The Effects of Task Complexity and Skill on Over/Under Estimation of Internal Control

Maureen Francis Mascha
Marquette University

Cathleen L. Miller
Wayne State University

Legislation and audit standards require auditors to assess internal control environments to reasonably assure the controls detect/prevent material misstatements. Incorrect assessments can lead to too much or too little audit testing, resulting in inefficient or ineffective audits. The *direction* of an incorrect assessment is equally important to its size.

We examine whether task complexity leads to over/under control risk assessments. Using Bonner's (1994) model, auditor skill, and audit work programs, we examine the relationship between task complexity and control risk assessment.

Using a laboratory experiment with senior accounting students enrolled in auditing at three Midwest universities, we find task complexity and skill level affect control risk assessments – high and low skilled auditors assess control risk too high when performing several, simple risk assessments. High and low skilled auditors improve risk assessments when performing several, complex risk assessments. We find work programs do not significantly affect control risk assessments.

Key Words: Internal Control, Task Complexity, Skill, Control Risk Assessment.

The Effects of Task Complexity and Skill on Over/Under-estimation of Internal Control

I. INTRODUCTION

Auditing has gone from a self-regulated profession to a regulated industry. This change resulted from several high-profile audit failures (e.g. Enron, WorldCom, and Tyco) that led to passage of the Sarbanes-Oxley Act of 2002 (Act). The Act mandates public oversight of the entire accounting and financial reporting process, with significant penalties for both management and auditors alike who fail to comply with the Act. For the first time, certified public accountants engaged in the audit of public companies (e.g. SEC listed) answer to an external, third party oversight board (PCAOB) that in turn answers to Congress. The PCAOB, charged with responsibility for overseeing the entire audit process from issuance of professional audit standards to regulation and review of audit firms, has broad powers, including the ability to levy fines and other penalties for non-compliance.

The standard setting function of the PCABO assumes great importance for CPAs engaged in the practice of auditing SEC clients since the PCAOB is required to audit the auditors and to report its findings to Congress in annual reports. Coupled with the Act's specific requirement that auditors review and assess a client's internal control environment to reasonably assure itself that the internal controls would detect and/or prevent material misstatements in the financial records, the Act increases both the responsibility and scope of auditors. Failure to detect a material error has assumed greater consequences for auditors, since the possibility of major penalties for noncompliant audit firms range from expanded civil liability to outright criminal prosecution. Although Audit Standard No. 5 (AS #5) provides auditors some relief in allowing risk-based substantive testing, it does not remove the requirement for internal control review nor lessen its impact on subsequent fieldwork. Clearly, assessment of internal control remains a pivotal point in the audit process, with increased risk for erroneous assessment.

Erroneous assessment consists of incorrectly evaluating the internal control environment. Such an error can lead to auditors overlooking serious issues that point to fraudulent misstatements and/or errors in addition to affecting the scope and nature of substantive testing. However, the erroneous assessment itself is not the only issue; instead, the *direction* of the error is equally, if not more important. For example, an auditor estimating internal control to be "average" might assign a score of 50 out of 100 to indicate client level of internal control. If the accurate assessment is 70/100 then an error of 20 points occurred (e.g. 70-50). However, a score of 50/100 is no less accurate than an auditor who assesses the same internal control environment to be 90/100 since both misjudged internal control by 20 points. The only difference between the two is that the first auditor underestimated internal control while the second auditor overestimated it. Borrowing from statistics, underestimates of internal control represent a Type I error since the auditor fails to accept the null hypothesis that internal control is adequate while overestimates are Type II errors since the auditor accepts the null (controls are adequate) when in fact they are not (Howell 1997).

Thus, equal errors in terms of score differences result in vastly different consequences. This issue is especially salient given the different risks and costs associated with each type of error. Type I errors result in over testing and potentially non-compensated audit work; while Type II errors can result in audit failure because of over reliance on internal control leading to reduced testing, ultimately resulting in failure to detect material misstatements. Even if overestimation of internal control does not lead to outright audit failure, it can lead to future restatements of client financial data, an embarrassing event for a profession seeking to regain investor confidence. Since costs are associated with either error, most firms are willing to take steps to prevent either from occurring. Therefore, understanding what causes over/under estimation is important for practicing auditors as well as academics educating the next generation of auditors.

This study examines one potential factor that leads auditors to over/under estimate internal control evaluation - task complexity. Prior research clearly finds that task complexity affects judgment decisions, including judgment accuracy (e.g. Anderson 1983; Bonner and Pennington 1990; Bonner 1994; Abdolmohammadi and Wright, 1987). Bonner (1994) develops a model for the relationship between task complexity, skill, motivation, and judgment accuracy. She identifies task complexity as the combination of two components – clarity of information and volume of information. This study examines these two components of task complexity, as well as skill, on the over/under evaluation of internal controls.

A second, but related issue, concerns the effect of a work program on decision accuracy, including over/underestimation of internal control. This issue assumes important given the Act's requirement that staff auditors be properly supervised and that this fact be documented. Since many CPA firms engaged in the practice of auditing SEC clients employ work programs, and use them to document work performed as well as its review, work programs can be used as a form of documentation for compliance with the Act. An additional benefit in the form of training potential may arise if work program use improves decision accuracy, including the reduction of over/underestimation of internal control. Prior research suggests that audit tools such as questionnaires assist auditors in identifying material internal control weaknesses by triggering knowledge of internal control from memory (Bierstaker, et al., 2004), implying that a tool which directs attention to one or more salient features (e.g. a work program) has the same potential at increasing auditor accuracy. However, since such a tool is not uniform in its effect on decision accuracy (Bierstaker, et al. 2004) and because task-specific knowledge is a critical covariate affecting decision accuracy, we examine the combined effects of task specific knowledge (skill) and a work program on decision accuracy and over/underestimation of internal control.

We examine these issues using a mixed factors laboratory experiment with work program use varied between subjects two ways (present/not present) and estimation of internal control measured at four discreet within-subject intervals: pretest, two with/without work program, and posttest. We use senior-level accounting students enrolled in the auditing course from three different universities to test our hypotheses. Since students enrolled in the auditing course generally matriculate in a relatively short time frame upon completion of the course and since these students represent the pool from which beginning staff auditors are selected, senior-level students are appropriate for purposes of this study. Additionally, the task used in this study is similar to the one used by Ashton and Kramer (1980) who note that senior-level accounting students are appropriate surrogates for novice auditors.

We find that although both clarity of information and quantity of information are important in determining task complexity, clarity of information appears to dominate the definition. Skill level is also key in interpreting experimental results: both high- and low-skill subjects improve their risk assessment judgments in complex conditions the more they perform such risk assessments, however, both high- and low-skill auditors make consistently worse and significantly overstated risk assessment judgments under repetitive simple conditions. In other words, when faced with several, simple risk assessment judgments, auditors revert to conservatism, assessing control risks too high.

With respect to work program use, our results indicate that decision accuracy, including over/underestimation of internal control, does not significantly differ for subjects provided the work program versus those in the control group. These results support the pattern of findings from Bierstaker, et al. (2004) and suggest that work program use without feedback is ineffective for beginning auditors.

The remainder of the paper is as follows. Part II discusses the literature and develops the hypotheses. Part III describes the experimental design and methodology, and Part IV presents the findings, implications, and directions for future research.

II. LITERATURE AND HYPOTHESES

Task Complexity: Bonner's Model

Bonner (1994) performed an extensive review of past studies in the decision-making literature with the aim of isolating reasons for the disparate, often contradictory findings. Her research led to a new model of task complexity, one that portrays the relationship between task complexity and judgment performance as a function of task complexity, skill, and motivation, depicted as follows:

$$\text{Judgment performance} = -f\left\{ \frac{\text{Task complexity}}{(\text{Skill, Motivation})} \right\}$$

The model predicts that task complexity has an inverse relationship with judgment performance, such that as tasks become more complex, judgment performance worsens. However, the decision maker's skill level and motivation mediate this relationship by lessening the negative effect of task complexity on judgment performance; essentially, the decline in judgment performance is less for higher skilled (more motivated) decision maker's than for lower skilled (less motivated) decision makers.[1]

Bonner (1994) further identifies two main components of task complexity–clarity of information and volume (i.e. quantity) of information. However, neither the model nor Bonner identify the individual effects of each component of task complexity on judgment performance nor whether one component dominates the other or whether the components contribute equal weight in determining task complexity. One goal of this study is to understand how the components–clarity and volume of information–affect judgment performance, both individually and together.

Task Complexity Components

Bonner (1994) demonstrates that the model components are slightly different depending on the phase of the judgment process. The judgment process consists of three phases–input, processing, and output. In the input phase, the decision maker encounters the information needed to make the decision; in the processing phase, the individual performs the cognitive work necessary for making a decision; the output phase is the method used to communicate the decision. Because the three phases link serially, information introduced during the input phase affects the next two phases; therefore, we examine task complexity effects for the first phase.

For the input phase, Bonner (1994) defines clarity of task as the degree to which information cues are consistent with each other and with information stored in memory. When cues are consistent, the task is clear; when cues are inconsistent, the task lacks clarity. Bonner defines volume of information (hereafter referred to as quantity) as the number of alternatives, number of cues, or number of procedures that must be processed as part of the task. When the

[1] Our paper focuses on task complexity and the components of task complexity. We also examine the mediating effects of skill; however, we leave the mediating effects of motivation for future research.

number of cues is low, the task is easier; as the number of cues increase, the task becomes more complex. Based on these definitions, the most complex task exists when clarity is low and quantity is high. Conversely, the least complex, or simple, task exists when clarity is high and quantity is low.

Clarity affects judgment performance in its effect on working memory. Since working memory is notoriously limited in capacity and duration (e.g. time; Anderson 1983), taxing its capacity causes detrimental effects on judgment performance. For example, as information clarity decreases because of inconsistent information cues, working memory decays from temporarily storing inconsistent cues while retrieving existing knowledge from long-term memory to assist in combining the inconsistent cues into a final judgment. For inconsistent cues, this judgment process requires more time and more processing effort as the decision maker searches long-term memory for cues to compare with the inconsistent cues to form a judgment. If working memory is insufficient to accommodate the entire process, or if information decays from working memory prior to the judgment point, then judgment performance suffers.

The same argument holds for tasks high in quantity of information. Research supports the premise that as the amount of information increases, working memory decays, inducing fatigue, causing an increase in errors (Iselin 1988). The higher number of information cues requires additional working memory in order to sort and process the cues. The additional processing demands from the large quantity of cues results in loss of working memory, therefore, compromising judgment performance.

Auditor Judgments

Due to the nature of auditing and the types of judgments auditors make, auditor judgments are rarely considered "right or wrong." Of more concern is whether the auditor's judgment is over- or understated, resulting in different consequences, particularly in performing risk assessment in the planning phase of an audit. For example, auditors assess control risk during the planning of an audit. If control risk is judged too high, then the auditors incur the risk of performing too much testing; however, if control risk is judged too low, then the auditors incur the risk of doing too little testing. Too much testing makes the audit inefficient, while too little testing makes the audit ineffective, both resulting in higher costs to the auditor.

Bonner's (1994) model defines judgment performance in terms of "better or worse" rather than over- and understatement; however, the model shows that when the judgment involves inconsistent evidence (i.e., low clarity) or lots of evidence (i.e., high quantity), the ability for working memory to process information properly declines. When auditors cannot properly process the information, they more likely overstate than understate their judgments of risk assessment due to conservatism. Conservatism predicts that when given a choice, auditors prefer to perform more testing than needed rather than less testing, since the costs associated with audit failure are so great. Therefore, this study examines the following hypotheses, stated in the alternate form.

H1a: Auditors performing tasks low in clarity will significantly overstate their judgments of risk assessment.

H1b: Auditors performing tasks high in quantity (i.e., large amounts of information) will significantly overstate their judgments of risk assessment.

Conversely, when information is consistent (high clarity) and/or only relevant information is provided (low quantity), working memory is not as heavily taxed, allowing for better processing of all available information. Since information is processed more clearly (i.e., high clarity) and more concisely (i.e., low quantity), auditor judgments concerning risk assessment should be more accurate; however, the clear and/or concise information may erroneously lead auditors to conclude that conditions are better than they really are. In such cases, auditors may understate their judgments of risk assessment. Since the effects of high clarity and/or low quantity are not clear, we test the following hypotheses, stated in the null form.

> **H2a: Auditors performing tasks high in clarity will not significantly overstate or understate their judgments of risk assessment.**

> **H2b: Auditors performing tasks low in quantity (i.e., small amounts of information) will not significantly overstate or understate their judgments of risk assessment.**

Because tasks seldom differ only in clarity or quantity of information, this study examines the effects of combined levels of clarity and quantity of information on judgment performance. Since Bonner (1994) defines low clarity and high quantity of information, individually, as more complex, then tasks consisting of both low clarity *and* high quantity of information are the most complex. Moreover, since hypotheses 1a and 1b predict low clarity and high quantity of information individually result in overstated judgments of risk assessment, then combining low clarity of information with a high quantity of information (i.e., complex task) should also result in overstated judgments of risk assessment.

In addition, conservatism theory indicates auditors are more likely to assess internal control risk too high rather than too low in order to reduce the costs of errors through increased audit testing. Conservatism is more likely to affect complex tasks than simple tasks because complex tasks often involve a large degree of auditor judgment (e.g. auditing contingencies). Even though good internal controls may exist in the area audited, the increase in judgment for that area requires more auditing. For example, auditors are more likely to increase testing in areas such as contingent liabilities, despite good internal controls for recording liabilities. The increased judgment required in recording contingent liabilities requires increased auditing. Therefore, auditors are more likely to overstate their risk assessments for internal controls for more complex audit areas (more complex tasks). Given Bonners' (1994) model predictions and conservatism theory, this study examines the following hypothesis, stated in the alternate form, for complex tasks.

> **H3a: Auditors performing a complex task (low clarity, high quantity) will significantly overstate their judgments of risk assessment.**

Likewise, the combination of high clarity with a low quantity of information is the least complex or a simple task. When tasks are simple, auditors prefer efficiency over effectiveness, since the cost of error is not as great for simple judgments as for complex judgments. For example, auditors are more likely to cut testing in areas such as prepaid expenses, where even poor internal controls most likely do not create a significant error; however, auditors are less likely to cut testing in accounts receivable (more complex judgments required) even when good internal controls exist. Therefore, auditors performing simple judgment tasks are more likely to understate their judgments of risk assessment.

However, Bonner's (1994) model predicts that consistent information (high clarity) and relevant information (low quantity) do not tax working memory as greatly, allowing better processing of information. This better processing of information should result in more accurate risk assessment judgments. Thus, auditors performing simple tasks could make risk assessment judgments similar to the expert system's risk assessment judgments.

On the other hand, auditors could find simple tasks to be too simple; thus, they become too confident in their knowledge and do not give adequate attention to the information presented. Mascha (2001) finds that auditing students did not pay attention to the information presented with a simple task, and they were overconfident in their judgments. Therefore, auditors may over rely on their previous knowledge and not pay attention to the current information presented. Since audit education focuses on conservative, auditors may overstate their risk assessment judgments. This education may override the information presented in simple tasks.

Since theory does not support a specific direction for auditors' risk assessment judgments for simple tasks, we test the following null hypothesis.

> **H3b: Auditors performing a simple task (high clarity, low quantity) will not significantly understate or overstate their judgments of risk assessment.**

Currently, no clear direction exists for the two tasks that combine the same amounts of clarity and quantity of information (i.e., low clarity, low quantity and high clarity, high quantity). Bonner (1994) does not provide guidance as to which component has a dominant effect. One goal of this paper is to examine whether one task component dominates the other in creating task complexity. If one component dominates the other, then a task complexity continuum forms between simple and complex tasks. Since prior research does not address this issue, this study examines the following research question:

> **RQ: Does clarity or quantity of information dominate in creating task complexity such that a task complexity continuum forms?**

Skill and Task Complexity

Cognitive research finds that subjects who are more familiar with a task integrate and process information more effectively because their task-specific knowledge reduces cognitive effort and any negative impacts on working memory (e.g. Anderson 1983). This literature suggests that auditors possessing high levels of task-specific knowledge should integrate and process information to make more accurate control risk assessments than auditors possessing low levels of task-specific knowledge. Therefore, as task-specific knowledge (i.e. skill) increases, auditors' overstatement or understatement of control risk decreases. We test the following hypothesis for skill.

H4a: High-skill auditors will make significantly better risk assessment judgments than low-skill subjects (as measured against expert system risk assessment judgments).

Bonner's (1994) model asserts that skill mediates task complexity effects on judgment performance. Defining "skill" as a person's level of task-specific knowledge, her model proposes that as skill-level increases, the negative effects of task complexity on judgment performance decrease, but the effects decrease *more* for low-skill levels than for high-skill levels. Therefore, Bonner's model proposes that the improvement in auditors' control risk assessments will be *greater* for low-skill auditors than for high-skill auditors.

Research examining skill effects reports that novice subjects (i.e. low-skill subjects) frequently overestimate control risk assessments and frequently are over confident about the accuracy of their internal control judgments (e.g. Hornick and Ruf 1997). These findings suggest that low-skill auditors will overstate their control risk assessments for all levels of task complexity. Combining this information with Bonner's (1994) model, we can expect that low-skill auditors will overstate their risk assessment judgments for both complex and simple tasks, but the overstatement will be *less* for simple tasks than for complex tasks. We test the following hypothesis concerning low-skill auditors.

H4b: Low-skill auditors performing a complex task will overstate their judgments of risk assessment significantly *more* than low-skill auditors performing a simple task.

Additionally, research examining skill effects reports that high-skill subjects often deskill when presented with simple tasks (Mascha and Smedley 2007; Sutton and Arnold 2003) or, at best, do not improve further (Mascha 2001). These more experienced subjects may become bored with simple tasks, and, therefore, they do not attend to the information cues or do not process all of the information. These findings coupled with Bonner's (1994) model suggest that high-skill auditors will improve their control risk assessment judgments for complex tasks, but may not improve, or may even make worse, their control risk assessment judgments for simple tasks. The following hypotheses summarize the mediating effects of skill on task complexity.

> **H4c: High-skill auditors performing a complex task will overstate their judgments of risk assessment *less* than low-skill auditors performing a complex task; however, high-skill auditors performing a simple task may overstate their judgments of risk assessment as much as, or even more than, low-skill auditors performing a simple task.**

Work Program Use and Decision Accuracy

An audit work program is a step-by-step list of instructions for the auditor to follow in auditing a specific area. As such, audit work programs assist auditors, especially novices such as newly graduated students, by directing their attention to the task at hand. Attention directing reduces cognitive load by providing guidance, thereby reducing the amount of working memory required in performing a task (Rose and Wolfe 1998). This should translate into better (e.g. more accurate) decision accuracy, defined here as little or no over/underestimation of risk assessment.

However, audit work programs are not true decision aids since they provide no feedback or explanation of how to amend the audit program if discrepancies are found. For this purpose, audit work programs may not be appropriate for auditors with virtually no task experience, such as those used here. In a study of questionnaire use on audit accuracy, Bierstaker, et al. (2004) note that questionnaire use did not assist low-experience auditors in making better decisions. They speculate that since these auditors had little task experience the questionnaire was ineffective at priming long-term memory since no memory existed. While the audit program used here differs from the questionnaire used by Bierstaker, et al., the parallel between the two instruments is similar in its intent at directing attention to salient audit steps that are important in the assessment of internal control. Given the uncertainty associated with use of audit work programs, we offer the following hypothesis, stated in the null.

H5: Use of an audit work program will not significantly affect decision accuracy nor will its use lessen over/understatement of risk assessment.

III. EXPERIMENTAL DESIGN

Experimental Design and Task

We tested the hypotheses using a mixed factors design.[2] This design allows us to examine the effects of task complexity and skill between subjects as well as within subjects. We test the hypotheses using between subjects analyses, and we include the within subjects analyses to evaluate the effects of skill and task complexity on judgment performance over more than one setting, as is done in practice when auditors plan more than one audit at a time.

Our mixed factors design combines two levels (high and low) for each component of task complexity (clarity and quantity) on a between-subjects basis, with four cases on a within-subjects basis. This combination yields four between-subjects cells with four within-subjects cells. Skill level is a continuous variable measure of the subject's task-related knowledge. This covariate measure equals the subject's case 1 internal control evaluation scores.

Consistent with prior studies examining internal control evaluation, we select an internal control review of the payroll cycle. This task is frequently assigned to beginning auditors (Ashton and Kramer, 1974) and is an internal control cycle that is generally familiar to our subjects, providing the potential for low- and high-skill level subjects.

[2] We designed the experiment to examine several issues concerning task complexity, some of which are not

We recruited subjects at three Midwestern state universities by offering bonus points for participation.[3] Subjects completed all experimental activities under the direct supervision of one of the authors. Subjects were senior level accounting majors enrolled in the auditing course. Students are appropriate since our primary purpose is to test theory.

Experimental Sequence

Upon arrival, students complete the consent form and a demographic questionnaire designed to measure prior experiences as well as course work. They begin the experiment by reading a brief description concerning the background and circumstances surrounding the payroll environment in general. Next, subjects review a specific payroll case scenario and provide internal control evaluations for five components of the payroll cycle, as well as an overall control evaluation. Subjects complete four payroll cases in the experimental task. The first case represents the student's pre-task knowledge; the remaining three cases represent repeated measures of the dependent variable designed to measure within-subject effects.

We assigned subjects randomly to either the audit work program group or the control group.. Within each audit work program group we further randomly subdivided subjects into one of four task complexity treatments, such that all cases presented to each subject remain consistent as to level of complexity (i.e. subjects in the simple treatment see four separate cases of simple complexity). While the complexity level remains the same, the cases are not identical. Subjects evaluated the internal control environments described in their cases four time: pretest (without the audit work program), two with (without the audit work program), and one at posttest. The subjects conclude the experiment by completing a brief post-experimental

---

included in this paper but may be included in future papers.

[3] Bonus points are a common reward given to students for participation in research studies. Our students received a minimal number of bonus points, representing less than 2-3 percent of the total points for the audit course.

questionnaire designed to measure attitudes toward the tasks.

For consistency and comparison, the experimental materials are from Eining and Dorr's (1991) expert system study. Each case presents descriptions of internal payroll controls for five components–initial hiring and termination, recording of time worked, calculation and preparation of payroll, payment and distribution of wages, and other controls. The subjects provide quantitative evaluations (from 0 = absence of all controls to 100 = all controls are present and working as designed) for each of the five component sections as well as an overall evaluation of payroll internal control.

Task Complexity

We manipulate the clarity component of task complexity by varying the consistency of internal controls described in the cases. High clarity exists when all cues provide a similar signal regarding the state of the internal control environment, e.g. all strong internal control cues or all weak internal control cues. Conversely, low clarity is a mix of cues such that no clear signal (i.e. strong internal control or weak internal control) can be determined. This manipulation is identical to the one successfully used by Mascha (2001).

We manipulate the quantity component (i.e. amount of information) of task complexity by varying the number of cues provided to the subject. The high quantity subjects receive 12 cues, whereas the low quantity subjects receive five cues. The cues provide both relevant information (e.g. the personnel department approves all new hires) and irrelevant information (e.g., payday is every Tuesday).

Skill Definition

Bonner (1994) stipulates that skill is primarily a function of task-related knowledge, i.e.

Instructors noted the bonus points provided sufficient motivation for students to attend to the tasks.

the knowledge and abilities related to a specific task, not knowledge or abilities in general. Since recent literature supports Bonner's definition (see Tan and Tao 1999; Tan et al. 2002), we adopt this definition as well. We use the subject's score for the overall evaluation of internal control over payroll from the first case as the measure of the subject's skill level. This measure is a continuous variable, representing the subject's task-related knowledge. Scores closer to the expert system scores are considered "high-skill," whereas scores further from the expert system scores are considered "low-skill".

We examined the stem-leaf diagram for the subject's overall evaluation of payroll controls for the first case. Since the dependent variable (discussed next) is the difference between the subject's evaluation score and the expert system's score, scores (the difference scores) closest to zero are more correct than scores further from zero. Therefore, we categorize the "high" skill level subjects as those whose scores are between positive 10 and negative 10 (i.e. $-10 \leq X \leq 10$, where X is the subject's difference score). We categorize all scores greater than positive 10 or less than negative 10 as "low" skill level subjects. This procedure results in 48 high-skill subjects and 83 low-skill subjects.

Dependent Variable

The dependent variable is the subject's internal control risk assessment for the payroll cycle (i.e. risk assessment), measured as the difference between the subject's overall evaluation of internal control and the overall evaluation determined with an expert system. This definition is consistent with prior research examining decision accuracy for judgment tasks (Mascha 2001; Odom and Dorr 1995; Eining and Dorr 1991; Libby and Libby 1989).

The range for the risk assessment scores is potentially -100 to +100. A negative score means the subject evaluated the payroll internal controls lower than the expert system score, equivalent to an auditor assessing control risk too low. A positive score means the subject evaluated the payroll internal controls higher than the expert system score, equivalent to an auditor assessing control risk too high.

IV. ANALYSES AND RESULTS

Demographics

One hundred forty one students participated in the experiment. We exclude ten subjects from the analyses because of missing data, resulting in a total sample of 131. Thirty-two have prior payroll accounting work experience, and six have prior internal control review experience. Every treatment group contains students with internal control and/or payroll accounting work experience. The final sample comprises 70 males (53%) and 61 females (47%). Chi-square tests indicate no significant differences (at the 0.05 level) between treatment groups for experience and gender.

Table 1 shows the number of subjects, average age, average GPA overall and by major, average credit hours and number of subjects with internal control and payroll accounting experience by treatment group. Analysis of variance or Chi-square tests show no significant differences (at the 0.05 level) between treatments on these demographic variables, suggesting successful randomization (Kerlinger 1989).

Insert Table 1 here

Table 1 also shows some notable characteristics about this subject group. First, the average age is approximately 25 years old, because subjects at two of the institutions have older students returning to school and part-time students. The average overall and major GPAs are quite stable at 3.25 or a B+.

Hypothesis Testing: Audit Work Program Use versus Non-use

Prior to testing the hypotheses related to risk assessments, we performed ANOVAs comparing the decision accuracy for each of the two work program groups. Our goals was to determine if there were significant difference in decision accuracy between the two groups thus requiring separate analyses for the task and skill hypotheses or whether there were no statistical differences and the two groups could be combined for further analysis.

Results indicate no significant difference in decision accuracy, including tendency to over/understated risk assessment between subjects who used the audit work programs and those who did not ($p > .10$). This effect held whether we used pretest versus posttest scores or compared all four internal control assessments for the two groups. As a result, we combined the two groups into one group for testing of the task and skill hypotheses.

Hypotheses Testing – Task Complexity and Components

Hypotheses 1a and 1b predict that subjects will significantly overstate risk assessments for tasks lacking in clarity or high in quantity of information. To test these hypotheses, we test the mean risk assessments for significance from zero. This test is appropriate because the means represent the difference between the subject's score and the expert system score. Therefore, scores closest to zero, either positive or negative, are more "accurate" while scores further from zero are overstated judgments when positive and understated judgments when negative.

A repeated measures analysis of variance test shows that both clarity and quantity significantly affect the subjects' risk assessment judgments (using Wilks' Lamda: $F = 23.62$ p-value $<.0001$ for clarity; $F = 16.66$ p-value $<.0001$ for quantity). Thus, a significant difference exists between high and low clarity and high and low quantity conditions. Table 2, Panels A and B, displays the means, significance values, and number of subjects for high/low clarity and high/low quantity conditions over the four cases of payroll internal controls.

As noted in Panel A, Table 2, hypotheses 1a and 1b are initially supported – subjects significantly overstate their risk assessment judgments for payroll controls for cases 1 and 2. However, for both low clarity and high quantity conditions, the risk assessment judgments improve from case 1 to case 4 to the point where the mean scores are not significantly different from zero, i.e. are not significantly overstated or understated (except low clarity, case 4, where the mean risk assessment score is significantly understated). These results suggest that as auditors repeatedly perform tasks low in clarity or high in quantity, their risk assessment judgments more closely agree with the expert systems' risk assessment judgments for payroll internal controls.

Insert Table 2 here

Table 2, Panel B, summarizes the results for testing hypotheses 2a and 2b. These hypotheses are stated in the null form, i.e. subjects' risk assessment judgments for tasks high in clarity or low in quantity will not be significantly overstated or understated. The results in Panel B show three significant means for both high clarity and low quantity conditions; thus, rejecting hypotheses 2a and 2b. For tasks high in clarity or low in quantity, subjects consistently overstate their risk assessment judgments for payroll internal controls. In addition, on average, subjects in the low quantity condition overstate or understate their risk assessment judgments more than subjects in the high clarity condition.

Overall, the results shown in Table 2 indicate that, initially, subjects overstate their risk assessment judgments (assessing control risk too high) over payroll internal controls under all four conditions of task complexity elements – high/low clarity and high/low quantity of information. However, on average, the subjects in the more difficult conditions, low clarity or high quantity, improve their risk assessment judgments over time (from case 1 to 4), while subjects in the more simple condition, high clarity or low quantity, continue to overstate their risk assessment judgments over time.

Hypotheses 3a and 3b and the research question examine the combinations of clarity and quantity together, creating simple versus complex task conditions. Table 3, Panel A, presents the results of complex and simple tasks, and Table 3, Panel B, presents the other two combinations of clarity and quantity (i.e. low clarity, low quantity and high clarity, high quantity) on the subject's risk assessment judgments.

Insert Table 3 here

Hypothesis 3a predicts that auditors significantly overstate their risk assessment judgments when performing complex tasks. Table 3, Panel A, shows that, initially, hypothesis 3a is supported. For cases 1 and 2, subjects significantly overstate their risk assessment judgments for payroll internal controls. However, as they continue, subjects in the complex condition change their risk assessment judgments to negative scores, i.e. understating their risk assessment judgment for payroll internal controls, significantly for case 3 ($p = .0001$) but not significantly for case 4 ($p = .1139$). Overall, as subjects performed the complex task multiple times, their risk assessment judgments improved. The last risk assessment judgment (case 4) was not significantly different from zero, indicating that, on average, the subjects' risk assessment judgment is close to the expert system's risk assessment judgment for this case.

Hypothesis 3b presents the null hypothesis that auditors do not significantly overstate or understate their risk assessment judgments of internal controls under simple task conditions. Table 3, Panel A, shows that for three of the four cases, subjects overstate their risk assessment judgments of payroll internal controls. This result supports the theory that auditors rely on their previous education or knowledge when making risk assessment judgments in a simple task condition. Therefore, we reject the null hypothesis 3b. When presented with multiple simple tasks, auditors overstate their risk assessment judgments of payroll internal controls.

Our research question asks whether clarity or quantity of information dominates the other in determining task complexity. Table 3, Panel B, presents the results of testing the two conditions of low clarity, low quantity and high clarity, high quantity. For three of the four cases, the pattern and significance of the two conditions are in the same direction – cases 1 and 3 result in significantly overstated risk assessment judgments while case 2 results in a non-significant understated risk assessment judgment. However, the low clarity, low quantity condition results in larger risk assessments, i.e. those further from zero (i.e. further from the expert system risk assessment), than the high clarity, high quantity condition, which results in risk assessment judgments closer to zero, i.e. closer to the expert system's risk assessment.

Comparing the risk assessment judgments from these high/high and low/low conditions to the risk assessment judgments of the complex and simple conditions, we find that the high clarity, high quantity condition resembles the simple task condition (high clarity, low quantity), while the low clarity, low quantity condition resembles the complex task condition (low clarity, high quantity). These comparisons suggest that clarity of information dominates quantity of information because the clarity element determines whether the high, high or low, low conditions are more simple or more complex. However, the results do not reflect a strong pattern such that a clear task complexity continuum occurred.

Hypotheses Testing – Task Complexity and Skill

Hypotheses 4a suggests a significant difference between high-skill and low-skill auditors making payroll internal control risk assessment judgments. The repeated measures analysis of variance test finds a significant effect for skill (p value = .0395) and for the interaction of skill with task complexity (p value = .0054). Since the first case is used to measure the subject's level of skill, only cases 2-4 are included in the repeated measures analysis of variance test. For all three cases, 2-4, the low-skill subjects significantly overstate their risk assessment judgments, while the high-skill subjects risk assessment judgments are not significantly different from zero, indicating, on average, the high-skill subjects' risk assessments are close to the expert system risk assessments. Table 4 summarizes the effect of skill on risk assessment judgments. This result supports hypothesis 4a.

Insert Table 4 here

Hypotheses 4b and 4c focus on skill and its effect on task complexity. Hypothesis 4b states that low-skill auditors performing complex tasks will *overstate* their risk assessment judgments *significantly more* than when they *overstate* their risk assessment judgments when performing simple tasks. Table 5, Panel A, summarizes the results of the repeated measures analysis of variance test for this hypothesis.

Insert Table 5 here

Comparing the results for low-skill subjects between Panel's A and B, we find partial support for hypothesis 4b. Low-skill subjects performing a complex task significantly perform worse than low-skill subjects performing a simple task for two of the three cases (case 2 and 3); however, the subjects do not consistently overstate their risk assessment judgments as prescribed in hypothesis 4b.

When performing a complex task, the low-skill subjects first significantly overstate their risk assessment judgments, and then they switch to significantly understating their risk assessment judgments. Although they flip from overstatement to understatement, these low-skill subjects improve their risk assessment judgments, moving closer to the expert system risk assessment judgment for a complex task.

On the other hand, when performing a simple task, the low-skill subjects begin with a risk assessment judgment not significantly different from the expert system risk assessment judgment. Then they progressively overstate their risk assessment judgments, resulting in significantly overstated risk assessment judgments. These low-skill subjects did not improve performance for a simple task as theorized; they performed worse over time.

Hypothesis 4c states the mediating effects of skill on task complexity. The expectation is that high-skill subjects will not significantly over or understate their risk assessment judgments when performing a complex task, but may significantly overstate their risk assessment judgments when performing a simple task. In fact, high-skill subjects may overstate their risk assessment judgments more than the low-skill subjects when performing a simple task. On the other hand, the expectation for low-skill subjects is that they will consistently overstate their risk assessment judgments for both complex and simple tasks. However, they will overstate their risk assessment judgments *less* for simple tasks, representing improved performance from the complex task.

The results show that high-skill subjects and low-skill subjects perform the complex task differently. First, the high-skill subjects overstate their risk assessment judgments only for case 4, while the low-skill subjects overstate their risk assessment judgments only for case 2. Second, the high-skill subjects are more consistent with the expert system in their risk assessment judgments for two cases, case 2 and 4. These risk assessment judgments are not significantly different from zero (case 2: $F = -1.60$, p value = .8425; case 4: $F = 4.00$, p value = .5700). Overall, two of the three cases show that high-skill subjects outperform the low-skill subjects performing the complex task, but they do not overstate their risk assessment judgments as predicted in hypothesis 4c. However, we acknowledge a potential limitation on these findings because of the small number of subjects in the high-skill group (5 subjects).

For the simple task, the results show that high-skill and low-skill subjects perform the same, supporting the hypothesis 4c expectation that high-skill subjects will perform the same as or worse than the low-skill subjects. As the subjects progress from case 2 to case 4, their risk assessment judgments become significantly more overstated. This finding supports prior research that high-skill subjects may de-skill when performing a simple task. Interestingly, both high-skill and low-skill subjects de-skill when performing a simple task over time.

V. DISCUSSION AND CONCLUSIONS

This study examines task complexity using a within- and between-subjects experimental design for specifically investigating the effect of skill on task complexity over time as well as by level of complexity. We use Bonner's (1994) model and cognitive research in developing the hypotheses. A key part of this study also involves exploratory research into the mixed components of task complexity (i.e. high clarity, high quantity and low clarity, low quantity) not generally examined in prior research.

We first examine the components of task complexity – clarity and quantity of information – separately. The results primarily show that for each condition – high clarity, low clarity, high quantity, and low quantity – subjects, on average, significantly overstate their risk assessment judgments for payroll internal controls. This finding holds when we examine the combinations of these components. For simple (high clarity, low quantity), complex (low clarity, high quantity), and mixed components (high clarity, high quantity and low clarity, low quantity), subjects significantly overstate the risk assessment judgments for payroll internal controls. These findings support the use of conservatism in assessing control risk, preferring to assess control risk too high, resulting in more auditing than may be necessary.

When skill is considered, the results show that high-skill subjects make risk assessment judgments for payroll internal controls similar to the expert system when performing a complex task; however, these high-skill subjects de-skill when performing a simple task. High-skill subjects progressively overstate their risk assessment judgments under the simple task condition. This finding is similar to prior research when studying effects of skill levels (Mascha 2001; Miller et al. 2006).

The low-skill subjects perform somewhat as expected when performing a complex task. While they significantly overstate their risk assessment judgments for payroll internal controls as predicted for one case, they also significantly *understate* their judgments under the complex task condition. Contrary to the expectation that low-skill subjects perform better when performing a simple task, our low-skill subjects perform like the high-skill subjects – significantly overstating their risk assessment judgments for the simple task.

A unique feature of our study is the performance of the tasks over time. Subjects perform risk assessment judgments for payroll internal controls for four cases within their assigned experimental condition. From this design, we note the following conclusions. When tasks are complex, or consist of only low clarity or high quantity of information, subjects' risk assessment judgments improve over time. In other words, their risk assessment judgments approximate the expert system risk assessment. For simple tasks, subjects' risk assessment judgments decline over time, resulting in larger, overstated judgments. This finding extends to both high-skill and low-skill subjects; both groups de-skill when performing a simple task multiple times.

Another goal of our study is to examine the effects on risk assessment judgments for internal controls for tasks designed as high clarity, high quantity and low clarity, low quantity.

We find that the high, high condition mimics the simple task condition, while the low, low condition mimics the complex task condition. This result indicates that clarity of information dominates quantity of information in defining task complexity. More research is needed in this area to definitively construct a task complexity continuum with the task complexity components.

Taken together these results imply the following for audit practice. First, risk assessment judgments vary between skill levels and complexity levels, even over a short period. This finding supports current practice where more experienced (presumably high-skill) auditors perform the more complex risk assessments. In addition, auditors performing complex risk assessments will improve their skills as they make more and more of these assessments. However, for both new and experienced auditors, performing risk assessments under conditions where the client's descriptions of internal controls are clear and concise may still result in assessing control risk too high. In addition, performing risk assessments over several audit areas (i.e. same task over short period of time) where the information is clear and concise (i.e. simple task) may mislead the auditor in not paying close attention to the information, resulting in assessing control risk too high and performing too much auditing. Secondly, conservatism clearly plays a dominant role in auditor's risk assessment judgments. Under all conditions, when the subjects' risk assessment judgments did not approximate the expert system risk assessment, the subjects overstated their risk assessment judgments, erring on the side of conservatism. For complex situations, this error is probably best, but for simpler situations, this error can increase the cost of the audit considerably and unnecessarily.

For researchers using judgment tasks, the findings for clarity and quantity of information are important for experimental task design. When incorporating task complexity as part of an experimental design, the clarity of information is more important than the quantity of information for measuring task complexity. In addition, skill levels of the subjects significantly affect the comparison between simple and complex tasks.

This study is not without its limitations, and future research could address these concerns. One is the artificial classroom setting in which the task was performed, which raises questions of subject attention to task. The researchers did not observe any severe lack of attention to the task. The subjects also performed the task several times, so they all were familiar with the task requirements. Second, the subjects are students, not practicing auditors. While students are adequate surrogates for beginning auditors (Ashton and Kramer 1980), results derived from using students may not generalize to auditors as a whole. We also had a very small sample size for our high-skill subjects performing a complex task. Despite the small sample size, the results are as expected. Finally, we used only one task. To alleviate these concerns and limitations, future research could use a variety of tasks performed by practicing auditors, such as audit area judgments (e.g. adequacy of allowance for doubtful accounts, adequacy of warranty expense/payable, etc.) and audit opinion determination (e.g. going concern, explanation of a significant matter, etc.).

Table 1

Average Age, GPA, Credit Hours and Experience by Treatment Group

Group	No. of subjects	Avg. Age (yrs)	Average Overall GPA (4.00=A)	Average Major GPA (4.00=A)	Avg. Credit Hours	Exp. with Internal Controls? Yes	No	Exp. with Payroll Accounting? Yes	No
1 - Simple[a]	34	25	3.25	3.27	114	4	30	10	24
2 - Mixed[b]	31	25	3.26	3.31	113	1	30	8	23
3 - Mixed[c]	34	28	3.25	3.16	120	4	30	6	28
4 - Complex[d]	32	24	3.26	3.29	111	2	30	7	25
Totals	131	25.6	3.25	3.25	114	11	120	31	100

[a]Group 1 Simple consists of low quantity and high clarity of information.
[b]Group 2 Mixed consists of high quantity and high clarity of information.
[c]Group 3 Mixed consists of low quantity and low clarity of information.
[d]Group 4 Complex consists of high quantity and low clarity of information.

Note: ANOVA and Chi-square tests show no significant differences between treatment groups on the above demographic variables.

Table 2

Repeated Measures Analysis of Variance Summary Tables and Judgment of Risk Assessment Means Tests of Hypotheses 1 and 2

Panel A: Hypotheses 1a and 1b – H_A= Judgments of risk assessment will be significantly overstated.

Number of subjects	65		63	
	Low Clarity		**High Quantity**	
Case	Mean	p value	Mean	p value
Case 1	**21.13**	**.0001**	**14.93**	**.0001**
Case 2	**8.19**	**.0008**	**9.55**	**.0001**
Case 3	3.16	.1868	-3.66	.1316
Case 4	**-4.57**	**.0215**	0.32	.8736

Bold values represent significant difference from zero at .05 alpha level.

Panel B: Hypotheses 2a and 2b – H_o= Judgments of risk assessment will not be significantly over- or understated.

Number of subjects	66		68	
	High Clarity		**Low Quantity**	
Case	Mean	p value	Mean	p value
Case 1	**8.28**	**.0001**	**14.49**	**.0001**
Case 2	-0.74	.7530	-2.10	.3675
Case 3	**6.77**	**.0048**	**13.59**	**.0001**
Case 4	**8.59**	**.0001**	***3.71***	***.0556***

Bold values represent significant difference from zero at .05 alpha level.
Bold & Italicized values represent significant difference from zero at .10 alpha level.

Table 3

Repeated Measures Analysis of Variance Summary Tables and Judgment of Risk Assessment Means Tests of Hypothesis 3 and Research Question 1

Panel A: Hypotheses 3a and 3b – Complex tasks: H_A= Judgments of risk assessment will be significantly overstated. **Simple tasks**: H_o= Judgments of risk assessment will not be significantly over- or understated.

Number of subjects	31		34	
	Complex		**Simple**	
Case	Mean	p value	Mean	p value
Case 1	**20.42**	**.0001**	**7.21**	**.0102**
Case 2	**19.29**	**.0001**	-1.76	.5805
Case 3	**-14.39**	**.0001**	**7.26**	**.0152**
Case 4	-4.52	.1139	**11.88**	**.0001**

Bold values represent significant difference from zero at .05 alpha level.

Panel B: Research question – Low, Low and High, High: No theoretical predictions for these tasks.

Number of subjects	34		32	
	Low, Low		**High, High**	
Case	Mean	p value	Mean	p value
Case 1	**21.76**	**.0001**	**9.41**	**.0012**
Case 2	-2.44	.4448	-0.03	.9939
Case 3	**19.91**	**.0001**	**6.78**	**.0276**
Case 4	-4.47	.1013	***5.21***	***.0646***

Bold values represent significant difference from zero at .05 alpha level.
Bold & Italicized values represent significant difference from zero at .10 alpha level.

Table 4

Repeated Measures Analysis of Variance Summary Tables and Judgment of Risk Assessment Means Tests of Hypothesis 4a

Test of Skill Effects - H_A= High-skill will make significantly better judgments of risk assessment than low-skill.

Number of subjects	48		83	
	High Skill		**Low Skill**	
Case	Mean	p value	Mean	p value
Case 2	-4.04	.1785	**6.62**	**.0016**
Case 3	-2.24	.4258	**7.27**	**.0002**
Case 4	1.22	.6407	***3.47***	***.0547***

Bold values represent significant difference from zero at .05 alpha level.
Bold & Italicized values represent significant difference from zero at .10 alpha level.

Table 5

Repeated Measures Analysis of Variance Summary Tables and Judgment of Risk Assessment Means Tests of Hypothesis 4b and 4c

Hypothesis 4b: Uses both Panel A and Panel B Low Skill Data: H_A= Low skill subjects' judgments of risk assessment for *complex task* will be *overstated significantly more* than low skill subjects' judgments of risk assessment for *simple task*.

Panel A: Hypothesis 4c – **Complex tasks**: H_A= High skill subjects' judgments of risk assessment will be significantly ***less overstated*** than low skill subjects' judgments of risk assessment.

Number of subjects	5		26	
	High Skill		**Low Skill**	
Case	Mean	p value	Mean	p value
Case 2	-1.60	.8425	**23.31**	**.0001**
Case 3	**-30.00**	**.0001**	**-11.38**	**.0008**
Case 4	4.00	.5700	**-6.15**	**.0479**

Bold values represent significant difference from zero at .05 alpha level.

Panel B: Hypothesis 4c – **Simple tasks**: H_A= High skill subjects' judgments of risk assessment will be significantly overstated as much as or more than low skill subjects' judgments of risk assessment.

Number of subjects	20		14	
	High Skill		**Low Skill**	
Case	Mean	p value	Mean	p value
Case 2	-1.95	.6284	-1.50	.7553
Case 3	***6.70***	***.0782***	***8.07***	***.0759***
Case 4	**10.30**	**.0040**	**14.14**	**.0010**

Bold values represent significant difference from zero at .05 alpha level.
Bold & Italicized values represent significant difference from zero at .10 alpha level.

REFERENCES

Abdolmohammadi, M., and A. Wright. 1987. An examination of the effects of experience and task complexity on audit judgments. *The Accounting Review* 62 (1): 1-13.

Anderson, J.R. 1983. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior* 22: 261-295.

Ashton, R., and S. Kramer. 1980. Students as Surrogates in Behavioral Accounting Research: Some Evidence. *Journal of Accounting Research* (Spring): 1-15.

Bonner, S. 1994. Model of the effects of task complexity. *Accounting, Organizations and Society* 19 (3): 213-234.

Eining, M., and P. Dorr. 1991. The impact of expert system usage on experiential learning in an auditing setting. *Journal of Information Systems* (Spring): 1-16.

Iselin, E. 1988. The effects of information load and information diversity on decision quality in a structured decision task. *Accounting, Organizations and Society* 13 (2): 146-164.

Kerlinger, F. 1989. *Foundations of Behavioral Research*, 3rd. edition; Fort Worth, TX: Harcourt Brace Jovanovich.

Libby, R. 1981. *Judgment and Decision Making*. New Jersey: Prentice Hall.

________, and P. Libby. 1989. Expert measurement and mechanical combination in control reliance decisions. *The Accounting Review* 64 (4): 729-747.

________, and J. Luft. 1993. Determinants of judgment performance in accounting settings: ability, knowledge, motivation, and environment. *Accounting, Organizations and Society* 18 (5): 425-450.

Mascha, M. F. 2001. The effect of task complexity and expert system type on the acquisition of procedural knowledge: Some new evidence. *International Journal of Accounting Information Systems* (2): 103-124.

Miller, C. L., D. Fador, and R. J. Ramsay. 2006. Effects of Discussion of Audit Reviews on Auditors' Motivation and Performance. *Behavioral Research in Accounting* (18): 135-146.

Odom, M., and P. Dorr. 1995. Epistemological issues of expert system use: an experimental investigation of knowledge acquisition and retention. *Journal of Information Systems* (Spring): 1-17.

Tan, H.T., and A. Kao. 1999. Accountability effects on auditors' performance: The influence of knowledge, problem-solving ability, and task complexity. *Journal of Accounting Research* 37 (Spring): 209-223.

_______, T.B. Ng, and B.W.Y. Mak. 2002. The effects of task complexity on auditors' performance: The impact of accountability and knowledge. *Auditing: A Journal of Practice and Theory* 21 (September): 81-96.