservatism bias when assessing the likelihood that individual internal controls are deficient. However, prior research has not examined whether auditors are likely to introduce additional bias when evaluating internal control combinations, or, exercise critical thinking skills and propose rational adjustments to conjoint probabilities associated with internal control combinations. In this study, we use a computer-based lab experiment to examine this issue. We find that experienced participants, compared to novices, are able to utilize performance feedback and make rational adjustments to conjoint probabilities that compensate for the conservatism bias in their initial assessments and avoid both a focalism bias and hypothesis framing bias. Our results suggest that by reflecting on their initial assessments, audit professionals are able to override judgment bias and make rational adjustments regarding the reliability of internal control combinations.



**Keywords:** cognitive biases, critical thinking, probability assessment, internal controls


**Data Availability:** Data are available upon request.

## INTRODUCTION

Per Auditing Standard No. 5 (AS5), auditors' evaluation of internal controls over financial reporting first considers whether individual controls are deficient (either in design or in operation). Auditors then consider whether there is a ***reasonable possibility*** that as a result of a deficiency, the internal control system will fail to prevent or detect a misstatement in an account balance or disclosure (PCAOB 2007). Furthermore, auditors must assess whether it is reasonably possible that two or more control deficiencies in combination will fail to prevent or detect a misstatement in the financial statements. PCAOB guidelines clearly emphasize the control risk posed by internal control combinations.

> Multiple control deficiencies that affect the same financial statement account balance or disclosure increase the likelihood of misstatement and may, ***in combination***, constitute a material weakness, even though such deficiencies may individually be less severe (emphasis added) (PCAOB 2007, 26).

Prior research indicates that auditors have difficulty assessing the likelihood of individual control deficiencies, raising questions regarding the quality of assessments for internal control combinations. For example, prior research finds that auditors' assessments of individual control deficiencies exhibit a conservatism bias: estimates are "too close" to prior probabilities (initial beliefs) regarding the likelihood of deficiencies and insufficiently approach extremities as indicated by normative theory, such as Bayes' Theorem (Joyce and Biddle 1981; Kinney and Uecker 1982; Holt 1987; Nelson 1995). This judgment bias is well documented in the extant literature (Brenner et al. 2005; Bazerman 2006; Griffin and Tversky 2003); however, its logical extension and implication for today's auditing profession might be overstated. For example, if auditors independently assess two internal controls and understate the likelihood that each control is reliable (e.g., assess controls as 70% and 80% reliable when in fact each control is 90% reliable), do auditors believe that the conjoint probability for **both controls** being reliable is only 56% (i.e., 70% * 80%)?[1] Or, alternatively, do auditors reflect on the implication of their initial assessments and override some of the bias compounded when considering a combination of internal controls?[2] Such reflection and overriding of biased judgment is indicative of critical thinking (Stanovich and West 2008;

[1] In this example, the normative conjoint probability of the reliability of the two controls is 90% * 90% = 81%, therefore the mathematical extension of the auditors' assessments is excessively understated. Alternatively, if auditors overstate the likelihood each control is reliable, (for example, 30% and 20% reliability versus the normative of 15% and 5%) then the auditors' conjoint probability (6%) is overstated compared to the normative (1%).

[2] For demonstration purposes, we are assuming that the example controls are complementary internal controls.

West et al. 2008). AS5, which requires auditors to assess the likelihood that combinations of internal controls represent material weaknesses, motivates us to understand whether auditors are capable of detecting and overriding their own biased inputs to assessments of internal control reliability. By reframing one of the traditional questions examined in audit judgment research, we shift the focus from documenting audit judgment bias to investigating auditors' ability to think critically. We examine whether audit professionals are capable of detecting biases in their initial judgments and overriding those biases with assessments that more closely approximate normative outcomes.

We utilize the classic book-bag and poker-chip game to examine how auditors' evaluation and revision of conjoint probabilities derived from conservative conditional probabilities are influenced by experience, feedback, and hypothesis framing. In particular, we are interested in whether conjoint probabilities remain conservative, become more conservative, or, alternatively, become less conservative. We present the game as described in other audit studies of subjective probability assessments: urns of marbles, where certain marbles (green marbles in this study) indicate reliable internal control processing and other marbles (red marbles) indicate internal control defects (Nelson 1995). However, we modify the game in three ways. First, for each case considered, participants view two separate samples (from two separate independent controls) and assess two independent subjective conditional probabilities. Second, we provide participants with three possible internal control combinations and the conjoint probabilities mathematically-derived from the subjective conditional probabilities. Participants are given the opportunity to reflect on the adequacy of those conjoint probabilities and enter adjustments to those conjoint probabilities. Third, hypothesis framing is manipulated by participants focusing on either control risk (the likelihood of having two reject urns) or control reliability (the likelihood of having two acceptable urns).

We utilize two participant groups for our study. Our experienced group consists of graduate students who have completed internships in auditing, are currently enrolled in a Master of Science in Accounting (MSA) program, and have completed multiple graduate courses in accounting and auditing. Our novice group consists of junior-level undergraduate students who have completed basic courses in accounting, but not auditing, and have no public accounting experience. We contend the MSA students are an adequate proxy for practicing auditors for two reasons. First, the experimental task (described below) is relatively low in integrative complexity; given the education and experience of the MSA students, prior research implies that practicing auditors would perform similarly on this task (Elliott et al. 2007 143 – 144). Second, as explained in the Experimental Methods section, MSA students at the host institution are likely to possess relatively high levels of critical thinking and analytical skills comparable to practicing auditors.

We argue that the primary challenge for rational adjustment of conjoint probabilities is recognizing the direction of the conservatism bias in initial subjective conditional probabilities (understated versus overstated) and revising in the appropriate direction. For example, when conservatively assessed conditional probabilities are ***understated***, conjoint probabilities mathematically-derived from individual conditional probabilities are ***excessively understated*** (due to a multiplicative effect). In this scenario, significant upward adjustments to such conjoint probabilities are rational adjustments. However, when conservatively assessed conditional probabilities are ***overstated***, downward adjustments to mathematically-derived conjoint probabilities are rational. We find that experienced participants are able to recognize the direction of conservatism bias and make rational adjustments to conjoint probabilities. Furthermore, we find that upon receiving feedback scores for conjoint probability assessments, experienced participants not only make rational adjustments to later assessments of conjoint probabilities, they also improve initial assessments of individual conditional probabilities, thereby diminishing the conservatism bias. Our results also indicate that experienced participants are not susceptible to framing effects. Experienced participants in a between-subjects design make final conjoint probability assessments that do not differ by hypothesis frame (focus on control risk versus control reliability).

We extend the accounting literature by examining auditors' assessments of conjoint probabilities and addressing a question posed by Bonner (2008): are biased assessments of individual discrete events troublesome for the auditing profession, or are their ill effects minimal? Specifically, we examine whether the conservatism bias is robust when experienced participants have the opportunity to reflect upon their initial assessments and adjust conjoint probabilities mathematically-derived from individual conditional probabilities (essentially re-assessing those initial conditional probabilities). Furthermore, our study complements a recent study by Ganguly and Hammersley (2009) that examines revisions of auditing covariation estimates (i.e., the degree of association between a "clue" and its potential "condition") after purchasing additional information. The results in both studies suggest that the opportunity to reflect on initial estimates and provide revisions to those estimates enables overriding bias in initial judgments.

We contribute to auditing practice by examining how auditors evaluate internal control combinations, as specified by AS5. Understanding whether auditors are likely to introduce additional bias and become more conservative when evaluating internal control combinations, or, exercise critical thinking skills and propose rational adjustments that reduce the bias in conjoint probabilities derived from internal control combinations, is important given that understated (overstated) control risk leads to ineffective (inefficient) audit work.

The remainder of the paper is organized as follows. We provide the theoretical background for our hypotheses in the second section, details of the experimental method in the third section, the results of statistical analyses in the fourth section, and conclusions based on experimental results in the fifth section.

## THEORY AND HYPOTHESES

### Rational versus Biased Adjustments to Conjoint Probabilities for Internal Control Combinations

West et al. (2008, 931) argues that "the traditional heuristics and biases studied by cognitive psychologists should be considered to be part of a broadened concept of critical thinking." Such a proposition is counterintuitive. Although heuristics and biases tend to result in poor judgments and decisions, as compared to those predicted by normative theory, West et al. posit that cognitive processes which override heuristics and minimize or avoid biased judgments are a component of critical thinking. The authors empirically demonstrate that the ability to avoid several forms of biased judgments, including insensitivity to sample size and non-Bayesian updating of belief, is, at a minimum, moderately correlated with traditional measures of critical thinking (Stanovich and West 2007; Cacioppo et al. 1996) and the ability to reason independently of prior belief (Markovits and Nantel 1989). Their results suggest that participants who possess stronger critical thinking skills are more likely to compensate for biased judgments emanating from well-known cognitive heuristics.

Perhaps more than most disciplines, the auditing profession focuses on developing critical thinking skills (AICPA 2009; Springer and Borthick 2007; Wyhe 2007). Thus, on the surface, it is puzzling that prior audit studies consistently report that auditors exhibit heuristic thinking and biased judgment. For example, relying on sample representativeness and insufficiently weighing base rate and sample size information leads to conservative assessment of conditional probabilities, such as the likelihood that a control deficiency is significant (Bonner 2008, pp. 149 – 152; Nelson 1995; Holt 1987; Kinney and Uecker 1982; Joyce and Biddle 1981). However, prior audit studies of probabilistic reasoning have not provided an environment in which participating auditors could exercise critical thinking skills. In the current study, we provide participants the opportunity to reconsider their initial assessments of conditional probabilities and revise upwards or downwards the conjoint probabilities mathematically-derived from those subjective conditional probabilities. We examine auditors' potential to propose rational adjustments to conjoint probabilities under two different scenarios: when subjective conditional probabilities are understated versus overstated vis-à-vis Bayes' Theorem. We propose that experienced participants will critically consider the reasonableness of conjoint probabilities for internal control combinations and act rationally by adjusting the conjoint probabilities in the proper direction: increasing (decreasing) odds when conjoint probabilities are understated (overstated).

Alternatively, rather than discriminating between understated and overstated conjoint probabilities for internal control combinations, we expect novices to adjust conjoint probabilities in a single direction: increasing the odds regardless of whether conjoint probabilities are understated or overstated. We attribute this behavior to the problem of focalism: "people focus too much on the occurrence in question (termed the focal event) and fail to consider the consequences of other events that are likely to occur" (Wilson et al. 2000 822).[3] In the context of the current study, participants' attention is focused on one of two outcomes: an internal control combination has no significant deficiencies, or both controls (in a combination of two controls) have significant deficiencies. We expect novices to focus "too much on the occurrence in question" and increase conjoint probabilities for the outcome in focus, regardless of whether the evidence suggests that the outcome is more likely or less likely. Such adjustments do not reflect critical thinking and, on their own, are biased (consistently increasing the odds regardless of the circumstances).

[3] Similar to Wilson et al.'s (2000) description of the problem of focalism, Schkade & Kahneman's (1998) describe the *focusing illusion bias* as one that occurs "when a judgment about an entire object or category is made with attention focused on a subset of that category, … whereby the attended subset is overweighted relative to the unattended subset" (p. 340).

Based on the preceding discussion, we expect to observe two distinct behaviors for experienced participants versus novices. Experienced participants should adjust conjoint probabilities in an appropriate direction by increasing (decreasing) the odds when initial conjoint probabilities are understated (overstated). On the other hand, novices are likely to increase odds assigned to conjoint probabilities regardless of understatement / overstatement. Stated formally:

> **H1:** Experienced participants, compared to novices, propose rational adjustments to conjoint probabilities for internal control combinations, increasing (decreasing) odds when conjoint probabilities are understated (overstated).

**Effect of Performance Feedback (Conjoint Probability Scores)**

Prior research on debiasing human judgment has focused on intensive, personalized feedback. For example, Bazerman (2006) outlines three detailed steps for debiasing judgment: (1) unfreezing ingrained thinking and behavior, (2) changing judgment processes by clarifying the existence and explaining the roots of specific judgmental deficiencies, and (3) refreezing new strategies by practicing over time and reviewing past training until new strategies become intuitive. These recommendations are in agreement with Lichtenstein and Fischhoff's (1980) seminal study which demonstrates that intensive, personalized performance feedback in the form of calibration graphs and detailed performance scores is effective in reducing the overconfidence bias in a forced-choice two-alternatives general knowledge task. A more recent study by Stone and Opel (2000) demonstrates that the use of intensive, personalized performance feedback tailored to a specific bias and followed by immediate retesting can be effective for bias reduction. Such an approach agrees with Larrick's (2004) recommendation that debiasing training should be bias specific and closely linked to subsequent testing. In summary, research on debiasing training suggest that personalized intervention is necessary before individuals will "unfreeze" their prior behavior and adopt new, less biased behavior.

This conclusion from debiasing judgment research stands in stark contrast to West et al.'s (2008) proposition that one form of critical thinking is overriding heuristics and avoiding biased judgment. Thus, not all individuals should require intensive, personalized feedback to debias their judgments. Rather, as West et al. emphasize, a component of critical thinking is to reason logically when logic conflicts with prior beliefs. In other words, when presented information that contradicts prior beliefs, individuals who think critically should override the heuristics they previously used and exercise judgment that is less biased. In our study we expect experienced participants, as opposed to novices, to think more critically and better interpret feedback information, generating less biased conjoint probabilities after receiving feedback. Specifically, we examine participants' ability to utilize performance feedback in the form of case-by-case conjoint probability scores that compare participants' subjectively-derived conjoint probabilities to normative (Bayesian-derived) conjoint probabilities.

Performance feedback in the form of conjoint probability scores should provide two signals to participants. First, the feedback indicates the direction of the bias (overstatement versus understatement) for specific conditions (e.g., likely versus unlikely significant control deficiencies). Second, the feedback indicates the magnitude of the bias. However, participants must not only critically observe the direction and magnitude of their bias, but they must appropriately modify at least one of two beliefs: (1) the belief that their initial assessment of individual conditional probabilities is appropriate, or (2) the belief that their adjustments to conjoint probabilities mathematically-derived from those conditional probabilities are appropriate. Critical thinking — thinking logically when logic conflicts with prior beliefs (West et al. 2008 930) — challenges one, if not both, of these beliefs. We expect that after receiving performance feedback experienced participants are more likely than novices to challenge their beliefs and substantively change their assessment of probabilities. By exercising critical thinking, experienced participants should increase the magnitude of their adjustments to conjoint probabilities and submit larger positive (negative) adjustments when conjoint probabilities are understated (overstated). On the other hand, because performance feedback in the form of conjoint probability scores is only indirectly related to the underlying individual conditional probabilities, we expect both participant types to remain conservative in their initial assessment of individual conditional probabilities. These arguments lead to a null hypothesis (H2) and a directional hypothesis (H3).

> **H2:** After receiving performance feedback in the form of conjoint probability scores for internal control combinations, experienced participants and novices do **not** improve the quality of their assessment of individual conditional probabilities (the likelihood that a single internal control is significantly deficient).

**H3:** After receiving performance feedback in the form of conjoint probability scores for internal control combinations, experienced participants are more likely than novices to adjust conjoint probabilities in the appropriate direction and increase the magnitude of those adjustments.

**Comparative Conjoint Probabilities and Hypothesis Framing Effects**

In the final analysis, auditors must assess whether internal control combinations have material weaknesses, or, are reliable. Thus, an issue for this study is whether hypothesis framing, in terms of control risk versus control reliability, matters. The audit risk model focuses on minimizing control risk (AICPA 2008, Messier et al. 2008), while control reliability frameworks, such as the SysTrust® framework utilized by information systems auditors, emphasizes maximizing control reliability (AICPA 2006; Romney and Steinbart 2009). Focusing on ***control risk*** versus ***control reliability*** is an example of attribute framing: "positive or negative wording of a key element of a single object or event about which subjects are asked to make a judgment" (Bonner 2008 179). In the current study, participants' initial conditional probabilities across frames are easily reconciled: the level of control risk associated with a single control is the inverse of the level of control reliability. For example, if an auditor assesses that an individual control has a 70% chance of being significantly deficient, he/she is implicitly stating that there is only a 30% chance that the control is reliable. A framing effect occurs if, given the same information, auditors assessing control reliability provide a different set of probabilities. Therefore, identifying whether a framing effect exists for initial conditional probabilities is straightforward.

However, identifying framing effects for combinations of internal controls is more complex. In a simple combination of two separate internal controls, the potential conjoint probabilities in the control reliability frame are **not** simple inverses of the control risk frame. Using the example probabilities previously stated, if an auditor assesses that each of two separate individual controls has a relatively high chance of being significantly deficient (say 70% and 80%, respectively), then the conjoint probability for **both controls being significantly deficient is 56%**. However, using the control reliability frame, equivalent initial assessments of control reliability would be 30% and 20%, respectively. Thus, the conjoint probability that **both controls are reliable is 6%**. These two outcomes (both controls are significantly deficient and both controls are reliable) are not inverses of each other because a third outcome is possible: **one, but not both, controls is significantly deficient (reliable) (38%)**. Therefore, in order to identify any framing effects, we ask participants to review and possibly adjust the conjoint probabilities for all three outcomes. Participants in the control risk frame focus on both controls being significantly deficient, yet still evaluate the conjoint probability that both controls are reliable. The reverse is true in the control reliability frame. Thus, we examine whether conjoint probabilities assigned to potential outcomes varies by whether the outcome is ***in focus*** or ***out of focus***.

A series of studies by Emby (1994) and Emby and Finley (1997) manipulate framing effects by asking auditors to evaluate the effect of additional evidence on the risk (strength) of the internal control system. In general, both of these studies find that framing effects are easily induced. However, Emby and Finley demonstrate that a debiasing technique (requiring participants to rate the direction and relevance of the evidence) can significantly mitigate framing effects in such contexts. Once again, the proposition that debiasing techniques are required in order to mitigate a specific bias (in this case, framing effects) is contradictory to West et al.'s (2008) proposition that some individuals engage in deeper levels of critical thinking and are able to override heuristics and avoid biased judgment. Applying this proposition to our study of framing effects bias, we would expect experienced participants to carefully consider each of the three outcomes previously identified and adjust their conjoint probabilities in an appropriate direction and magnitude regardless of whether the outcome is **in focus** or **out of focus**.[4] Thus, we expect experienced participants to differ from novices in overriding the hypothesis framing bias. Specifically, experienced participants, but not novices, should assign similar conjoint probabilities to equivalent outcomes, regardless of whether the outcome is *in focus* or *out of focus*. However, this effect is more likely to occur after receiving performance feedback, in which experienced participants critically evaluate their conjoint probability scores and override any hypothesis framing bias.

> **H4:** After receiving performance feedback in the form of conjoint probability scores, experienced participants are more likely than novices to submit conjoint probabilities for comparable outcomes that do not differ by hypothesis frame.

[4] In the experiment, one of the outcomes (one, but not both, controls is significantly deficient) is neither *in focus* or *out of focus*. Conjoint probability adjustments to this disjunctive set are simply the differences between adjustments to the two conjunctive sets. Thus, we do not provide additional discussion or data analysis for the disjunctive set.

## EXPERIMENTAL METHOD

### Participants

All participants are enrolled at a large public university in the United States. The primary group of participants consists of 52 graduate students enrolled in a Master of Science in Accounting (MSA) program who completed an audit internship in public accounting, as well as advanced courses in Accounting, Auditing, and Statistics. 50% are enrolled in a graduate capstone course (the "final course in the MSA program"), while the remaining 50% are enrolled in a graduate auditing course, focusing on information systems auditing. The MSA students are classified as *experienced participants* since they have completed an audit internship and represent those accounting students who most likely posses the analytical and critical thinking skills required to succeed in both the MSA program and the accounting profession.[5] The experienced participants identify themselves as having a GPA above 3.0; 67% indicate having a GPA between 3.6 and 4.0.

The second group of participants serves as a control group and consists of 54 junior-level accounting students with no accounting or auditing work experience who are enrolled in an introductory accounting information systems course. Only 43% of the undergraduate students indicate having a GPA between 3.6 and 4.0, and 11% indicate having a GPA between 2.5 and 3.0. These participants are identified as "novices" as they have completed only a minimum number of courses in accounting, have minimum exposure to courses teaching analytical and critical thinking skills, have relatively lower GPAs, and have no experience in public accounting (have not completed an auditing internship).

The MSA graduate students volunteer to participate outside of class time. To encourage participation, MSA students receive a minimum payment of $20. The MSA students play the computer-based game with a reward pool of $1,600. The median payout per participant is $30, with a range from $20 to $40. Undergraduate students participate during class time and receive a 10 point credit. Since their effort to attend the experimental session is significantly less than the MSA students, their reward

[5] At the participating university, students intending to take the CPA exam must qualify and enroll in the MSA program (in order to complete courses required for the CPA exam). A relatively small percentage of undergraduate accounting majors (less than 30%) qualify for the MSA program based on GPA, GMAT scores, and letters of recommendations. The participating university uses these factors to identify students that most likely possess the analytical and critical thinking skills required to succeed in the MSA program, as well as the accounting profession. Essentially, compared to the general population of accounting majors, students with the necessary skills to succeed in the accounting profession self-select to enroll in the MSA program.

pool is set at $800. The median payout per participant is $15, with a range from $10 to $20. Students earn a higher percentage of the reward pool by obtaining higher conjoint probability scores. We calculate the percentage of points earned by each participant in each participant group, and then standardized those percentages to ensure that minimum and maximum payouts are met.

One participant from each group was unable to complete the experiment due to computer-based technical problems and received the median compensation for that group. Members of each participant group are randomly assigned to one of two treatment groups based on hypothesis frame (control risk or control reliability). The MSA students are randomly split 49% - 51% between the two treatment groups, while the undergraduate students are split 53% - 47%. In total, we conduct four experimental sessions: one session for each treatment group and each participant type.

**Experimental Task and Procedures**

We conduct the experimental sessions in a computer lab; participants access a web-based computer program for task instructions and completing the probability assessments. Similar to Nelson (1995), the experimental task is derived from the classic book-bag and poker-chip game in which participants view sample evidence and assess the likelihood that the sample is drawn from a specific bag. However, we modify the game by having participants view two separate samples (from two separate independent controls) and evaluate the reasonableness of conjoint probabilities mathematically-derived from their initial assessments. In the current study, participants focus on whether sample evidence comes from one of two urns, both of which have an equal chance of being selected (i.e., prior probabilities are set at 50%). *Acceptable Urns* contain 80% green marbles and 20% red marbles, while *Reject Urns* contain 60% green marbles and 40% red marbles. For MSA students with audit training and experience, the urns are labeled to represent different levels of internal control reliability. *Acceptable Urns* indicate that an internal control is functioning at an acceptable level (working 80% of the time); while *Reject Urns* indicate that an internal control is **not functioning** at an acceptable level (working only 60% of the time).[6]

[6] In order to avoid extreme values for conjoint probabilities, we set higher rates for both prior probabilities and defect rates, compared to prior studies. For example, Nelson's (1995) study employs prior probabilities and defect rates that generate Bayesian conditional probabilities that are at the extreme (either greater than 90% or less than 20%). Although two independent samples with Bayesian conditional probabilities of 90% and 95% would yield a reasonable conjoint probability target of 85.5%, the alternative hypothesis frame, based on Bayesian conditional probabilities of 10% and 15%, would yield an extremely low conjoint probability of 1.5%. We recognize that low prior probabilities and low defect rates are characteristic of audit sampling and evaluation; however, in this study, in order to provide a wide

For novices, the urns are labeled either acceptable or reject, with no reference to internal control conditions.[7]

In each case presented, participants view sample evidence from two separate and independent controls (two separate sets of urns). As shown in Figure 1, participants consider sample information and enter their subjective conditional probability for a single internal control condition, revising either upwards or downwards the prior probability of 50%. The procedure is repeated for a second and separate internal control condition. Essentially, we use this procedure to induce uncertainty regarding whether each internal control individually is significantly deficient, or not. However, our primary interest is participants' evaluation of internal control combinations, given the uncertainty that one or both internal controls is functioning.

<Figure 1`>

As shown in Figure 2, we have participants consider three different internal control combinations whose conjoint probabilities are uncertain. We use terminology from the auditing literature to focus experienced participants' attention on different levels of internal control deficiencies: two Acceptable Urns represents "no control deficiency", one Acceptable and one Reject Urn (together) represent a "significant deficiency," and two Reject Urns represents a "material weakness".[8] Thus, participants must consider the uncertainty associated with each internal control combination and judge whether the conjoint probabilities associated with each combination is appropriate, or should be revised. Participants enter their "best estimate" for conjoint probabilities, which could be the conjoint probabilities mathematically-derived from their initial conditional probabilities (as displayed on the computer screen), or alternatively, a revised estimate of conjoint probabilities. The computer program enforces the rule that odds assigned to the three outcomes total to 100%.

<Figure 2>

---

range of Bayesian-based conjoint probabilities and avoid participants expecting only extreme conjoint probabilities, we set prior probabilities at 50% and utilized relatively high defect rates.

[7] As previously noted, novice participants have no accounting or auditing work experience and have not taken an auditing course. As such, novices are unfamiliar with the terms related to internal control conditions. Therefore, in order to avoid confusion over terminology, we use a context free setting for novices to perform the experiment.

[8] We examine whether participants exhibit critical thinking or biased judgment when evaluating the uncertainties associated with internal control combinations. We do not examine or manipulate *misstatement magnitude*, which further differentiates material weaknesses and significant control deficiencies.

Participants focusing on internal control reliability assess the odds of having an acceptable urn (Figure 1), then assess the conjoint probabilities for the three internal combinations, viewing those outcomes in the order listed above (no control deficiency, significant deficiency, and then material weakness) (Figure 2). However, participants focusing on control risk assess the odds of having a reject urn and view the three outcomes in reverse order (i.e. material weakness, significant deficiency, and then no control deficiency). Thus, we manipulate hypothesis framing as a between-subjects effect. Participants assess one of two different types of conditional probabilities (i.e., either the likelihood of acceptable urn or the likelihood of reject urn). Furthermore, participants in different hypothesis frames view different orderings of internal control combinations (i.e., "no control deficiencies" listed first versus "material weakness" listed first). For novice participants, the outcomes are simply identified by the number of acceptable versus reject urns.

We develop twenty-four unique samples for this study and utilize them to manipulate two within-subject effects: sample size and sample representativeness. Half of the samples are relatively small-size (n = 10, 15, or 20) and the other half are relatively large-size (n = 30, 35, or 40). For sample representativeness, half of the samples have proportions of green marbles that are greater than 70%, and thus more representative of an acceptable urn (i.e., 80% green marbles), while the other half have proportions less than 70%, and thus more representative of reject urns (i.e., 60% green marbles). Furthermore, sample representativeness generates Bayesian conditional probabilities that are either greater than 50% (signaling reliability) or less than 50% (signaling control risk). These two manipulations (sample size and sample representativeness) are counterbalanced within subjects. Table 1 displays the twenty-four unique samples used in this study.

<Table 1>

Participants complete 48 cases across four rounds (12 cases per round). During Round 1, participants assess subjective conditional probabilities for two independent controls, then review conjoint probabilities derived from those assessments and enter adjustments to reflect "their best estimate" (Figure 2). During Rounds 2 and 3, we provide participants with performance feedback in the form of case-by-case conjoint probability scores that compare their "best estimate" of conjoint probabilities to Bayesian-based conjoint probabilities. See Figure 3. In Round 4 (the final round), participants assess probabilities in the same manner as Round 1 and receive no performance feedback. Thus, in a pretest-posttest experimental design, we examine whether performance feedback in the form of conjoint probability scores impacts either (1) the assessment of subjective conditional probabilities for individual internal control conditions, or (2) the adjustment of conjoint probabilities mathematically-derived from those conditional probabilities.

<Figure 3>

To calculate conjoint probability scores (labeled “Total Points Earned” in Figure 3), we convert each participant’s final estimates and Bayesian-based estimates (i.e., labeled “Math Estimate” in Figure 3) to whole numbers and calculate the absolute difference for each outcome set (each internal combination). It is theoretically possible for the sum of the absolute differences across outcome sets to total to 200 points. Therefore we calculate conjoint probability scores as the total absolute difference from 200. The goal for participants is to maximize the number of points earned out of 200, on each of the internal control combination cases.

In each of the four rounds, participants view 12 cases, with each case consisting of two separate and independent samples. For Round 1, we utilize each sample once, forming 12 unique pairings. These same pairings are also utilized in Round 4 after participants receive performance feedback in Rounds 2 and 3. We also utilize each sample in Rounds 2 and 3; however, each case is unique. For example, sample #3 is paired with sample #6 in Rounds 1 and 4, but paired with sample #1 in Round 2 and sample #5 in Round 3.

Table 2 displays the 36 unique cases along with their conjoint probabilities for both hypothesis frames. We control for order effects in two different fashions. First, we vary the cases assigned to Rounds 2 and 3. Approximately half of the participants view “Round 2” cases in Round 2, while the other half view “Round 3” cases in Round 2. Second, we vary the case number on which participants begin each round. For each of the cases identified in Table 2, approximately one-twelve of the participants begin each round on that case. We subsequently test for both order effects. Furthermore, throughout the experiment, the computer program randomly selects which of two samples in each case is viewed first versus viewed second.

<Table 2>

Participants complete 48 cases across four rounds, viewing 96 samples in total. Prior to playing the game, we provide participants with instructions regarding two key attributes of sample evidence (sample size and sample representativeness) and navigation of the web-based computer game. Instructional materials that map acceptable and reject urns combinations to internal control combinations are presented in the Appendix. Each experimental session is held in a college of business computer lab at a large public university. At the end of each experimental session, participants complete a computer-based debriefing questionnaire.

**Experimental Design**

We manipulate two between-subjects variables: experience level (experienced participants versus novices) and hypothesis frame (focus on acceptable urns versus reject urns). As described above, hypothesis framing is reinforced through the labeling and ordering of internal control combinations. We also manipulate three within-subjects variables: sample size (large versus small), sample representativeness (Bayesian conditional probabilities greater than 50% versus less than 50%), and order (cases in “Round 2” viewed prior to “Round 3”, and vice-versa). The prior probability of 50% is held constant across all rounds and cases.

We analyze multiple dependent variables. For H1 through H3, we limit our analysis to internal control combinations *in focus*, such as *no control deficiencies* (two acceptable urns) displayed in Figure 2). We calculate *raw adjustment* as the difference between participants’ “best estimate” of conjoint probability (displayed as a data entry area in Figure 2) minus the conjoint probability mathematically-derived from participants’ two subjective conditional probabilities (displayed on the first line under “Overall Likelihood” in Figure 2). However, interpreting whether *raw adjustments* are **rational** (improving the conjoint probability scores) or **biased (**e.g., arbitrarily increasing odds when control combinations are *in focus*) requires comparing *raw adjustments* to *initial distance*. We calculate *initial distance* as the difference between conjoint probabilities mathematically-derived from participants’ two subjective conditional probabilities (displayed in Figure 2 on the first line under “Overall Likelihood”) and those conjoint probabilities mathematically-derived from Bayesian conditional probabilities (displayed in Figure 3 on the first line under “Math Estimate”). Positive *raw adjustments* when *initial distance* is negative (conjoint probabilities are understated) are rational; however, positive *raw adjustments* when *initial* *distance* is positive (conjoint probabilities are overstated) are **not** rational and reflect holding onto initial beliefs rather than critically considering alternatives.

For H4, we utilize *initial distance* (described above) and *final distance*. We calculate *final distance* as the difference between participants' "best estimate" of conjoint probabilities (their "Revised Estimate" shown in Figure 2) and Bayesian-based conjoint probabilities (i.e., the "Math Estimate" shown Figure 3). We examine for differences attributed to participant experience level, hypothesis framing, and focus type. Focus type is whether a specific internal control combination (such as no control deficiencies) is listed first or last when considering all three possible combinations. Although focus type is tied to hypothesis framing (no control deficiencies is listed first in the control reliability frame, but last in the control risk frame), each treatment level for hypothesis framing requires participants to assess the conjoint probability for each of the three possible internal control combinations. Therefore, for H4, we examine whether *final distance* for identical control combinations (such as "no control deficiency") varies by hypothesis frame (i.e., control risk versus control reliability).

## RESULTS

### Manipulation Checks

Prior to Round 1, we perform two manipulation checks to verify that participants understand the implications of sample representativeness and sample size. Participants have to provide the correct answer before proceeding. On their first attempt at answering a question regarding sample representativeness, 98% and 92% of experienced and novice participants, respectively, recognize which sample is more representative of the targeted object (e.g., an acceptable urn). Regarding sample size, 96% of the experienced participants, versus only 75% of the novice participants, recognize on their first attempt that a larger sample has a stronger signal. After receiving performance feedback (conjoint probability scores) in Rounds 2 and 3, participants complete a second verification prior to Round 4. These two questions on sample representativeness and sample size are more challenging and require participants to consider **disconfirming information**: singals that the sample does **not** come from the targeted urn. Regarding sample representativeness, 95% of the experienced participants recognize the correct answer on either their first or second attempt, compared to only 79% of novice participants. Regarding sample size, 90% of the experienced participants recognize the correct answer on either their first or second attempt, compared to only 74% of novice participants. In summary, participants easily process information regarding sample representativeness and less easily process information regarding sample size, especially when viewing disconfirming information. Also, in general, experienced participants more carefully consider sample information, demonstrating more critical thinking compared to novice participants.

**Test of Hypotheses**

Initial statistical analysis indicates that, for half of the 48 cases, participants submit subjective conditional probabilities that are very close to Bayesian conditional probabilities and make very small adjustments to conjoint probabilities. For example, the average *initial distance* in Round 1 is less than 5 percentage points and the average *change in initial distance* from Round 1 to Round 4 is less than 2 percentage points. Furthermore, the revised conjoint probabilities (participants' "best estimates") in Round 4 are very close to Bayesian, ranging from 1.5 to 5.0 in *final distance*. These 24 cases are directly tied to the 12 samples indicating acceptable urns (i.e., low defect rates). For these samples, participants could submit sample proportions as their subjective conditional probabilities and be relatively close to Bayesian conditional probabilities. For example, in Table 1 Sample #17 has 15 green marbles in a sample of 20 marbles; the sample proportion is 75% and the Bayesian conditional probability is 70.1%. Participants could employ the strategy of anchoring on sample proportions and making modest adjustments when defect rates are relatively low. However, this strategy is ineffective when defect rates are relatively high. For example, Sample #5 has 12 green marbles out of 20 (a sample proportion of 60%), but its Bayesian conditional probability is much lower (10.9%). For cases based on samples with relatively high defect rates, if participants anchor on sample proportions, then larger adjustments are necessary to adequately approach Bayesian conditional probabilities. Providing relatively large adjustments for these cases is rational and indicative of critical thinking.

Because of relatively close *initial distances* and modest *raw adjustments* to conjoint probabilities for the 24 cases generated by samples indicating acceptable urns (i.e., relatively low defect rates), we do not present statistical analyses for these cases. However, we suggest that these cases, which are randomly intermingled with those having relatively high defect rates, create a distractor task (Davelaar et al. 2006; Mulligan et al. 2007) that interferes with participants' cognitive processing. In other words, when presented samples with relatively low defect rates, participants might anchor on sample proportions and achieve relatively high conjoint probability scores; however, when assessing samples with relatively high defect rates, this heuristic would result in relatively low conjoint probability scores. Thus, to maximize conjoint probability scores, participants must recognize when to suppress the heuristic “sample proportions approximate conditional probabilities.”

To clarify: the statistical analyses presented below are for cases in which the sample signal (the combination of sample representativeness and sample size) **indicates having a reject urn**. When the hypothesis frame is control risk, we expect participants’ subjective conditional probabilities to be understated vis-à-vis Bayesian conditional probabilities. Conversely, when the hypothesis frame is control reliability, we expect participants’ subjective conditional probabilities to be overstated vis-à-vis Bayesian conditional probabilities. As described above, we are not able to observe variations in judgments for the inverse condition (i.e., sample evidence indicates having an acceptable urn and the hypothesis framing is either control risk or control reliability). We further discuss this limitation in the conclusion section of this paper.

***Hypotheses H1: Conjoint Probability Adjustments prior to Feedback***

The first hypothesis posits that, prior to receiving any performance feedback, experienced participants, compared to novices, propose rational adjustments to conjoint probabilities for internal control combinations, increasing (decreasing) odds when conjoint probabilities are understated (overstated). Table 3 Panel A presents cell means for *initial distances* and *raw adjustments* in Round 1, along with related two-tailed t-tests for whether *raw adjustments* are statistically different from zero. Panel B presents the overall ANOVA statistics for *raw adjustments*. Experienced participants make rational adjustments to conjoint probabilities in Round 1, prior to receiving performance feedback in Rounds 2 and 3. In the reliability frame (focus on acceptable urns), their average *initial distance* is a positive 16.63 and their average *raw adjustment* to conjoint probabilities is significantly negative (-1.17, $p < .006$). On the other hand, in the control risk frame (focus on reject urns), their average *initial distance* is -32.19 and their average *raw adjustment* to conjoint probabilities is significantly positive (+1.93, $p < .005$). Thus, experienced participants in a between-subjects design make appropriate adjustments to conjoint probabilities: positive (negative) adjustments when initial conditional probabilities are understated (overstated). The results suggest an awareness of the conservatism bias in initial assessments of conditional probabilities and an ability to override this bias when evaluating conjoint probabilities derived from those initial assessments.

<Table 3>

Conversely, in Round 1, novices make adjustments to conjoint probabilities that reflect the focalism bias: consistently increasing odds assigned to conjoint probabilities regardless of hypothesis frame. Novices' average *initial distances* are in the same direction as experienced participants: 17.86 and -23.41, for the acceptable urn frame and reject urn frame, respectively. However, their average *raw adjustment* to conjoint probabilities is positive in both frames: +2.82 ($p < .001$) and +4.99 ($p < .001$), respectively. Novices' consistently increasing odds assigned to conjoint probabilities regardless of hypothesis frame does not reflect critical thinking. In summary, the statistical results support H1.

***Hypotheses 2 and 3:* Effect of Performance Feedback**

H2 is a null hypothesis and states that after receiving performance feedback in the form of conjoint probability scores, experienced participants and novices do **not** improve their assessment of initial conditional probabilities. For H2, we calculate the *change in initial distance* as *initial distance* in Round 4 minus *initial distance* in Round 1, such that a negative changes (positive changes) move closer to Bayesian for Control Reliability / Acceptable Urns cases (Control Risk / Reject Urns cases). Since our pre-test post-test format utilizes the same cases in Round 1 and Round 4, we calculate *change in initial distance* for individual cases. This variable indicates whether participants' conservatism bias in assessing individual conditional probabilities is modified by performance feedback on conjoint probabilities, or remains unchanged.

Table 4 Panel A presents treatment cell means for *change in initial distances* from Round 1 to Round 4, along with related two-tailed t-tests and contrast of cell means; Panel B presents the overall ANOVA statistics for *change in initial distances*. Furthermore, since insensitivity to sample size is a key bias generating conservative subjective conditional probabilities, we present cell means partitioned by *sample size*, as well as by *experience level* and *hypothesis frame*. For experienced participants, each of the four treatment groups defined by *hypothesis frame* and *sample size* has an average *change in initial distance* that is significantly different from zero (largest $p$ value < .01, two-tailed test) and in the appropriate direction. Thus, for experienced participants, the null hypothesis H2 is not supported. Rather, after receiving performance feedback on the quality of their conjoint probabilities for internal control combinations, experienced participants modify their assessments of conditional probabilities associated with individual internal controls, by increasing or decreasing the odds in the appropriate direction. Such adjustments require substantial critical thinking because feedback on conjoint probabilities for internal control combinations does not directly address the quality of initial conditional probabilities for individual internal controls. Experienced participants appear to reason (correctly) that one way to improve their assessment of conjoint probabilities is to modify (in the appropriate direction) their assessment of individual conditional probabilities. Clearly, such judgment demonstrates a rational response that mitigates, but does not eliminate, the conservatism bias. However, we find no evidence that experienced participants improve their assessment of subjective conditional probabilities by incorporating sample size information; linear contrasts indicate that the magnitude of *change in initial distance* does not differ between the two levels of sample size.

<Table 4>

*Change in initial distances* for novice participants are relatively small and only moderately significant ($p < .05$, two-tailed test) in one treatment cell (acceptable urn with large sample size). These results support the null hypothesis (H2) in regards to novice participants: performance feedback in the form of conjoint probability scores does **not** enable novice participants to improve their assessment of individual conditional probabilities. The results suggest that novices engage in less critical thinking and make less effort to override the conservatism bias.

H3 states that after receiving performance feedback in the form of conjoint probability scores for internal control combinations, experienced participants are more likely than novices to adjust conjoint probabilities in the appropriate direction and increase the magnitude of those adjustments. Specifically, we expect experienced participants to compensate for conservative biased assessments of individual conditional probabilities by substantially decreasing (increasing) conjoint probabilities that are overstated (understated). Table 4 Panel A presents treatment cell means for *raw adjustments* in both Round 1 and Round 4, as well as *change in raw adjustments*, with related t-tests and contrast of cell means. Panel B presents the overall ANOVA statistics.[9] Similar to *change in initial distance*, we calculate *change in raw adjustments*, which indicates whether participants' modify the magnitude of *raw adjustments* from Round 1 to Round 4.

Experienced participants significantly increase the magnitude of their raw adjustments in all four treatment cells. Their *changes in raw adjustments* are significantly negative for large and small sample sizes (-2.95 and -2.56, respectively) when initial distances are overstated in the control reliability frame, and significantly positive (+7.86 and +5.71, respectively) when initial distances are understated in the control risk frame. The relevant t-tests indicate that each of these averages is greater than zero ($p < .01$, two-tailed test). It is important to note that these improvements occur in addition to the previously mentioned improvement: a reduction in the conservatism bias when assessing individual conditional probabilities. Thus, in Round 4, experienced participants provide less conservative initial individual conditional probabilities and larger raw adjustments to conjoint probabilities derived from those conditional probabilities.

As presented in Table 4, linear contrasts indicate that novices' *change in raw adjustments* in the Acceptable Urns frame are equivalent to experienced participants in the Control Reliability frame. However, the source of the change is qualitatively different. Novices' *raw adjustments* in Round 1 are positive (increasing the odds assigned to conjoint probabilities that are already overstated). In Round 4, after receiving performance feedback, novices essentially stop increasing those odds, but, they do not significantly decrease the odds: their *raw adjustments* for Acceptable Urns in Round 4 are not statistically different from zero (-0.82 and 0.08).

On the other hand, linear contrasts indicate that novices' *changes in raw adjustments* in the Reject Urns frame are statistically different from experienced participants in the Control Risk frame. When conjoint probabilities are understated, novices' *changes in raw adjustments* are small and statistically insignificant (1.13 and -0.04). In other words, novice participants in the Reject Urn frame do not mitigate the conservatism bias in any fashion. They do not improve their assessment of initial individual conditional probabilities nor do they increase the odds assigned to the conjoint probabilities derived from

[9] We test for two order effects. For *round order*, half of the participants complete "Round 2" cases prior to "Round 3", and vice versa. For *case order*, participants start each round on one of 12 different cases. We examine potential order effects on *raw adjustments* and *change in raw adjustments* in Round 4. The ANOVA results indicate that neither *round order* or *case order*, or any of their interactions with other treatment variables, are significant.

those understated conditional probabilities. In general, the results support H3. Experienced participants, compared to novices, adjust conjoint probabilities in the appropriate direction and increase the magnitude of those adjustments after receiving performance feedback in the form of conjoint probability scores.

***Hypotheses 4:*** **Comparable Conjoint Probabilities and Hypothesis Framing Effects**

The final issue is whether hypothesis frame affects **comparable** conjoint probabilities. As previously discussed, in the computer-based experiment, participants adjust odds assigned to two internal control combinations that differ regarding *in focus* versus *out of focus*. H4 predicts that after receiving performance feedback in the form of conjoint probability scores, experienced participants are more likely than novices to submit conjoint probabilities for comparable outcomes that do not differ by hypothesis frame. Alternatively, we expect novice participants to demonstrate systematic differences between hypothesis frames for comparable internal control combinations. Table 5 presents cell means and ANOVA statistics for both *initial distances* and *final distances* in Round 4.

<Table 5>

The cell means and linear contrasts presented in Panel A of Table 5 indicate that in Round 4 both experienced participants and novices provide *initial distances* that do not vary by focus type (*in focus* versus *out of focus*). In fact, linear contrasts indicate that *initial distances* are similar for *in focus* versus *out of focus* internal control combinations, regardless of participant type. For example, in the experienced participant group, the *initial distances* for the internal control combination *no control deficiencies* does not differ significantly between focus type: 12.79 and 14.04, for *in focus* and *out of focus*, respectively; in the novice group, the comparable *initial distances* are also statistically similar: 15.65 and 13.26, for *in focus* and *out of focus*, respectively.

However, *final distances* differ substantially by participant type, focus type and hypothesis frame, as evidenced by the statistically significant ($p < .001$) three-way interaction term reported in Panel B of Table 5. Linear contrasts indicate that experienced participants provide similar *final distances* for the internal control combination *no control deficiencies* whether the control combination is *in focus* versus *out of focus* (8.87 versus 8.36, respectively). Similarly, for the internal control combination *material weakness*, experienced participants submit similar *final distances* whether the control combination is *in focus* versus *out of focus (*-15.93 versus -15.91, respectively). In other words, experienced participants in different hypothesis frames provide similar conjoint probabilities for specific internal control combinations, regardless of whether the control combinations are *in focus* or *out of focus*. These results indicate that experienced participants are able to adjust conjoint probabilities in the appropriate direction and magnitude, regardless of framing: whether the internal control combination is *in focus* or *out of focus*.

The results are very different for novices. For the *two acceptable urns* combination, novices provide *final distances* that are statistically different ($p < .01$) across hypothesis frames: 15.29 versus 10.87, for the Acceptable Urn frame versus Reject Urn frame, respectively. Novices also provide statistically different ($p < .05$) *final distances* for the *two reject urns* combination, -19.77 versus -16.59, for Acceptable Urn frame versus Reject Urn frame, respectively. The differing performance indicates framing matters for novices: focusing on "Acceptable Urns" led to conjoint probabilities that have significantly greater *final distances* than focusing on "Reject Urns". Alternatively, experienced participants appear to demonstrate genuine critical thinking: adjusting the conjoint probabilities for comparable internal control combinations in a similar manner regardless of whether the hypothesis frame places the control combination *in focus* or *out of focus*. The pattern of results supports H4.

## CONCLUSION

In this study, we extend audit judgment research by shifting the focus from biased audit judgment to auditors' ability to think critically: the ability to reason logically when logic conflicts with prior beliefs (West et al. 2008). Specifically, we examine whether audit professionals are capable of detecting biases in their initial conditional probability judgments and overriding those biases with conjoint probability adjustments that more closely approximate normative outcomes. We find that, prior to receiving performance feedback, experienced participants adjust conjoint probabilities associated with internal control combinations in a manner that compensates for the conservatism bias in their initial assessments and avoids the focalism bias. Furthermore, after receiving performance feedback, these participants are able to reduce the conservatism bias in their initial assessments and avoid a framing effects bias when adjusting conjoint probabilities derived from those probabilities. We argue that the ability to override biased judgments and avoid introducing additional bias is indicative of critical thinking and warrants further research in the accounting and auditing literatures.[10]

In the current study, we provide participants with two opportunities to compensate for the conservatism bias or override the bias directly. Each of these opportunities requires participants to demonstrate critical thinking when evaluating conjoint probabilities for internal control combinations. First, in all four rounds of the experiment, participants have the opportunity to compensate for the bias by adjusting conjoint probabilities for internal control combinations derived from conservatively assessed individual internal controls. We find that in Round 1, **prior to receiving performance feedback**, experienced participants provide rational adjustments to conjoint probabilities for internal control combinations by increasing (decreasing) odds assigned to understated (overstated) conjoint probabilities. On the other hand, in Round 1, novices consistently generate positive adjustments for conjoint probabilities regardless of whether their initial assessments are overstated or understated (which is evidence of a focalism bias). We suggest that novices' responses reflect the use of a simple heuristic: increase the odds assigned to conjoint probabilities, especially when sample evidence indicates that both events are likely to occur. Use of such a heuristic is the antithesis of critical thinking.

The second opportunity to inhibit or override the conservatism bias takes place after participants receive performance feedback in the form of conjoint probability scores. Potentially, all participants could utilize this feedback to avoid conservative assessment of individual conditional probabilities; however, overriding the basic conservatism bias in initial assessments is challenging from a critical thinking perspective because performance feedback is directly tied to conjoint probabilities and indirectly associated with individual conditional probabilities. Given prior research documenting the robustness of the conservatism bias in probabilistic reasoning, we did not expect participants to modify their assessment of individual initial conditional probabilities. However, across four different treatment cells (two levels of hypothesis frame and two levels of sample size), after receiving performance feedback, experienced participants provide assessments of individual conditional probabilities that are significantly less conservative than originally assessed in Round 1. We emphasize that the performance feedback in this study is not the personalized, intensive feedback provided in prior studies examining debiasing of

[10] A series of papers by Stanovich and West (1997, 1998, 2000, 2007) provide a framework for such an investigation of the nature of rational thought.

probabilistic judgments (Bazerman 2006; Stone and Opel 2000; Lichtenstein and Fischhoff's 1980). Rather, the feedback is simple conjoint probability scores that participants can utilize, or not utilize, in evaluating the quality of their initial beliefs regarding individual internal controls. Experienced participants demonstrate an ability to critically consider the feedback and modify their initial beliefs regarding the likelihood that individual internal controls are significantly deficient. On the other hand, novices do not modify the conservatism bias in their initial assessments.

Also, after receiving performance feedback, experienced participants increase the magnitude of their adjustments to conjoint probabilities for internal control combinations, thus further compensating for some of the conservatism bias that remains in individual internal control assessments. These adjustments to conjoint probabilities continue to be in the appropriate direction, increasing (decreasing) understated (overstated) conjoint probabilities. Thus, by **both inhibiting** the conservatism bias in initial assessments and **overriding** the remaining bias in conjoint probabilities derived from those initial assessments, experienced participants, compared to novices, provide final conjoint probabilities for internal control combinations that are significantly closer to normative conjoint probabilities.

It is important to emphasize that this inhibiting and overriding of the conservatism bias occurs when performance feedback is no longer available (the final round of the experiment). Experienced participants in this study exercise judgments that appear to detect a need to override biased responses, and, they sustain overriding biased responses at a later point in time. Furthermore, we find that experienced participants, regardless of hypothesis frame, submit final conjoint probabilities in the final round that do not differ by focus type: whether the internal control combination is *in focus* versus *out of focus*. These results are important because they emphasize that, in addition to inhibiting and overriding the conservatism bias noted above, experienced participants also inhibit and override a potential bias attributed to hypothesis framing. In other words, experienced participants demonstrate multiple instances in which they inhibit and override judgment bias, generating assessments that are more rational and reflect critical thinking. However, this behavior does not generalize to novices; novice participants' assessments remain much more biased along the three dimensions that experienced participants avoid or override: conservatism, focalism, and hypothesis framing.

The results from this study have specific implications for audit practice. Based on prior audit judgment research, one would logically deduce that auditors' biased assessment of individual internal control reliability leads to excessive conservatism in the assessment of internal control combinations. This excessive conservatism would be attributed to a multiplicative effect arising from considering combinations of internal controls. The penalty for excessively understating internal control reliability is either audit inefficiency (unnecessarily increasing substantive testing) and/or opining an inappropriately high likelihood for material weakness in the internal control system over financial reporting. However, based on the results from the current study, we suggest that audit professionals are capable of compensating for conservative assessment of individual control reliability by adjusting their assessment of the reliability of internal control combinations. Using the framework provided by Stanovich and colleagues, we argue that auditing professionals' detection of bias in initial assessments (or, at a minimum, appearing to consistently detect bias in initial assessments) demonstrates critical thinking. Perhaps a key component for facilitating critical thinking in the auditing profession is to provide individuals with the opportunity to "be reflective." For example, the results in both our study and a recent audit study by Ganguly and Hammersley (2009) indicate that the opportunity to reflect on initial estimates and provide revisions to those estimates is critical for overriding bias in initial judgments. In a recent paper, Stanovich (2009) outlines the relative importance of the "reflective mind" in inhibiting heuristic thinking and biased judgments from the "autonomous mind" and facilitating activation of analytical and more reasoned responses from the "algorithmic mind." While prior audit research has extensively studied heuristic thinking and biased judgment, relatively little is known regarding the role of the reflective mind in enabling critical thinking. Future audit research should explore the theoretical role of the reflective mind and its impact on critical thinking and the practice of auditing.

Our study is subject to several limitations. First and most notably, we utilize graduate students as proxies for practicing auditors. Nonetheless, these participants have some audit experience and have completed advanced courses emphasizing critical thinking. Comparing their judgments to novices (undergraduate accounting students with less training and experience) suggests that we would expect practicing auditors to demonstrate even higher levels of critical thinking and less biased responses. Secondly, the novice participants perform the experiment in a context free setting, which was necessary to avoid confusion over unfamiliar internal control terminology. As such, we are unable to determine the full extent to which observed differences between participant groups are attributable to experience or context. Third, the current analysis is also limited to audit conditions in which participants receive signals that internal control deficiencies are likely. When participants receive signals that internal control deficiencies are unlikely, we find that most participants (experienced participants as well as novices) utilize sample proportions as an approximation of individual conditional probabilities. This heuristic is successful because low defect rates in such samples happen to approximate Bayesian conditional probabilities. Thus, we are unable to assert that the critical thinking we observe when internal control deficiencies are likely would hold when such deficiencies are unlikely. Lastly, we note that probabilistic reasoning remains less than normative. Although experienced participants significantly improve their probability assessments given performance feedback and the opportunity to revise conjoint probabilities mathematically-derived from conservatively assessed individual conditional probabilities, nonetheless, their final assessments remain significantly different from normative theory (Bayesian-based conjoint probabilities).

## REFERENCES


American Institute of Certified Public Accountants (AICPA). 2006. *Trust Services Principles, Criteria and Illustrations for Security, Availability, Processing Integrity, Confidentiality, and Privacy (Including WebTrust® and SysTrust®).* New York, NY: AICPA.

_____. 2008. *Audit Risk and Materiality in Conducting an Audit*. Statements on Auditing Standards No. 107. New York, NY: AICPA.

_____. 2009. *CPA Vision Project.* http://www.aicpa.org/vision

Bazerman, M. H. 2006. *Judgment in managerial decision making* (6th ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Bonner, S. E. 2008. *Judgment and Decision Making in Accounting*. Upper Saddle River, NJ: Pearson Prentice Hall.

Brenner, L., D. Griffin, and D. J. Koehler. 2005. Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes, 97,* 64 – 81.

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

Davelaar, E. J., H. J. Haarmann, Y. Goshen-Gottstein, and M. Usher. 2006. Semantic similarity dissociates short- from long-term recency effects: Testing a neurocomputational model of list memory. *Memory & Cognition, 34*, 323 – 334.

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, 139 – 168.

Emby, C. 1994. Framing and presentation mode effects in professional judgment: Auditors' internal control judgments and substantive testing decisions. *Auditing: A Journal of Practice and Theory, 13,* 102 – 115,

Emby, C. and D. Finley. 1997. Debiasing framing effects in auditor's internal control judgments and testing decisions. *Contemporary Accounting Research, 14,* 55 – 77.

Ganguly, A. R. and J. S. Hammersley. 2009. Covariation assessments with costly information collection in audit planning: An experimental study. *Auditing: A Journal of Practice and Theory, 28*, 1 – 27.

Griffin, D., and A.Tversky. 2003. The weighing of evidence and the determinants of confidence. In T. Gilovich, D. Griffin and D. Kahneman (Eds.), *Heuristics and biases* (pp. 230 – 249). Cambridge: Cambridge University Press.

Holt, D. 1987. Auditors and base rates revisited. *Accounting, Organizations and Society, 6,* 571 – 578.

Joyce, E., and G. C. Biddle. 1981. Are auditiors' judgments sufficiently regressive? *Journal of Accounting Research, 19,* 323 – 349.

Kinney, W. R., Jr. and W. C. Uecker. 1982. Mitigating the consequences of anchoring in auditor judgments. *The Accounting Review, 57,* 55 – 69.

Larrick, R. P. 2004. Debiasing. In D. J. Koehler and N. Harvey (Eds.), *Blackwell Handbook of Judgment and Decision Making*. Oxford, England: Blackwell.

Lichtenstein, S., and B. Fischhoff. 1980. Training for calibration. *Organizational Behavior and Human Performance, 26,* 149 – 171.

Markovits, H., and G. Nantel. 1989. The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition, 17,* 11 – 17.

Messier, W. F., Jr., S. M. Glover, and D. F. Prawitt. 2008. Auditing & Assurance Services: A Systematic Approach (6th Ed.). New York, NY: McGraw-Hill Irwin.

Mulligan, N. W., M. Duke, and A. W. Cooper. 2007. The effects of divided attention on auditory priming. *Memory & Cognition, 35*, 1245 – 1254.

Nelson, M. K. 1995. Strategies of auditors: Evaluation of sample results. *Auditing: A Journal of Practice & Theory* (Spring): 34 – 49.

Public Company Accounting Oversight Board (PCAOB). 2007. *An Audit of Internal Control Over Financial Reporting that is Integrated with an Audit of Financial Statements*. Auditing Standard No. 5. Washington, D.C.: PCAOB.

Romney, M. B. and P. J. Steinbart. 2009. *Accounting Information Systems* (11th Ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Schkade, D. A., and D. Kahneman. 1998. Does living in california make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science, 9,* 340 – 346.

Springer, C. W., and A. F. Borthick. 2007. Improving performance in accounting: Evidence for insisting on cognitive conflict tasks. *Issues in Accounting Education, 22,* 1 – 19.

Stanovich, K. E. 2009. Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. St. B. T. Evans and K. Frankish (Eds.), *In Two Minds* (pp. 55 – 88). Oxford: Oxford University Press.

Stanovich, K. E., and R. F. West. 1997. Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89,* 342 – 357.

_____ and _____. 1998. Individual differences in rational thought. *Journal of Experimental Psychology: General, 127,* 161 – 188.

_____ and _____. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23,* 645 – 726.

_____ and _____. 2007. Natural myside bias is independent of cognitive ability. *Thinking & Reasoning, 13,* 225 – 247.

_____ and _____. 2008. On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94,* 672 – 695.

Stone, E. R., and R. B. Opel. 2000. Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes, 83,* 282 – 309.

West, R. F., M. E. Toplak, and K. E. Stanovich. 2008. Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology* 100: 930 - 941.

Wilson, T. D., T. Wheatley, J. M. Meyers, D. T. Gilbert, and D. Axsom. 2000. Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 78,* 821 – 836.

Wyhe, G. V. 2007. A history of U. S. higher education in accounting, Part II: Reforming accounting within the academy. *Issues in Accounting Education, 22,* 481 – 501.

**Figure 1: Assessment of Subjective Conditional Probabilities**

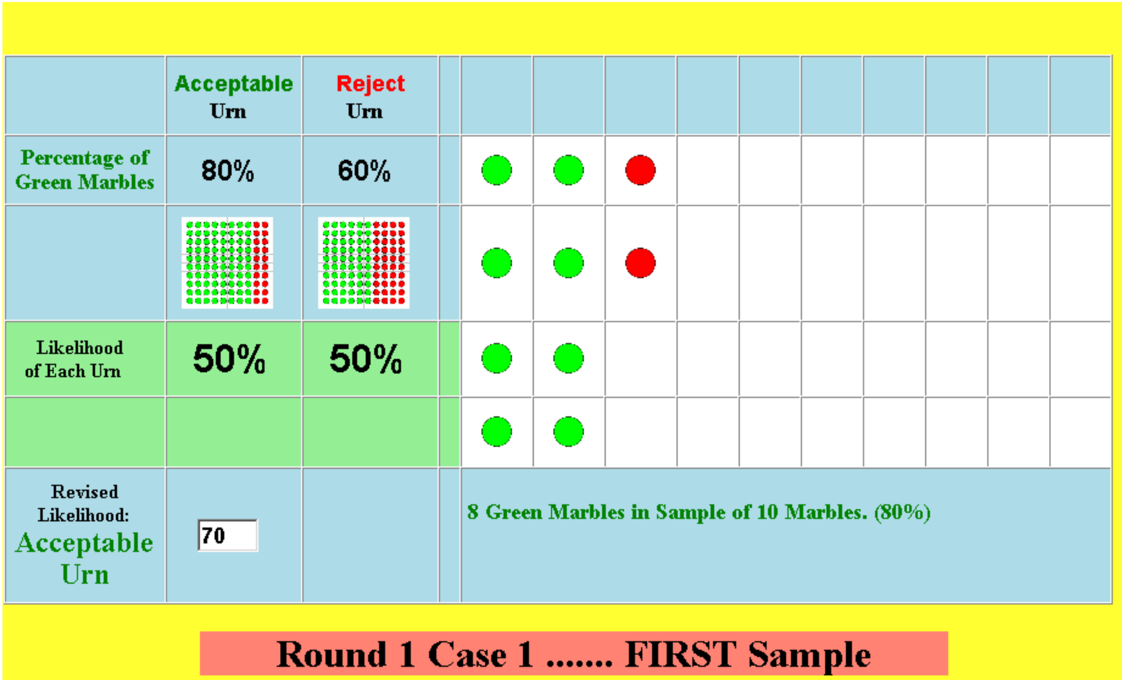

**Figure 2: Adjustment of Conjoint Probabilities**

| | | |
|---|---|---|
| Acceptable Urns | | |
| FIRST Sample | 70% | |
| SECOND Sample | 80% | |
| Possible Outcomes | Overall Likelihood | Best Estimate |
| Two Acceptable Urns (NO Control Deficiency) | 56% | % |
| One Acceptable Urn One Reject Urn (Significant Deficiency) | 38% | % |
| Two Reject Urns (Material Weakness) | 6% | 0 % |
| | 100% | 0 % |
| | | Click to Submit |

**Figure 3: Performance Feedback via Conjoint Probability Scores**

**Round # 2: Case # 1**

| Possible Outcomes | Your Estimate | Math Estimate | Over / Under | | |
|---|---|---|---|---|---|
| Two Acceptable Urns (NO Control Deficiency) | 55% | 88% | -33 | | |
| One Acceptable Urn One Reject Urn (Significant Deficiency) | 45% | 12% | 33 | | |
| Two Reject Urns (Material Weakness) | 0% | 0% | 0 | Total Points Earned | 134 |

CLICK HERE FOR NEXT CASE

Table 1
Unique and Independent Samples

| | | Sample ID | Sample Size | Number of Green Marbles | Number of Red Marbles | Focus on Reliability (Acceptable Urn) Sample Proportion | Focus on Reliability (Acceptable Urn) Bayesian Odds | Focus on Risk (Reject Urn) Sample Proportion | Focus on Risk (Reject Urn) Bayesian Odds |
|---|---|---|---|---|---|---|---|---|---|
| Samples Representative of **Reject Urn** | Small Samples | 1 | 10 | 5 | 5 | 50% | 11.6% | 50% | 88.4% |
| | | 2 | 10 | 6 | 4 | 60% | 26.0% | 40% | 74.0% |
| | | 3 | 15 | 9 | 6 | 60% | 17.2% | 40% | 82.8% |
| | | 4 | 15 | 10 | 5 | 67% | 35.7% | 33% | 64.3% |
| | | 5 | 20 | 12 | 8 | 60% | 10.9% | 40% | 89.1% |
| | | 6 | 20 | 13 | 7 | 65% | 24.7% | 35% | 75.3% |
| | Large Samples | 7 | 30 | 19 | 11 | 63% | 10.3% | 37% | 89.7% |
| | | 8 | 30 | 20 | 10 | 67% | 23.5% | 33% | 76.5% |
| | | 9 | 35 | 23 | 12 | 66% | 15.4% | 34% | 84.6% |
| | | 10 | 35 | 24 | 11 | 69% | 32.7% | 31% | 67.3% |
| | | 11 | 40 | 26 | 14 | 65% | 9.7% | 35% | 90.3% |
| | | 12 | 40 | 27 | 13 | 67% | 22.3% | 33% | 77.7% |
| Samples Representative of **Acceptable Urn** | Small Samples | 13 | 10 | 8 | 2 | 80% | 71.4% | 20% | 28.6% |
| | | 14 | 10 | 9 | 1 | 90% | 86.9% | 10% | 13.1% |
| | | 15 | 15 | 11 | 4 | 73% | 59.7% | 27% | 40.3% |
| | | 16 | 15 | 12 | 3 | 80% | 79.8% | 20% | 20.2% |
| | | 17 | 20 | 15 | 5 | 75% | 70.1% | 25% | 29.9% |
| | | 18 | 20 | 16 | 4 | 80% | 86.2% | 20% | 13.8% |
| | Large Samples | 19 | 30 | 22 | 8 | 73% | 68.7% | 27% | 31.3% |
| | | 20 | 30 | 23 | 7 | 77% | 85.4% | 23% | 14.6% |
| | | 21 | 35 | 26 | 9 | 74% | 77.6% | 26% | 22.4% |
| | | 22 | 35 | 27 | 8 | 77% | 90.2% | 23% | 9.8% |
| | | 23 | 40 | 29 | 11 | 73% | 67.2% | 28% | 32.8% |
| | | 24 | 40 | 30 | 10 | 75% | 84.6% | 25% | 15.4% |

Table 2
Cases Derived by Combining Samples

| | | | | Conjoint Probability | |
|---|---|---|---|---|---|
| Round | Case | Sample A ID | Sample B ID | Focus on Reliability (2 Acceptable Urns) | Focus on Risk (2 Reject Urns) |
| 1 and 4 | 1 | 13 | 18 | 61.6% | 3.9% |
| 1 and 4 | 2 | 20 | 24 | 72.2% | 2.3% |
| 1 and 4 | 3 | 19 | 21 | 53.3% | 7.0% |
| 1 and 4 | 4 | 14 | 16 | 69.4% | 2.6% |
| 1 and 4 | 5 | 7 | 10 | 3.4% | 60.4% |
| 1 and 4 | 6 | 22 | 23 | 60.7% | 3.2% |
| 1 and 4 | 7 | 2 | 5 | 2.8% | 65.9% |
| 1 and 4 | 8 | 1 | 4 | 4.1% | 56.8% |
| 1 and 4 | 9 | 9 | 12 | 3.4% | 65.8% |
| 1 and 4 | 10 | 15 | 17 | 41.8% | 12.1% |
| 1 and 4 | 11 | 8 | 11 | 2.3% | 69.1% |
| 1 and 4 | 12 | 3 | 6 | 4.3% | 62.3% |
| 2 | 1 | 4 | 5 | 3.9% | 57.3% |
| 2 | 2 | 21 | 24 | 65.6% | 3.5% |
| 2 | 3 | 8 | 12 | 5.2% | 59.4% |
| 2 | 4 | 1 | 3 | 2.0% | 73.2% |
| 2 | 5 | 10 | 11 | 3.2% | 60.8% |
| 2 | 6 | 19 | 22 | 62.0% | 3.1% |
| 2 | 7 | 7 | 9 | 1.6% | 75.9% |
| 2 | 8 | 14 | 15 | 51.9% | 5.3% |
| 2 | 9 | 2 | 6 | 6.4% | 55.7% |
| 2 | 10 | 13 | 17 | 50.0% | 8.6% |
| 2 | 11 | 20 | 23 | 57.4% | 4.8% |
| 2 | 12 | 16 | 18 | 68.8% | 2.8% |
| 3 | 1 | 14 | 17 | 60.9% | 3.9% |
| 3 | 2 | 19 | 23 | 46.2% | 10.3% |
| 3 | 3 | 1 | 6 | 2.9% | 66.5% |
| 3 | 4 | 8 | 10 | 7.7% | 51.5% |
| 3 | 5 | 7 | 12 | 2.3% | 69.7% |
| 3 | 6 | 22 | 24 | 76.3% | 1.5% |
| 3 | 7 | 9 | 11 | 1.5% | 76.4% |
| 3 | 8 | 15 | 18 | 51.4% | 5.6% |
| 3 | 9 | 13 | 16 | 57.0% | 5.8% |
| 3 | 10 | 3 | 5 | 1.9% | 73.7% |
| 3 | 11 | 2 | 4 | 9.3% | 47.6% |
| 3 | 12 | 20 | 21 | 66.3% | 3.3% |

Table 3
Round 1 Initial Distance and Raw Adjustments for Subjective Conjoint Probabilities

Panel A: Cell Means

| Predictor Variables | Round 1 Initial Distance[1] *M* | Round 1 Raw Adjustments[2] *M* | Linear Contrasts[3] | t test ≠ 0 *p*[4] |
|---|---|---|---|---|
| Participants with Audit Experience | | | | |
| Frame: Control Reliability | 16.63 | -1.17 | a | 0.006 |
| Frame: Control Risk | -32.19 | 1.93 | b | 0.005 |
| Novices | | | | |
| Frame: Acceptable Urns | 17.86 | 2.82 | b | 0.001 |
| Frame: Reject Urns | -23.41 | 4.99 | c | 0.001 |

[1] *Initial Distance* is the difference between conjoint probabilities mathematically-derived from subjective conditional probabilities and those mathematically-derived from Bayesian conditional probabilities.

[2] *Raw adjustment* is the difference between participants' "best estimate" of conjoint probability minus the conjoint probability mathematically-derived from participants' two subjective conditional probabilities.

[3] Different letters for linear contrasts across treatment cells indicate statistically different means ($p < 0.01$), such that a ≠ b, etc.

[4] Two-tailed tests.

Panel B: ANOVA Statistics for Raw Adjustments

| Predictor Variables | Round 1 Raw Adjustments *F* | *p* |
|---|---|---|
| Between-Subjects Variables | | |
| Experience Level | 11.44 | 0.001 |
| Frame | 6.40 | 0.013 |
| Experience Level * Frame | 0.20 | 0.658 |

Table 4

Initial Distances and Raw Adjustments for Subjective Conjoint Probabilities : Round 4 versus Round 1

Panel A: Cell Means

| Predictor Variables | Initial Distance[1] Round 1 *M* | Round 4 *M* | Change in Initial Distance[3] *M* | LC[4] | Raw Adjustments[2] Round 1 *M* | LC[4] | Round 4 *M* | LC[4] | Change in Raw Adjustment[3] *M* | LC[4] |
|---|---|---|---|---|---|---|---|---|---|---|
| Participants with Audit Experience | | | | | | | | | | |
| Frame: Control Reliability | | | | | | | | | | |
| Large sample size | 20.57 | 16.81 | -3.76** | a | -1.83** | a | -4.77*** | a | -2.95** | a |
| Small sample size | 12.68 | 8.79 | -3.89*** | a | -0.51 | a | -3.07*** | a | -2.56** | a |
| Frame: Control Risk | | | | | | | | | | |
| Large sample size | -36.77 | -30.17 | 6.60*** | b | 0.22 | a | 8.08*** | b | 7.86*** | b |
| Small sample size | -27.82 | -19.12 | 8.70*** | b | 3.64*** | b | 9.35*** | b | 5.71** | b |
| Novices | | | | | | | | | | |
| Frame: Acceptable Urns | | | | | | | | | | |
| Large sample size | 21.51 | 18.18 | -3.33* | c | 2.92** | c | -0.82 | c | -3.74** | a |
| Small sample size | 14.21 | 13.13 | -1.08 | c | 2.71*** | c | 0.08 | c | -2.63* | a |
| Frame: Reject Urns | | | | | | | | | | |
| Large sample size | -28.59 | -28.41 | 0.18 | c | 5.08*** | d | 6.21*** | d | 1.13 | c |
| Small sample size | -18.23 | -15.85 | 2.38 | c | 4.89*** | d | 4.85*** | d | -.04 | c |

[1] *Initial Distance* is the difference between conjoint probabilities mathematically-derived from subjective conditional probabilities and those mathematically-derived from Bayesian conditional probabilities.

[2] *Raw adjustment* is the difference between participants' "best estimate" of conjoint probability minus the conjoint probability mathematically-derived from participants' two subjective conditional probabilities

[3] Differences reflect Round 4 values minus Round 1 values. Negative changes (positive changes) move closer to Bayesian for Control Reliability / Acceptable Urns cases (Control Risk / Reject Urns cases).

[4] Separate Linear Contrasts (LC) are performed on *Change in Initial Distance*, *Raw Adjustments* Round 1, and *Raw Adjustments* Round 4, and *Change in Raw Adjustment*. Different letters for linear contrasts indicate statistically different means ($p < 0.01$), such that a ≠ b, etc.

*, **, *** Significant t-test ≠ 0 at the 5%, 1%, and .1% levels, respectively, two-tailed test.

Panel B: ANOVA Statistics

| Predictor Variables | Change in Initial Distance | | Raw Adjustments | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Round 1 | | Round 4 | | Change in Raw Adjustment | |
| | *F* | *p* | *F* | *p* | *F* | *p* | *F* | *p* |
| Between-Subjects Variables | | | | | | | | |
| Experience Level | 5.18 | 0.025 | 11.44 | 0.001 | 0.02 | 0.897 | 3.76 | 0.055 |
| Frame | 0.68 | 0.412 | 6.40 | 0.013 | 41.09 | 0.001 | 0.22 | 0.644 |
| Experience Level * Frame | 1.82 | 0.181 | 0.20 | 0.656 | 5.42 | 0.022 | 4.97 | 0.020 |
| Within-Subjects Variables | | | | | | | | |
| Sample Size | 0.30 | 0.585 | 4.22 | 0.041 | 1.44 | 0.230 | 2.63 | 0.105 |
| Sample Size * Experience Level | 0.33 | 0.568 | 5.86 | 0.016 | 2.67 | 0.103 | 0.01 | 0.930 |
| Sample Size * Frame | 2.60 | 0.108 | 1.00 | 0.318 | 1.66 | 0.198 | 0.38 | 0.537 |
| Sample Size * Experience Level * Frame | 0.38 | 0.539 | 0.97 | 0.325 | 0.76 | 0.384 | 0.33 | 0.567 |

Table 5

Round 4 Conjoint Probability Distances: "In Focus" versus "Out of Focus" Conjunctive Sets

Panel A: Cell Means

| | Round 4 | | | |
|---|---|---|---|---|
| | Initial [1] Distance | | Final [2] Distance | |
| Predictor Variables | *M* | | *M* | |
| Participants with Audit Experience | | | | |
| Frame: Control Reliability | | | | |
| In Focus: No Control Deficiencies | 12.79 | a | 8.87 | a |
| Out of Focus: Material Weakness | -22.74 | b | -15.91 | b |
| Frame: Control Risk | | | | |
| In Focus: Material Weakness | -24.64 | b | -15.93 | b |
| Out of Focus: No Control Deficiencies | 14.04 | a | 8.36 | a |
| Participants Novices | | | | |
| Frame: Acceptable Urns | | | | |
| In Focus: 2 Acceptable Urns | 15.65 | a | 15.29 | c |
| Out of Focus: 2 Reject Urns | -25.08 | b | -19.77 | d |
| Frame: Reject Urns | | | | |
| In Focus: 2 Reject Urns | -22.13 | b | -16.59 | e |
| Out of Focus: 2 Acceptable Urns | 13.26 | a | 10.87 | f |

[1] *Initial Distance* is the difference between conjoint probabilities mathematically-derived from subjective conditional probabilities and those mathematically-derived from Bayesian conditional probabilities.

[2] *Final Distance is* the difference between participants' "best estimate" of conjoint probabilities and Bayesian-derived conditional probabilities.

Note: Linear contrasts are performed on three variables: *Change in Initial Distance*, *Raw Adjustments* Round 1, and *Raw Adjustments* Round 4.

Note: Different letters for Linear Contrasts indicate statistically different means ($p < 0.01$), except for *final distance* d ≠ e ($p < 0.05$).

Panel B: ANOVA Statistics

| Predictor Variables | Initial Distance F Value | *p* | Final Distance F Value | *p* |
|---|---|---|---|---|
| Between-Subjects Variables | | | | |
| Experience Level | 0.91 | 0.344 | 2.40 | 0.124 |
| Frame | 0.00 | 0.969 | 0.39 | 0.535 |
| Experience Level * Frame | 0.26 | 0.608 | 0.06 | 0.805 |
| Within-Subjects Variables | | | | |
| Focus Type (In Focus v. Out of Focus) | 0.31 | 0.581 | 4.22 | 0.040 |
| Focus Type * Experience Level | 4.54 | 0.033 | 3.27 | 0.071 |
| Focus Type * Frame | 1420.00 | 0.001 | 805.56 | 0.001 |
| Focus Type * Experience Level * Frame | 0.23 | 0.632 | 11.70 | 0.001 |

## APPENDIX

### Instructional Materials

In the company you are auditing, sales people are allowed to modify sales contract terms. The following two controls detect and account for any influence these modifications have on revenue recognition.

- 1. The accounting function reviews changes to contract terms

- 2. The accounting function reviews changes to shipping terms.

1. If the Sample Evidence suggests that BOTH Controls are Rejected, then the Internal Control System has a Material Weakness. For example…….

   If BOTH controls are REJECTED (i.e., fail to work), a material misstatement could result in the financial statements. Thus, BOTH controls failing indicates a MATERIAL WEAKNESS in internal controls.

2. If the Sample Evidence suggests that ONE Control (but not both) is Rejected, then the Internal Control System has a Significant Deficiency. For example…….

   If ONE of the controls is REJECTED and the other is ACCEPTABLE, this could cause a significant, but not material, misstatement in the financial statements. Thus, ONE control failing, but NOT both, indicates the internal control system has a SIGNIFICANT DEFICIENCY.

3. If the Sample Evidence suggests that BOTH Controls are Acceptable, then the Internal Control System has NO Control Deficiency. For example……

   If BOTH controls are ACCEPTABLE, the internal control system has NO CONTROL DEFICIENCY.

In this game, each Internal Control is represented by an URN filled with marbles.

Red marbles indicate when the control is NOT working,

Green marbles indicate when the control is WORKING.

We use URNS filled with marbles to indicate REJECT controls and ACCEPTABLE controls

A Reject Urn has 40% Red Marbles and 60% Green Marbles.

The Control fails 40% of the time.

The control is NOT functioning at an acceptable level.

An Acceptable Urn has 20% Red Marbles and 80% Green Marbles.

The Control fails only 20% of the time.

The control is functioning at an acceptable level.