

# An Experimental Study of Incentive Pay

Abel Winn  
Market-Based Management Center,  
Wichita State University

**This is a Preliminary Draft. Comments and criticisms are welcome.**

**Abstract:** We use experimental methods to explore the influence of personality traits, risk preference and skill on agents' self-selection into payment contracts and their level of effort under those contracts. When contracts are imposed on subjects, roughly half respond to their incentives in accordance with agency theory, exerting effort only when it can increase their pay. The other half of subjects can be characterized as incentive indifferent, exerting maximum effort regardless of their financial incentives. Personality traits have some explanatory power in predicting incentive indifference. When subjects are allowed to self-select among contracts, behavior conforms much more closely to agency theory.

Personality traits also affect subjects' choice of payment contract. We find some support for the agency theoretic prediction that risk-preferring subjects self-select into contracts with a variable pay component. However, the prediction that variable pay contracts will also attract highly skilled subjects is not supported.

## I. Introduction

Economists typically envision the relationship between a manager and employee as a problem of asymmetric information and imperfect or costly supervision. The employee is assumed to have a utility function defined positively over wealth and negatively over effort. He is also assumed to have an opportunity cost in the form of a reservation wage. The manager is assumed to be a profit maximizer with the authority to define labor contracts. She has full knowledge of the employee's utility function and reservation wage, but imperfect information on his execution of duties. *Ex post* inference of the employee's effort is complicated by the fact that effort is mapped into outcomes by a stochastic function. The manager's challenge is to define a contract that will maximize the difference between the value of the employee's production and his earnings.

Agency theory provides a rigorous mathematical analysis of the manner in which the manager will meet this challenge, and there is no doubt that it is a useful tool for focusing the practitioner's attention on many of the salient issues of compensation. If the manager wishes to align the employee's incentives with the firm, agency theory reminds her that the cost of doing so may exceed the gains. If the manager wishes to attract a talented workforce, agency theory informs her that those with the highest skills have the most to gain from incentive pay. Yet when the manager considers how an employee's personal characteristics may affect his response to (and acceptance of) various pay contracts, standard theory has little to say. Given its assumptions that wealth and effort motivate the agent's behavior, agency theory advises the manager to investigate the applicant's utility function.

In practice, however, managers are much more likely to consider personality traits when designing compensation packages and making hiring decisions. Rather than (or in addition to) seeking the degree to which they are motivated by pay, managers tend to screen applicants for personal qualities like self-motivation, goal-orientation, and attention to detail. Firms' substantial expenditure of resources to investigate personality traits strongly suggests that there is a real return to doing so. Experimental methods are a useful tool to investigate whether or not this is the case.

Researchers have used experimental methods to investigate the validity of standard agency theory for at least a quarter century. In conducting experimental investigations of this sort, the researcher must necessarily make difficult tradeoffs between the precision of the theoretical predictions being tested by the experiment and the degree to which the experiment resembles the real world, what one might call external fidelity. For example, agency theory requires that the agent select a level of effort to exert on behalf of the principal given his labor contract and utility function. A highly abstract experimental design which induces the subject's utility function and allows him to select a numerical representation of effort from a menu of options can provide a rigorous test of the theoretical model. Yet choosing a number from a menu is not a close approximation of real-world labor. Alternatively, the experimenter could present the subject with an actual task (like solving a puzzle or constructing simple goods) that requires real effort.

However, this comes at the expense of some control over the experimental environment, in that the experimenter does not know the subject's utility for wealth, disutility for labor or level of skill or effort.

The literature produced by experimental research on agency theory can be roughly divided into two broad categories: experimental economics (EE) literature and management and accounting (MA) literature. Scholars of the EE strain (e.g., Baiman & Lewis (1989), Epstein(1992), Berg, et al. (1992), Guth, et al. (1998), Ghosh & John (2000) and Eriksson & Villeval (2004)) have tended to favor precision and specificity. They typically employ experimental designs that offer very high levels of environmental control: precise model predictions, public knowledge of utility functions, exogenous outcome uncertainty, induced levels of risk aversion via Berg's et al. (1986) lottery wheels, and abstract representations of effort. Scholars of the MA strain (e.g., Chow (1983), Waller & Chow (1985), Dillard & Fisher (1990) and Cadsby, Song & Tapon (2007)) generally favor external fidelity. Subjects are given an actual task to perform, risk preference is either assumed or measured, utility functions are unknown, and outcome uncertainty is often endogenous, in that the only uncertainty is whether the subject will successfully complete his task. A reasonably accurate (though perhaps simplistic) characterization of the two strains of literature is that the EE literature asks how accurately agency theory predicts actual behavior, while the MA literature focuses on whether it is qualitatively accurate. Both strains have found general (though not unqualified) support for the theory.

Given that a major goal of agency theory is to provide a useful tool to managers, we see substantial value in the MA literature's emphasis on external fidelity, even at the expense of formal precision. The average employer is highly unlikely to know the utility function

of her employees, the precise conditional outcome distributions of their effort, or even their reservation wages. Thus, what appears to be lack of rigor to the theorist is likely a valuable précis to the practitioner. Moreover, given our intent to examine the effect of personality traits on effort and contract selection, a design that favors external fidelity is appropriate.

Nevertheless, there are two weaknesses (both raised in Young & Lewis (1995)) in the MA literature that we attempt to address in this paper. First, measures of risk preference have tended to be ad-hoc or nonexistent.<sup>1</sup> Chow (1983), for instance, elicited certainty equivalents for 17 hypothetical lotteries, then summed them to measure each subject's level of risk aversion. Waller & Chow (1985) simply assume risk aversion with no attempt at measurement. In this paper, we implement Holt & Laury's (2002) method (described below) to obtain a more precise risk aversion metric.

Second, experiments in the MA literature have tended to confound effort and skill by measuring only productivity. Subjects are typically instructed to complete as many iterations of a given task as possible in a set amount of time, with output defined as the number of successfully completed iterations. Notice, however, that the number of successful iterations is a function of both the subject's effort *and* his skill. A subject who attempts 10 iterations and is 70% successful is indistinguishable from a subject who attempts 7 iterations and is 100% successful. Thus, a subject's self-selection into an incentive pay contract may reflect either a willingness to exert effort (i.e., a low disutility for effort), a high level of skill, or some combination of the two. Our experimental

---

<sup>1</sup> One notable exception is Cadsby, Song & Tapon (2007), who use the method adopted by this study.

design, in contrast, separates skill from effort by presenting subjects with a set number of iterations per round, and offering the option of attempting to complete the iteration or not. The number of times a subject attempts completion is the measure of his effort. The percentage of attempts that are successful is the measure of his skill.

In addition to these innovations, we look at psychological factors that may have explanatory power for subjects' behavior. We gather measures of personality characteristics in order to consider their possible impact on the agents' behavior. This is not intended as an assault on the assumption of rational self-interest, nor on agency theory. Rather, we are asking whether it is reasonable to assume that employees' utility functions are defined strictly over wealth and effort, and whether personality traits offer explanatory power in addition to agency theory.

## II. Experimental Design

### **Labor**

Labor was modeled using a visual challenge task derived from the “match-to-sample task” used in Houser, et al. (2006). In each iteration of a task, a string of five letters was displayed on subjects' screens for 0.3 seconds, after which they were shown a second string of three letters and asked whether all three of these letters were contained in the first string.<sup>2</sup> Subjects had five seconds to indicate an answer of yes or no by clicking the appropriately labeled button on their screens. Subjects could also click a third button labeled “skip,” which would allow them to move on to the next task without submitting

---

<sup>2</sup> All three- and five-letter strings were randomly generated before a session was conducted. The software used in the experiment constrained the second string of letters so that it always shared at least one letter with the five-letter string.

an answer. If no button was clicked within five seconds, the current task was skipped automatically.

In each round  $t$ , subject  $i$  received a score,  $\sigma_{it}$ , based on his performance in the tasks. For every correct answer,  $\sigma_{it}$  was incremented by one point. Incorrect answers lowered  $\sigma_{it}$  by one. Skipping a task or allowing time to run out left the score unchanged. Each round of the experiment consisted of ten tasks, so that  $\sigma_{it}$  could take any integer value between -10 and 10 inclusive.

Although there was no reason to believe *ex ante* that subjects would enjoy solving the visual challenges (thus violating the assumption of labor disutility), the possibility could not be ruled out. We therefore imposed a monetary penalty to replicate disutility of labor. Subjects were charged a cost  $c = \$0.05$  for every answer of yes or no they submitted. Skipping a task or allowing time to run out incurred no cost. We would therefore expect subjects to submit an answer only when doing so increased their expected payoff.

### **Treatment Variables**

Three different contracts were used to pay subjects for their labor, each of which constituted a treatment variable. For a given round, subject  $i$ 's profit  $\pi_{it}$  took the following functional form:

$$\pi_{it} = \begin{cases} s + \mu(\sigma_{it} - \tau) - lc & \text{if } \sigma_{it} \geq \tau \\ -lc & \text{otherwise} \end{cases}$$

Where:

$s$  is a fixed pay parameter

$\mu$  is a variable pay parameter

$\sigma_{it}$  is  $i$ 's score in round  $t$

$\tau$  is some minimum performance threshold

$l$  is the number of iterations in which the subject submitted an answer

$c$  is the cost of submitting answers

Informally, subjects received some payoff minus the cost of exerting effort so long as their score met some threshold. If they failed to meet this threshold, subjects received no payment for the round, but still bore their costs of labor.

The three contracts were defined by the values given to the parameters  $s$ ,  $\mu$ , and  $\tau$  (see Table 1). Under one contract, payment was entirely dependent on the fixed pay parameter  $s$ . In a second, it was entirely dependent on the variable pay parameter  $\mu$ . In the final contract, both  $s$  and  $\mu$  played a role in determining the payoff. These three plans are analogous to salary, piece rate, and salary plus bonus compensation. However, to avoid overtly triggering mental biases among the subjects, the experimental instructions referred to the contracts as “Threshold,” “Multiple,” and “Combo” respectively.

In the Threshold treatment,  $s = \$1.15$ ,  $\mu = 0$  and  $\tau = 3$ . That is, a subject would earn \$1.15 (minus his induced disutility of labor) so long as his score was at least three at the end of the round. A subject would earn no payment, and still bear the disutility of labor,

if he scored below three. The Threshold plan emulated the incentives commonly presented to salaried employees: subjects' optimal level of effort would be just enough to achieve the minimum threshold. Additional effort would incur additional costs with no chance of a monetary gain.

In the Multiple treatment, the parameters were set at  $s = 0$ ,  $\mu = \$0.20$ , and  $\tau = 0$ . In this treatment, *all* of a subject's payment was variable. A subject's earnings at the end of a round were simply  $\$0.20$  multiplied by his score (provided he had a positive score). The cost of submitting answers was subtracted from this amount to determine the round profit. The Multiple contract mirrors purely piece rate (or commission-based) compensation plans. A subject should submit an answer so long as the expected utility of a correct answer exceeds the marginal cost of labor.

For the Combo treatment,  $s = \$0.90$ ,  $\mu = \$0.12$ , and  $\tau = 3$ . As with the Threshold treatment, subjects received a fixed payment for achieving a score of three. Additionally, similar to the Multiple treatment, subjects could also receive a variable payment for exceeding the minimum performance threshold.

Table 1 displays the minimum and maximum  $\pi_{it}$  in each treatment. Under any labor contract a subject's lowest earnings per round would be  $-\$0.50$  for submitting 10 incorrect answers. In the Threshold treatment, a subject could earn up to  $\$1.00$ . This would require submitting three accurate answers and skipping all subsequent questions. In the Combo and Multiple treatments, subjects would maximize earnings by submitting

10 correct answers. Payment for three correct answers in these treatments were substantially lower (\$0.75 in Combo and \$0.45 in Multiple), but they also offered significantly greater rewards for more effort, provided the answers submitted were correct. Per round earnings in the Combo and Multiple treatments had the potential for an increase over those in the Threshold treatment by 24% and 50% respectively.

The visual challenge tasks were organized into “sections.” Each section consisted of five rounds under the same compensation plan. Subjects were required to complete one section of each treatment. Because subjects were assigned the payment method in these sections, we refer to them as the “compulsory” sections. The order in which each subject encountered the treatments was randomized to avoid bias due to ordering effects.

### **Risk Aversion**

Risk aversion was measured via a method developed by Holt & Laury (2002). Subjects were shown a series of ten lottery pairs. Both lotteries in a lottery pair consisted of a high and low payoff,  $\pi_H^k$  and  $\pi_L^k$  respectively, where the superscript  $k \in \{1,2\}$  indicates the lottery. The  $\pi_H^k$  and  $\pi_L^k$  were chosen so that Lottery 1 was the less risky choice, with  $\pi_H^1 < \pi_H^2$  and  $\pi_L^1 > \pi_L^2$ . The payoffs for each lottery remained constant in every pair in the series.

For a given lottery pair in the series, both lotteries shared a common probability for the high and low payoffs. The probability of realizing the high payoff was  $p_H$ , with a

corresponding probability of realizing the low payoff equal to  $1 - p_H$ . For the first pair in the series,  $p_H = 0.1$ . In each subsequent pair  $p_H$  was increased by 0.1, so that the high payoff was a certainty by the tenth pair.

Subjects were told to indicate their preferred lottery in each pair. After submitting their preferences, one of the pairs was selected at random and the lottery in that pair that the subject had indicated a preference for was executed for a cash payment.

In the first four pairs, Lottery 1 has the higher expected payoff (see Table 2). From the fifth pair on the expected payoff advantage belongs to Lottery 2. The lottery pair at which a subject began to prefer Lottery 2, or his “crossover point,” provided a measure of risk preference. A risk neutral subject would have a crossover point of 5. Crossing over prior to the fifth lottery pair indicates risk proclivity, while crossing over at some point after the fifth lottery pair indicates risk aversion. Subject to the assumption that subjects’ utility functions for money,  $u(\pi)$ , follow the form  $u(\pi) = \pi^{1-r}$ ,<sup>3,4</sup> ranges for the risk preference parameter  $r$  can be calculated. These are contained in Table 3 with an informal risk preference classification for each range.<sup>5</sup>

---

<sup>3</sup> For  $r > 1$  the function is divided by  $(1 - r)$  to preserve increasing utility in  $\pi$ . For  $r = 1$  the functional form becomes  $u(\pi) = \ln(\pi)$ .

<sup>4</sup> Holt and Laury report increasing risk aversion as the payoffs are scaled up, indicating that this functional form is not strictly accurate. However, the more complex method of measurement they develop is both expensive and more conducive to determining the average risk preference of a group than estimating it at the individual level. Moreover, within the range of payoffs in this experiment the utility function derived by their complex method is extremely close to the assumed function.

<sup>5</sup> The contents of the range of  $r$  and risk preference classification in Table 3 come chiefly from Table 3 in Holt & Laury (2002). Note, however, that for our purposes the crossover point is the point of reference, while Holt and Laury focus on the number of less risky lotteries chosen. Choices among our subject pool are highly consistent with those reported in Holt & Laury (2002).

In subjects' instructions this portion of the experiment was referred to as the "lottery section." This section was presented randomly among the three visual challenge sections to avoid ordering effects.

### **Elective Section**

After completing all three visual challenge sections and the lottery section, a display on the subjects' screen showed them the total profits they had earned under each of the three compensation plans. Subjects were then required to complete a fourth visual challenge section, this time under the contract of their choice. Earnings from this self-selected or "elective" section were added to the subjects' total payment.

This portion of the experiment replicates agents' selection of a contract. *Ceteris paribus*, agency theory predicts that risk preferring, high skilled subjects will self-select into one of the incentive-pay contracts, while risk averse, low skilled subjects will self-select into the Threshold contract. Recording subjects' choice of contract in addition to their risk preferences and levels of skill allows us to test this prediction.

### **Psychological Inventories**

In addition to the visual challenge and lottery sections, subjects were required to complete an online psychological profile known as the Watterson Personality Inventory (WPI). The WPI consists of 291 questions posed in the form of single sentence self-evaluative statements. For each statement, the subject must indicate that the statement is true or false with regards to himself or indicate that he is not certain. The results of the

WPI generate in excess of 40 measures of personality, but we focus on five “Global Factors,” which are analogous to the personality dimensions measured in the Five Factor Model (FFM) of personality. The FFM has been studied by organizational psychologists for more than 40 years and is generally accepted to be the most useful comprehensive taxonomy of personality (see McCrae & Costa, 1987). Table 4 displays the Global Factors and the implications of a high or low score in each. The Global Factors are measured by sten scores, which are bounded by 1 and 10 inclusive and distributed  $N(5.5,2)$ . An intuitive description of each Global Factor follows:

- **Extraversion** – Represents an individual’s response to social settings, especially large groups. Those with high Extraversion scores tend to gather energy from interpersonal activity, and are likely to be assertive and talkative in group settings. Those with low Extraversion scores tend to be drained of energy from interpersonal activity. They are more likely to be reserved in group settings.
- **Emotional Resilience** – Represents an individual’s emotional stability and ability to adjust to the unexpected. High Emotional Resilience is correlated with a relaxed, confident attitude. Low Emotional Resilience indicates self-doubt and anxiety.
- **Self-Control** – Represents an individual’s attitude and drive toward accomplishing a goal. Individuals who score high on Self-Control are likely to be highly organized, persistent, and goal-oriented. Individuals who score low on Self-Control are likely to be disorganized, less focused, and impulsive.
- **Independence** – Represents an individual’s attitude toward others in interpersonal transactions. Highly Independent individuals tend to be assertive,

suspicious of the motives of others, and focused on achieving their own ends.

Individuals who score low on Independence exhibit high levels of trust, forgiveness and altruism.

- **Practicality** – Represents whether an individual's thinking style is dominated by a factual or emotional orientation. Practical individuals may be characterized as conventional, and having mainstream interests. They have a tendency to assess situations with set formulas. Individuals who are less practical can be described as innovative, nonconformist and imaginative. They are more likely to follow “inspiration” than analytical reasoning when making decisions.

Because there is no precedent for psychological measures in the agency theory literature of which we are aware, we consider all five Global Factors in our statistical analysis. The results may serve as a baseline for future research in this area.

## **Procedures**

All experiments were conducted at the Computer Laboratory for Experimental Economic Research at Wichita State University. The subject pool consisted of 80 subjects composed of undergraduate and graduate students as well as a single recent alumna. 60 of the subjects were recruited from economics and accounting classes. Word of mouth from this subject pool attracted an additional 20 subjects who were members of a business fraternity on campus.

With the exception of 10 subjects, the WPI was administered on a different day than the experiment.<sup>6</sup> 39 subjects completed the experiment prior to filling out the WPI. Of these, four did not return on the second day to complete the WPI. 41 subjects had the WPI administered first. One of these subjects did not return to complete the experiment.

On the day of the WPI, subjects were told that they would be filling out a psychological inventory and to work at their own pace. Subjects would typically complete the WPI in 30 – 40 minutes, with a few subjects taking an hour. They were paid \$15 for their time.

On the day of the experiment, subjects read through a set of electronic instructions. After the written instructions, the lab monitor read through a summary of the rules. Subjects were encouraged to ask clarifying questions at any time during the instruction period. Once the summary had been read, subjects completed between one and three practice rounds of the visual challenge before continuing on to the actual experiment.<sup>7</sup> The practice rounds were conducted solely to familiarize subjects with the software interface and had no direct impact on their earnings. Subjects were paid \$5 for participating in addition to their earnings from the experiment. Excluding this \$5 and the \$15 for filling out the WPI, subjects earned an average of \$15.38. Given that most subjects spent 35 – 45 minutes in this portion of the experiment, we consider these earnings salient. (The range of possible earnings was \$0.00 to \$30.05).

---

<sup>6</sup> These 10 subjects participated in the experiment near the end of the semester, and therefore time constraints prevented them from completing the two parts on separate days.

<sup>7</sup> All subjects were required to complete one practice round. They were allowed to complete up to two more, but could continue on to the experiment at the end of any practice round. Following each practice round a subject was shown what his payment would have been under each of the three payment plans given his score.

### III. Hypotheses

Our experimental design allows us to test a number of hypotheses. First, the Threshold contract design should incent subjects to submit only enough answers to achieve a score of 3. Although we have no *ex ante* theoretical prediction for effort under the Combo and Multiple contracts, a reasonable hypothesis is that the greater the reliance of  $\pi_{ir}$  on the variable component  $\mu$ , the greater the level of effort exerted. Thus we propose hypothesis H<sub>1</sub>: Subjects will exert more effort in the Multiple treatment than the Combo treatment, and will exert more effort in the Combo treatment than the Threshold treatment.

Second, agency theory predicts that variable pay contracts will attract high-skilled employees who know that their expected earnings are higher with variable pay than fixed pay. We therefore propose hypothesis H<sub>2</sub>: Subjects will tend to self-select into contracts on the basis of skill, with higher skilled subjects choosing the Combo and Multiple contracts and lower skilled subjects choosing the Threshold contract. Agency theory further predicts a similar self-selection of contracts based on risk aversion, as variable pay contracts offer a more risky income stream. This suggests hypothesis H<sub>3</sub>: Subjects with higher levels of risk aversion will tend to self-select into the Threshold contract, while risk-neutral and risk-preferring subjects will self-select into the Combo and Multiple contracts.

Third, we may consider whether incentive pay is effective in maximizing profit to the firm. Our design makes no assumption about the value of the output generated by employees. However, we can measure the average cost per unit (ACPU) of production of each subject in each treatment. Thus we propose hypothesis H<sub>4</sub>: The mean ACPU in the Threshold treatment will exceed the mean ACPU in both the Combo and Multiple treatments.

One could propose any number of hypotheses with regards to the results of the psychological profiles. For example, highly Self-Controlled subjects' tendencies toward perfectionism may lead them to exert extra effort regardless of their contract. Subjects who score low in Emotional Resilience may prefer fixed pay contracts due to heightened anxiety. However, as this research can be considered an exploratory experiment of the effects of personality traits on the principal-agent problem, we propose only two broad hypotheses: H<sub>5</sub>: Psychological characteristics will be a factor in the effort exerted by subjects; and H<sub>6</sub>: Psychological characteristics will be a factor in the contracts selected by subjects.

## IV. Results

### **Compulsory Sections**

An informal assessment of the data from the compulsory sections reveals a number of stylized facts with respect to the level of effort subjects exerted under each payment method.

Figure 1 displays the distribution of effort by payment treatment across all rounds and all subjects. The Combo and Multiple payment methods clearly elicited high levels of effort, with modal effort at 10 answers in both treatments. Of interest is the fact that there is no obvious distinction between the two distributions, suggesting that a  $\mu$  parameter of \$0.12 was equally effective at incenting maximum effort as a  $\mu$  parameter of \$0.20. Given the large (67%) disparity in the value of a marginal increase in  $\sigma$  under the two payment methods, this is surprising.

Also surprising is the distribution of effort in the Threshold treatment, which is roughly bimodal. In just over 30% of rounds, subjects answered only three questions, which is consistent with the agency theory prediction for a perfectly accurate employee. In an additional 15.7% of rounds, subjects submitted either 5 or 7 answers, which is consistent with a profit-maximizing agent who submits 1 or 2 incorrect answers respectively. However, in the greatest percentage of rounds (38.7%), subjects submitted an answer to every question. The distribution is virtually identical if we consider only attempts made in round 5 of each treatment (31.6% of subjects answered 3 questions, 40.5% answered 10 – see Figure 2). Thus, learning effects appear to be absent.

One might suppose that low levels of subject accuracy can account for the unexpectedly high levels of effort in the Threshold treatment, but the distribution of final scores (contained in Figure 3) discredits this hypothesis.<sup>8</sup> This distribution is roughly normal, and is clearly centered on 3, which accounts for 47.8% of its mass. However, the right

---

<sup>8</sup> Moreover, given that the score is incremented or decremented each time an answer is submitted, an even number of answers submitted implies an even score. Therefore, submitting 10 answers could never result in a score of 3 and is ruled out by agency theory.

tail of the distribution is quite thick, accounting for 38.2% of the mass. That is, nearly 40% of the time, subjects submitted answers when the expected value of doing so was strictly *negative*. Standard agency theory fares little better when we restrict the sample to round 5 scores. In this case, 49.4% of subjects ended the round with a score of 3 and the right-hand tail accounts for 35.4% of rounds.

The bimodal distribution of effort under the Threshold payment method leads us to suspect that the subject pool was comprised of two distinct types of subject, each with their own incentive disposition. The first type, which we will refer to as incentive responsive (IR), exerted effort only when they had a monetary incentive to do so. The second type, which we will refer to as incentive indifferent (II), exerted effort essentially without regard to their monetary incentives.

Examining the subject-level data allows us to investigate this hypothesis. We use the following rule set to classify each subject as IR or II. A subject who skipped all remaining questions after achieving a score of 3 in the fifth round of the compulsory Threshold section was categorized as IR. If a subject failed to achieve a score of 3 before the 10<sup>th</sup> question of the fifth round, then classification was based on behavior in round four.<sup>9</sup> With one exception, all subjects who could not be categorized as IR in this way were categorized as II.<sup>10</sup>

---

<sup>9</sup> In no instance was it necessary to consider behavior in round three.

<sup>10</sup> In the single exception, the subject submitted two answers after achieving a score of three in round five, one of which was correct, one incorrect. The subject then skipped all remaining questions. His behavior in round four was perfectly consistent with IR behavior. He was therefore categorized as IR for purposes of analysis.

Using this rule set, the subject pool was almost evenly split, with 51.25% of subjects categorized as IR versus 48.75% categorized as II. In addition, behavior patterns in the Threshold section were remarkably consistent across periods. Of those categorized as IR based on final round behavior, only 19.5% failed to display IR behavior from the very first round of the section. Only 9.8% failed to exhibit IR behavior by the second round. Thus, the overwhelming majority of subjects either behaved in accordance with agency theory for the entire compulsory Threshold section or failed to behave in accordance with it at all.

Before turning to what may account for the high level of II behavior, we first use regression analysis to statistically test our hypothesis that subjects can be broadly categorized into two distinct groups. The model characterizes subject behavior under each of the compulsory treatments, given that some subjects were IR. It is specified as:

$$\begin{aligned}
 Effort_{it} = & \alpha + \beta_1 Threshold + \beta_2 IR * Threshold + \beta_3 Combo + \beta_4 Accuracy_{it} \\
 & + \beta_5 Threshold * Accuracy_{it} + \beta_6 IR * Threshold * Accuracy_{it} \\
 & + \beta_7 Combo * Accuracy_{it} + \gamma_i + \varepsilon_{it}
 \end{aligned} \tag{1}$$

where  $Effort_{it}$  is the number of answers submitted by subject  $i$  in round  $t$ ,  $Threshold$  and  $Combo$  are dummy variables indicating the payment treatment (with the Multiple treatment as the baseline),  $IR$  is a dummy variable indicating whether or not subject  $i$  is categorized as incentive responsive, and  $Accuracy_{it}$  is the percent of answers submitted in round  $t$  that were correct. (The  $Accuracy_{it}$  variable is scaled between 0 and 1, so that the

reported coefficient indicates the marginal effect of perfect accuracy.) The  $\gamma_i$  term is a subject-specific random effect.<sup>11</sup> The error term was corrected for an AR(1) process. Results from (1) are displayed in .

The statistical analysis confirms our suspicions from informal observation of the data. In the Multiple treatment, subjects' baseline effort was 9.5 answers ( $p < 0.001$ ), with the level of accuracy having no distinguishable impact. The insignificant coefficients for *Combo* and *Combo \* Accuracy<sub>it</sub>* indicates that this behavior remained unchanged under the Combo payment method. Likewise, *Threshold* and *Threshold \* Accuracy<sub>it</sub>* have coefficients that are indistinguishable from zero. Only incentive responsive subjects in the Threshold treatment deviate from this pattern of maximum exertion of effort. The estimated coefficients on *IR \* Threshold* and *IR \* Threshold \* Accuracy<sub>it</sub>* are 4.88 and -11.03 respectively ( $p < 0.001$  in each case). A Wald test cannot reject the null hypothesis that a perfectly accurate incentive responsive subject will submit exactly three answers ( $p = 0.282$ ).

These results offer only weak evidence to confirm hypothesis H<sub>1</sub>. Incentive pay does elicit more effort than fixed pay for roughly half of our subject sample, but has no effort effect on the remaining half. Moreover, the Multiple and Combo pay contracts elicit indistinguishable levels of effort across the entire sample.

---

<sup>11</sup> A Hausman test fails to reject the null hypothesis of equivalent coefficients between a fixed- and random-effects specification at the 5% confidence level, but only by a slim margin ( $p = 0.0637$ ). Despite the marginal significance of the Hausman test we are comfortable using the results of the random-effects model, as the estimates from the fixed- and random-effects specifications are qualitatively equivalent.

It is tempting to disregard this failure among a large portion of subjects to exhibit agent theoretic behavior as subject confusion, but such an explanation seems implausible. Subjects' instructions explicitly stated that scoring above 3 in a Threshold round would not improve their earnings, and that every answer submitted cost \$0.05. This was stated explicitly again during the oral summary. Moreover, in every round of the experiment a reminder of the rules by which the subjects' scores would determine their payoffs in that round were displayed at the bottom of the screen. Yet in *every* session some subjects exerted maximum effort in the Threshold section.<sup>12</sup> Subject confusion seems insufficient to explain II behavior among *nearly half* of all participants.

An alternative (or at least parallel) explanation is that II subjects were not motivated strictly by their payoffs. These subjects may have been motivated by a desire to achieve a high score. They may have found the dissatisfaction of leaving a task incomplete (by skipping it) sufficiently distasteful to bear the cost of submitting answers with no expectation of monetary reward. They may have been so caught up in the task-solving level of the experiment that they failed to act strategically when their payoffs were dependent only on achieving a minimum threshold. In short, subjects had ample opportunity to act on the basis of non-pecuniary stimuli, and they may have taken it.

In order to test this hypothesis, we specify a logistic regression model with incentive responsiveness as the dependent variable. In order to account for possible differences

---

<sup>12</sup>This trend held up even in one session in which a subject announced publicly, during the oral summary, that one's optimal strategy in the Threshold section was stop submitting answers after achieving a score of 3.

across gender we specify the self-explanatory dummy variable *Male*. We also include the five Global Factors as independent variables.<sup>13</sup>

Our experimental design also allows us to include a variable which captures the possibility of subject confusion. Recall that in the Holt and Laury risk aversion measurement subjects should choose a single crossover point above which they prefer Lottery 1 and below which they prefer Lottery 2. While the majority of subjects did this, some crossed over between the lotteries multiple times, suggesting that they failed to understand their incentives. A handful also indicated a preference for Lottery 1 in all cases, even though in the tenth lottery pair they were guaranteed the high payoff and  $\pi_H^1 < \pi_H^2$ . We constructed a dummy variable, *Comprehend*, which takes a value of 1 if the subject expressed preferences in the lottery section roughly consistent with expected utility theory and takes a value of 0 otherwise (see Appendix). We include this variable in the logistic regression model to account (at least partially) for the possibility of subject confusion.

The results are contained in Table 6. The positive and significant coefficient for the *Comprehend* variable ( $p = 0.005$ ) suggests that subject confusion was indeed a factor in generating II behavior among the subjects. Participants whose behavior in the Lottery section was roughly consistent with expected utility theory were more likely to be incentive responsive.

---

<sup>13</sup> The sten scores contained in the WPI results were converted to standard deviations by subtracting 5.5 and dividing this difference by 2. The estimated coefficients of the Global Factors can therefore be interpreted as the marginal effects of a one standard deviation change from the mean in the personality trait of interest.

Personality traits also affected subjects' incentive dispositions. Two of the five Global Factors have coefficients that are at least marginally significant, confirming hypothesis H<sub>5</sub> that personality traits affect effort. Subjects who had low Emotional Resilience were less likely to be incentive responsive ( $p = 0.025$ ). Recall that low levels of Emotional Resilience are correlated with anxiety and self-consciousness. It is likely that such subjects did not trust themselves to understand the experimental environment sufficiently to be comfortable exerting minimal effort. They may have therefore answered every question in an attempt to insure against this possibility.

The coefficient for *Practicality* is marginally significant ( $p = 0.063$ ) and positive, indicating that more Practical subjects may have been more likely to be IR. Although we cannot quite be 95% confident, this result makes a kind of intuitive sense. Practical individuals tend to assess situations with straightforward, utilitarian formulas. Such an assessment of the payment institution of the Threshold section should logically lead one to the conclusion that answering questions once a score of 3 has been achieved is wasted (and costly) effort.

### **Elective Section**

Agency theory predicted subject behavior substantially better in the elective section than in the compulsory sections.

Figure 4 displays the distribution of effort by treatment for all rounds of the elective section. As in the compulsory sections, effort under the Combo and Multiple contracts substantially exceeds effort under the Threshold contract. Moreover, subjects who self-selected into the Multiple payment method appear to have exerted slightly more effort than their Combo counterparts.

The dramatic reduction in incentive indifferent behavior in the Threshold treatment also strongly supports agency theory. In 42.9% of all rounds subjects submitted the absolute minimum number of answers to receive their payment. In only 19% of rounds did subjects exert maximum effort. The evidence is even more stark when we look at the distribution of final scores (see Figure 5). Subjects' behavior resulted in a score of 3 in 71.4% of rounds of the Threshold section. Scores exceeded 3 in only 20% of rounds.<sup>14</sup> This behavior is reflected in the fact that 71.4% of subjects (15 of 21) in the Threshold treatment of the elective section can be categorized as IR.

There are two possible explanations for the superior predictive power of agency theory in the elective Threshold section. It may reflect learning effects, as all subjects who chose to repeat the Threshold section were experienced in the payment method. Alternatively, it may reflect selection effects, in that subjects who are predisposed to exert maximum effort are less likely to self-select into a payment method that does not reward them for doing so.

---

<sup>14</sup> Behavior was almost completely consistent across rounds. We therefore display results from the entire elective section, rather than from the fifth round alone.

Observation of the subject-level data leads us to conclude that learning effects have very little explanatory power. Of 15 subjects who are categorized as IR in the elective Threshold section, only two were II in the compulsory section.<sup>15</sup> Thus, 86.7% of the agent theoretic behavior in the elective sections can be attributed to the selection effect. Moreover, statistical analysis demonstrates that, holding incentive disposition constant, subject behavior is unaffected by self-selection. Fitting data from the elective section to (1) yields results that are nearly identical to those from the compulsory sections (see Table 7).<sup>16</sup> The baseline effort under the Multiple payment method was 9.59 attempts ( $p < 0.001$ ), and a subject's accuracy in the round had no significant impact. Behavior was not significantly different in the Combo treatment, nor in the Threshold treatment provided the subject was II. However, IR subjects' effort was negatively correlated with their accuracy in a given round. A Wald test cannot reject the hypothesis that IR subjects with perfect accuracy submitted the required three answers to receive payment ( $p = 0.2972$ ).

We conclude that hypothesis  $H_1$  is better supported under employee-selected contracts than imposed contracts. Close to three quarters of subjects who chose to repeat the Threshold section behaved exactly as agency theory would predict. However, there is still no statistically significant difference in effort levels elicited by the Multiple and

---

<sup>15</sup> One of these two subjects could almost have been categorized as IR in the compulsory Threshold section. In that section, she consistently achieved a score of four before skipping the remainder of the questions. Curiously, two subjects who were IR in the compulsory sections exhibited II behavior in the elective section.

<sup>16</sup> A Hausman test indicates that random effects are justified ( $p = 0.9975$ ). The fixed-effects model used for comparison in the Hausman test constrained the fixed effect for three of the subjects to be zero in order to avoid perfect multicollinearity between the fixed effects and the treatment dummies.

Combo contracts. Within the range of the variable pay parameter used in this experiment ( $\mu \in \{\$0.12, \$0.20\}$ ), the amount of pay at risk is inconsequential.

### **Contract Selection**

A visual examination of the data suggests that agency theory was of little value as a predictor of contract selection. Figure 6 displays the distributions of skill level conditional on contract selection in the elective section. All three distributions are centered around 75% accuracy and are roughly normal, though the distribution for subjects who self-selected into the Multiple payment method is somewhat skewed to the right. Self selection does not appear to have separated subjects by skill.

Two stylized facts surfaced while conducting the experiments. First, subjects seemed to have a preference for a salary plus bonus contract. Of 80 subjects, 37 (46%) chose to repeat the Combo section. The remaining 43 subjects were split virtually evenly: 21 subjects (26%) repeated the Threshold section and 22 (28%) repeated the Multiple section. Second, subjects very often seemed to make their contract selection largely on the basis of previous earnings. That is, subjects often chose to repeat that payment method under which they had earned the greatest amount during the compulsory sections.

In order to statistically analyze the relevant influences on contract selection, we fit the data to three logit models, each with the selection of one of the payment methods as the binary dependent variable. In order to test agency theory directly, we included risk preference and skill. The former was represented by the subject's crossover point in the

lottery section.<sup>17</sup> The latter was measured by the percentage of answers submitted that were correct in the practice rounds and compulsory sections. To test whether subjects' previous earnings under a payment method affected their decision, we included his normalized earnings. If we define  $\theta$  to be the treatment serving as the dependent variable and  $\bar{\theta}$  to be the more profitable of the two treatments not serving as the dependent variable, then we can express subject  $i$ 's normalized earnings for  $\theta$ ,  $\Pi_i^\theta$ , as

$$\Pi_i^\theta = \sum_{t=1}^5 (\pi_{it}^\theta - \pi_{it}^{\bar{\theta}}).$$

Finally, we included a gender dummy (*Male*) and the Global

Factors<sup>18</sup> to capture any effects of gender or personality traits.

The results (displayed in Table 8) provide only minimal support for agency theory. The hypothesis that agents of higher skill levels will self-select into incentive pay contracts (H<sub>2</sub>) is not supported. The coefficient for *Accuracy* is marginally significant in the Combo model ( $p = 0.059$ ), but is of the opposite sign predicted by theory, suggesting that higher skilled subjects did not prefer that method of incentive pay. The *Accuracy* coefficient is not significant at any conventional level of confidence in the Threshold and Multiple models.

The hypothesis that agents will self-select into fixed and incentive pay contracts on the basis of risk aversion (H<sub>3</sub>) is weakly supported. The coefficient for *Crossover* is negative and significant ( $p = 0.028$ ) in the Combo model and positive and marginally significant ( $p = 0.069$ ) in the Threshold model. This suggests that risk neutral and risk preferring

---

<sup>17</sup> Eleven subjects whose choices were highly incompatible with expected utility theory are excluded from the analysis.

<sup>18</sup> Stens were converted to standard deviations for the analysis.

subjects tended to repeat the Combo section, while risk averse subjects may have favored the Threshold section. However, the coefficient for *Crossover* is not remotely significant ( $p = 0.912$ ) in the Multiple model.

This failure to find substantial support for the contract selection implications of agency theory is surprising. We note that other scholars have findings opposite to our own (see Eriksson & Villeval (2004) and Cadsby, Song & Tapon (2007)) and thus encourage the reader not to consider our results to be definitive. They must, instead, be considered in the light of the experimental agency theory literature in general.

Our two stylized facts are confirmed by the statistical analysis. The constant term in the Combo model is positive and significant ( $p = 0.016$ ) and indistinguishable from zero in the other models, indicating that subjects had a bias toward choosing the salary plus bonus pay contract. This is not without precedent. Waller & Chow (1985) and Ghosh & John (2000) report results in which contracts with both fixed and variable pay elements were chosen more often than purely fixed or variable contracts. It could be that subjects see a mixed payment method as a “best of both worlds” contract, as the fixed component acts as something of an insurance policy against poor performance and the variable pay offers a chance to boost earnings.

Our intuition that subjects were basing their contract selection on the comparative profitability of the compulsory sections is strongly supported. The *NormEarnings*

coefficient is positive and highly significant in all models.<sup>19</sup> In fact, for the Multiple model, *NormEarnings* is the *only* variable with any significant explanatory power whatsoever, suggesting that past earnings dominated all other considerations for those subjects willing to accept a piece rate contract. The general tendency of past results to influence contract selection decisions suggests that subjects had difficulty predicting earnings under the three payment methods on the basis of their demonstrated skill level.

Gender played no role in guiding contract selection, but personality traits did in the case of Threshold and Combo contracts, therefore  $H_6$  is supported. Subjects with high Self-Control scores tended not to repeat the Combo section ( $p = 0.049$ ), but this result does not lend itself to easy interpretation given that the *Self-Control* variable is not significant in the Threshold or Multiple models. Subjects with a high Practicality score were significantly less likely to select fixed pay only ( $p = 0.033$ ) and significantly more likely to select salary plus bonus ( $p = 0.014$ ). Recall that Practical subjects (tending to be more utilitarian) were also more likely to be IR, exerting the minimum effort possible to earn their payment.

To the extent that the results of this experiment are externally valid, this suggests that firms with Practical agents and a fixed pay policy are likely to see substantial gains from offering incentive pay contracts. Their employees are likely exerting minimal effort under the status quo, but are also likely to opt into a more incentive compatible payment method.

---

<sup>19</sup> For the Threshold, Combo and Multiple models,  $p = 0.006$ ,  $p = 0.001$  and  $p = 0.011$  respectively.

The coefficient for *Emotional Resilience* is positive and marginally significant ( $p = 0.077$ ) in the Combo model, suggesting that subjects with relaxed, confident dispositions tended to prefer a salary plus bonus contract. Because low Emotional Resilience scores are significantly correlated with IR behavior, the results for *Emotional Resilience* and *Practicality* in the Combo model suggest that IR subjects are drawn to the Combo contract. One possible explanation for this is that high Emotional Resilience makes such subjects confident that they can achieve a score in excess of 3 and high Practicality makes them aware that even in the event of poor performance the fixed pay component will serve to cut their losses. However, given that the *Emotional Resilience* coefficient is not significant at the 5% level we are cautious with this interpretation.

### **Cost Effectiveness of Incentive Pay**

We measure the cost effectiveness of the treatments by calculating subjects' average cost per unit (ACPU) of production. For subject  $i$  in treatment  $\theta$ , the ACPU is defined as

$$\frac{\sum_{i=1}^5 \rho_{it}^{\theta}}{\sum_{i=1}^5 \frac{\sigma_{it}^{\theta}}{5}}, \quad (2)$$

$$\text{where } \rho_{it}^{\theta} = \begin{cases} s^{\theta} + \mu^{\theta}(\sigma_{it}^{\theta} - \tau^{\theta}) & \text{if } \sigma_{it}^{\theta} \geq \tau^{\theta} \\ 0 & \text{otherwise} \end{cases}$$

Informally, the ACPU is  $i$ 's average compensation in  $\theta$  divided by his average score in  $\theta$ .

The distributions of subjects' ACPU in the compulsory sections are displayed by treatment (and incentive disposition in the case of the Threshold treatment) in Figure 7. The distributions center around \$0.20 in the Combo and Multiple treatments, suggesting that the two payment methods are equally cost effective. If subjects in the Threshold treatment generated a similar ACPU, averaging \$0.197. The distribution among IR subjects in the same treatment, however, centers around \$0.38, and exhibits virtually no overlap with the other treatments. Results in the elective section are similar to those in the compulsory section.

The mean ACPU for all treatments across both assignment methods is displayed in Table 9 along with pair-wise differences in mean ACPU. T-tests of the pair-wise differences render three main results. First, holding incentive disposition constant, cost effectiveness is stable across contract assignment methods. No difference in mean ACPU across assignment methods within a treatment exceeds \$0.0426, and none is significant at conventional levels, even using one-tailed tests. We may therefore conclude that subject behavior was directed by their incentive disposition and the incentives that they faced. Learning effects and self-selection effects (holding incentive disposition constant) had no influence.

Second, we may conclude that, given our experimental parameters, the Combo and Multiple payment methods are equally cost effective. The mean ACPU in the Combo and Multiple treatments under compulsory assignment were \$0.2039 and \$0.2026 respectively, and statistically indistinguishable. Under elective assignment the respective

mean ACPUs were \$0.2465 and \$0.2151. While the difference between treatment mean ACPUs of \$0.0314 is relatively large (nearly 15% of the mean ACPU in the Multiple treatment), it is not significant at the 5% confidence level ( $p > 0.20$ ). This does not imply that it would be impossible to find values for the salary and incentive pay parameters that would support a significant difference between the two contracts. However, an employer in our experimental environment should be indifferent between them.

Finally, it is clear that the comparative cost effectiveness of the incentive pay contracts relative to the Threshold contract is dependent on the subject's incentive disposition. II subjects in the Threshold treatment had a mean ACPU of \$0.1971 under compulsory assignment and \$0.1905 under elective assignment. This is lower in both cases than the corresponding variable pay contracts, though not at conventional significance levels. IR subjects, on the other hand, generated ACPUs roughly double those of the incentive pay methods: \$0.3772 in the compulsory sections and \$0.403 in the elective section. In fact, the IR Threshold mean ACPUs are significantly higher than any possible comparison mean ACPU at very high levels of confidence ( $p < 0.01$  in two cases,  $p < 0.001$  in ten cases).

Recall that hypothesis  $H_4$  predicted that the variable pay contracts would have a lower mean ACPU than the Threshold contract. In the compulsory (elective) sections,  $H_4$  is strongly upheld for roughly half (three quarters) of subjects, but strongly rejected for the remaining half (one quarter).

This implies that, under the Threshold payment method, employee selection is paramount. Employers hiring IR employees would be at a severe cost disadvantage without some form of incentive pay. Employers who could reliably hire II employees could achieve low levels of ACPU with a simple salary compensation plan. Given subjects' overwhelming adherence to agency theory in the elective sections, however, it seems unlikely that firms offering salary only compensation could attract large numbers of II employees.

## V. Summary & Conclusions

Agency theory predicts that pay contracts with a variable component will increase a firm's productivity by giving employees an incentive to exert more effort and by attracting highly skilled workers while encouraging low-skilled workers to find employment elsewhere. It also predicts that risk averse employees will be attracted to firms that offer fixed pay contracts, while firms that offer variable pay contracts will attract risk neutral and risk loving employees.

Our experimental research confirms that incentive pay contracts will indeed elicit higher levels of effort from a large portion of employees. For many, such incentives may be unnecessary when fixed wages are the only option. But when employees are allowed to select among alternatives, those who choose fixed wages are likely to be those who will exert minimal effort in accordance with agency theory.

We are unable to find evidence that offering different levels of variable pay leads employees to self-select by skill level, and only weak evidence that it leads them to self-select by risk preference. We emphasize, however, that previous work by other scholars does find such evidence. We therefore do not consider our findings on these two hypotheses to be conclusive.

We consider the following three results from this research to be of the greatest potential value to firms and future research. First, it is clear that some personality traits have explanatory power with regard to the level of effort individuals will exert and the payment methods into which they will self-select. This implies that psychological evaluations may be useful tools in considering what types of pay contracts should be offered to employees. It also implies the reverse: that pay contracts may be a useful tool in attracting employees with personality traits that are beneficial beyond their impact on effort. In our experimental environment, for instance, the Combo payment method attracted employees with more confident, relaxed attitudes and utilitarian thinking styles. These traits may be measures of skill (or potential skill) in certain industries. We welcome further research into both of these implications.

Second, our results suggest that, at least within a certain range, the fact that compensation is offered at the margin may be more important than the magnitude of such compensation. Subjects did not systematically exert less effort in the Combo section than the Multiple section, despite the substantial difference in relative marginal compensation. If this is true outside the laboratory as well, managers need not be overly concerned with

adjusting each employee's labor contract to match his utility function. General rules of thumb may be equally effective.

Finally, our subjects tended to be backward looking when choosing a pay contract, very often selecting the contract that had been most beneficial to them in the past. This suggests that demonstrating the effectiveness of a compensation plan in increasing employees' pay may be crucial in convincing them to voluntarily adopt it. To this end, it may be advantageous for firms to use the financial results of early adopters of variable pay as a positive example to reticent employees. Another strategy may be to give employees "shadow" paychecks showing them the compensation they would earn if they had selected a contract with a variable pay component.

## Appendix: Taxonomy of Violations of Expected Utility Theory

Violations of expected utility theory observed in our data set can be organized according to five classifications: minor oscillation, oscillation, major oscillation, reversal or irrational risk aversion. Below, we define each of these classifications. Table 10 contains examples of each classification.

**Minor Oscillation** – The subject selected the low variance lottery in the first several pairs. He then selected the high variance lottery in the next pair, the low variance lottery in the following pair, and the high variance lottery in the remaining pairs. The brief switch back to the low variance lottery does not conform to expected utility theory, but could be interpreted as an oversight. Two subjects' decisions were consistent with minor oscillation.

**Oscillation** – As with minor oscillation, the subject switched back to the low variance lottery after crossing over to the high variance lottery, then later returned to preferring the high variance lottery. However, unlike minor oscillation, oscillation implies that the switch back to the low variance lottery was extended in duration. Two subjects' choices displayed oscillation.

**Major Oscillation** – The subject consistently switched back and forth between the low and high variance lotteries. Four subjects' choices showed major oscillation.

**Reversal** – The subject had a single crossover point, but crossed over from the high variance lottery to the low variance lottery. One subject displayed a reversal.

**Irrational Risk Aversion** – The subject preferred the low variance lottery in all cases. Since in the 10<sup>th</sup> lottery pair the high payoff is a certainty, no subject should ever prefer the low variance lottery, as they are guaranteed a higher payoff by selecting the high variance lottery. Four subjects' behavior could be categorized as irrational risk aversion.

With the exception of minor oscillation, we took violations of expected utility theory as evidence that the subject did not understand his incentives in the lottery section. In the 11 instances of subjects displaying oscillation, major oscillation, reversal or irrational risk aversion, we set *Comprehend* = 0. In all other cases, *Comprehend* = 1. In the two cases of minor oscillation, we follow Holt & Laury's (2002) lead and count the first switch from the low variance lottery to the high variance lottery as the crossover point.

## References

- Baiman, Stanley, and Barry L. Lewis. "An Experiment Testing the Behavioral Equivalence of Strategically Equivalent Employment Contracts." *Journal of Accounting Research* (Spring, 1989): 1 – 20.
- Berg, Joyce E., Lane Daley, John Dickhaut, John O'Brien. "Controlling Preferences for Lotteries on Units of Experimental Exchange." *Quarterly Journal of Economics*, (May, 1986): 281 – 306.
- \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_, \_\_\_\_\_. "Moral Hazard and Risk Sharing: Experimental Evidence." R. M. Isaac, ed. *Research in Experimental Economics*, vol. 5 JAI Press, Greenwich, CT, 1992: 1 – 34.
- Cadsby, Bram C., Fei Song, and Francis Tapon. "Sorting and Incentive Effects of Pay for Performance: An Experimental Investigation." *Academy of Management Journal* (April 2007): 387 – 405.
- Chow, Chee W. "The Effects of Job Standard Tightness and Compensation Scheme on Performance: An Exploration of Linkages." *The Accounting Review* (Oct., 1983): 667 – 685.
- Dillard, Jesse F., and Joseph G. Fisher. "Compensation Schemes, Skill Level, and Task Performance: An Experimental Examination." *Decision Sciences* (Winter, 1990): 121 – 137.
- Epstein, Seth. "Testing Principal-Agent Theory." R. M. Isaac, ed. *Research in Experimental Economics*, vol. 5 JAI Press, Greenwich, CT, 1992: 35 – 60.
- Eriksson, Tor, and Villeval, Marie-Claire, "Other-Regarding Preferences and Performance Pay – An Experiment on Incentives and Sorting." IZA Discussion Paper No. 1191 (Jun., 2004).
- Ghosh, Mrinal, and George John. "Experimental Evidence for Agency Models of Salesforce Compensation." *Marketing Science* (Autumn, 2000): 348 – 365.
- Guth, Werner, Wolfgang Klose, Manfred Konigstein, Joachim Schwallbach. "An Experimental Study of a Dynamic Principal-Agent Relationship." *Managerial and Decision Economics* (Jun. – Aug, 1998): 327 – 341.
- Holt, Charles A., and Susan K. Laury. "Risk Aversion and Incentive Effects." *American Economic Review* (Dec. 2002): 1644 – 1655.
- Houser, Daniel, Kevin McCabe, Steve Saletta, Erte Xiao, and Vernon Smith. "Working for Self vs. Working for Other." Unpublished working paper, 2006.
- Waller, William S., and Chee W. Chow. "The Self-Selection and Effort Effects of Standard-Based Employment Contracts: A Framework and Some Empirical Evidence." *The Accounting Review* (Jul., 1985): 458 – 476.
- Young, Mark, and Barry Lewis. "Experimental Incentive-Contracting Research in Management Accounting." Robert H. Ashton and Alison Hubbard Ashton ed. *Judgment and Decision-Making Research in Accounting and Auditing*, Cambridge University Press, New York, NY, 1995: 55 – 75.

## Tables and Figures

**Table 1. Parameter Values by Treatment**

	Parameter	$S$	$\mu$	$\tau$	$c$	Min $\pi_H$	Max $\pi_H$
Treatment							
Threshold		\$1.15	\$0.00	3	\$0.05	-\$0.50	\$1.00
Multiple		\$0.00	\$0.20	0	\$0.05	-\$0.50	\$1.24
Combo		\$0.90	\$0.12	3	\$0.05	-\$0.50	\$1.50

**Table 2. Ten Lottery Pairs**

		Lottery 1		Lottery 2		Expected Value	
$p_H$	$p_L$	$\pi_H$	$\pi_L$	$\pi_H$	$\pi_L$	Lottery 1	Lottery 2
10%	90%	\$2.00	\$1.60	\$3.85	\$0.10	<b>\$1.64</b>	\$0.48
20%	80%	\$2.00	\$1.60	\$3.85	\$0.10	<b>\$1.68</b>	\$0.85
30%	70%	\$2.00	\$1.60	\$3.85	\$0.10	<b>\$1.72</b>	\$1.22
40%	60%	\$2.00	\$1.60	\$3.85	\$0.10	<b>\$1.76</b>	\$1.60
50%	50%	\$2.00	\$1.60	\$3.85	\$0.10	\$1.80	<b>\$1.98</b>
60%	40%	\$2.00	\$1.60	\$3.85	\$0.10	\$1.84	<b>\$2.35</b>
70%	30%	\$2.00	\$1.60	\$3.85	\$0.10	\$1.88	<b>\$2.73</b>
80%	20%	\$2.00	\$1.60	\$3.85	\$0.10	\$1.92	<b>\$3.10</b>
90%	10%	\$2.00	\$1.60	\$3.85	\$0.10	\$1.96	<b>\$3.48</b>
100%	0%	\$2.00	\$1.60	\$3.85	\$0.10	\$2.00	<b>\$3.85</b>

**Table 3. Crossover Points and Implied Utility Functions**

Crossover Point	Implied Range of $r$	Risk Preference Classification	Percentage of Choices <sup>20</sup>
1	$r < -1.71$	God Complex	2.73%
2	$-1.71 < r < -0.95$	Highly Risk Loving	0.00%
3	$-0.95 < r < -0.49$	Very Risk Loving	4.11%
4	$-0.49 < r < -0.15$	Risk Loving	5.48%
5	$-0.15 < r < 0.15$	Risk Neutral	19.18%
6	$0.15 < r < 0.41$	Slightly Risk Averse	24.66%
7	$0.41 < r < 0.68$	Risk Averse	17.81%
8	$0.68 < r < 0.97$	Very Risk Averse	19.18%
9	$0.97 < r < 1.37$	Highly Risk Averse	0.00%
10	$1.37 < r$	Stay In Bed	1.37%
No Crossover	--	Irrational Risk Aversion	5.48%

<sup>20</sup> Seven subjects' choices are not accounted for due to multiple crossover points (see Appendix).

**Table 4. Summary of the Five Global Factors**

Global Factor	Analogous FFM Dimension	Characteristics of Individual With	
		Low Factor Score	High Factor Score
Extraversion	Extraversion	Reserved Quiet Independent	Assertive Energetic Cheerful Enthusiastic
Emotional Resilience	Neuroticism <sup>‡</sup>	Self-confident Relaxed Even Tempered	Anxious Hostile Self-conscious
Self Control	Conscientiousness	Spontaneous Impulsive Irresolute Fickle	Organized Persistent Motivated
Independence	Agreeableness <sup>‡</sup>	Trusting Forgiving Altruistic	Self-centered Suspicious Ruthless
Practicality	Openness to Experience <sup>‡</sup>	Creative Curious Imaginative Untraditional	Conventional Narrow in interests Utilitarian Straightforward
<sup>‡</sup> Indicates that high scores of the Global Factor correspond with low scores of the FFM Dimension			

Figure 1. Distribution of Attempts by Treatment in Compulsory Sections

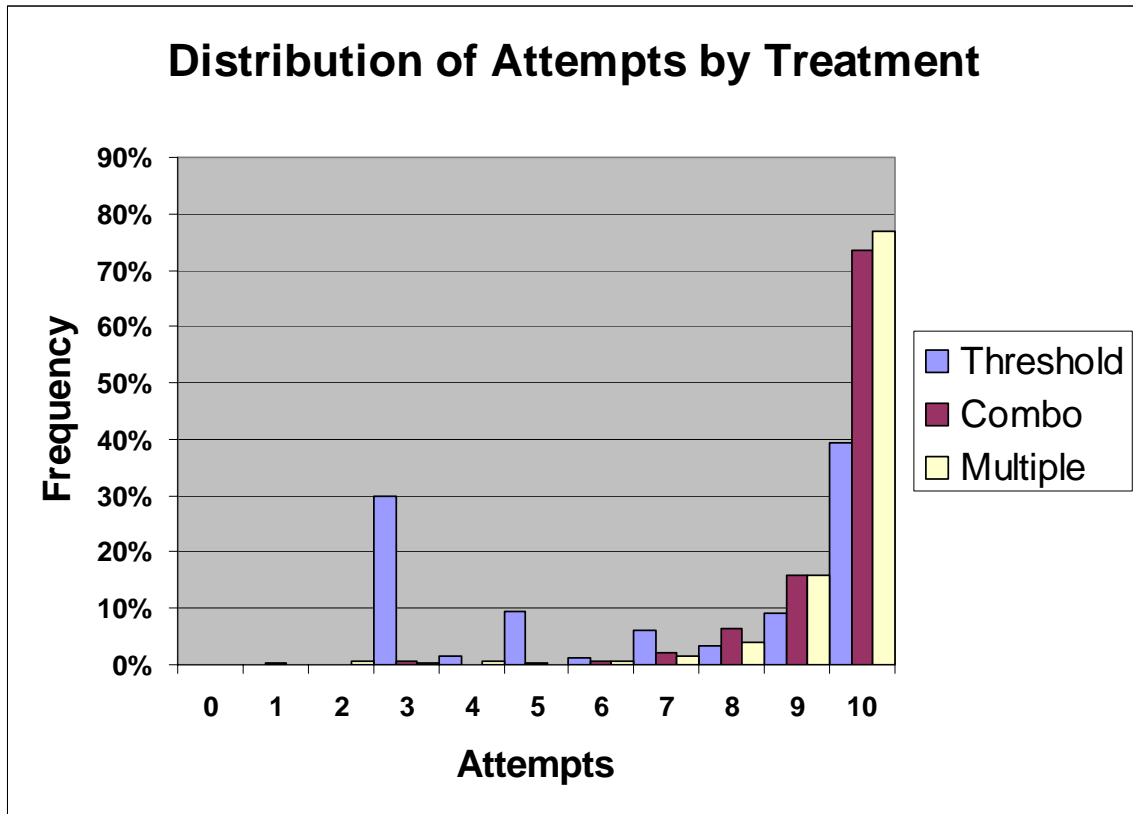


Figure 2. Distribution of Attempts by Treatment in Compulsory Sections, Round 5 Only

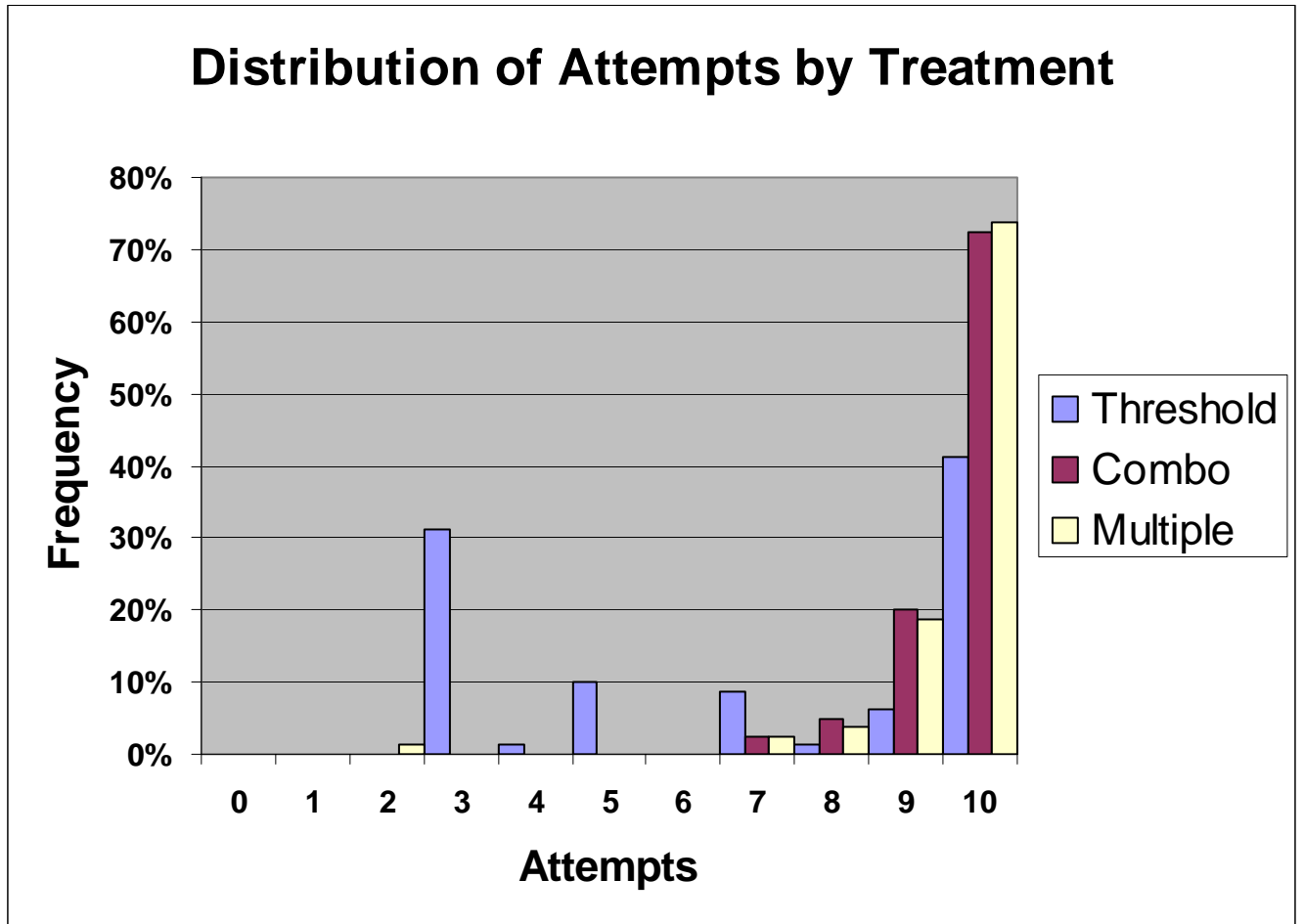
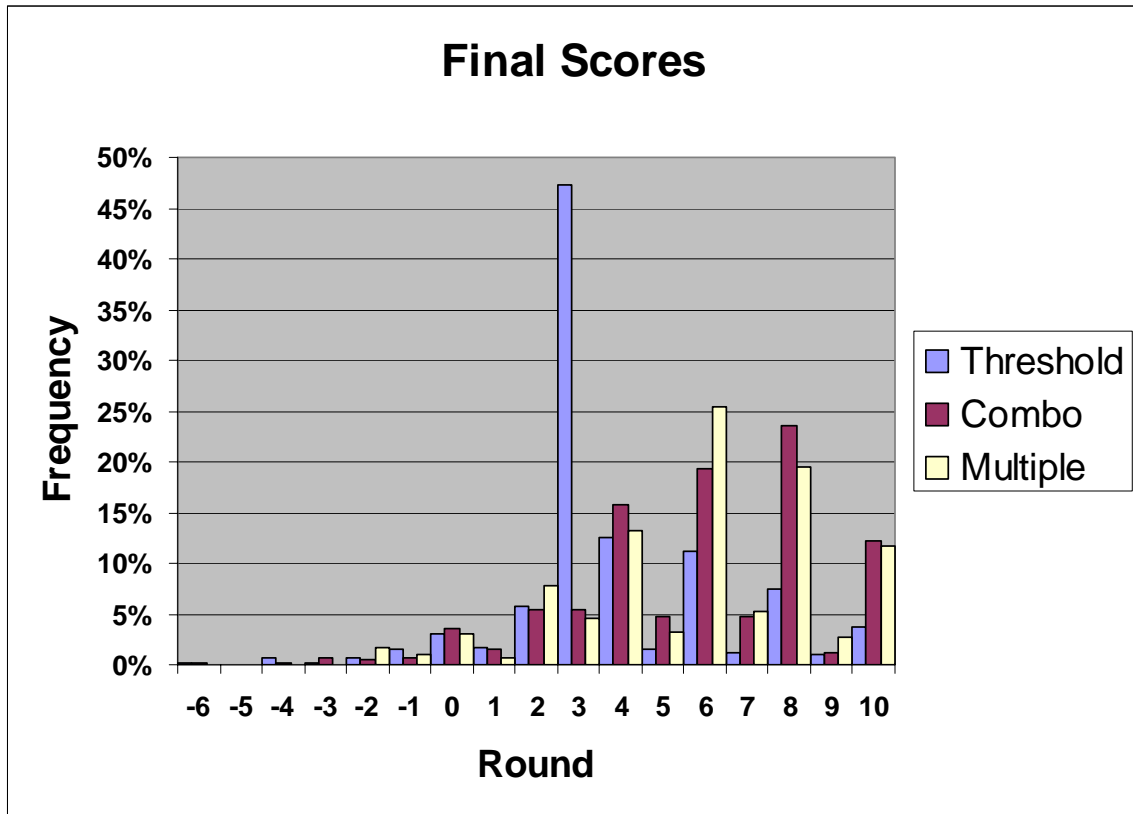


Figure 3. Distribution of Final Scores by Treatment in Compulsory Sections



**Table 5. Results of Regression on Effort in Compulsory Sections**

Variable	Coefficient (Std. Error)		P-Statistic		
$\alpha$	9.5 <sup>***</sup> (0.2)		0.000		
<i>Threshold</i>	0.33 (0.37)		0.361		
<i>IR * Threshold</i>	4.88 <sup>***</sup> (0.5)		0.000		
<i>Combo</i>	-0.05 (0.07)		0.502		
<i>Accuracy<sub>it</sub></i>	-0.81 (0.71)		0.254		
<i>Threshold * Accuracy<sub>it</sub></i>	-0.6 (0.46)		0.192		
<i>IR * Threshold * Accuracy<sub>it</sub></i>	-11.03 <sup>***</sup> (0.59)		0.000		
<i>Combo * Accuracy<sub>it</sub></i>	1.16 (0.82)		0.161		
Observations:	1200	R <sup>2</sup> :	0.8107	Wald $\chi^2$	4134.84
* Indicates Significance at 10% confidence level ** Indicates Significance at 5% confidence level *** Indicates Significance at 1% confidence level					

**Table 6. Results of Logistic Regression on Incentive Responsiveness**

Variable	Coefficient (Std. Error)	P-Statistic
$\alpha$	2.986** (1.139)	0.009
<i>Comprehend</i>	3.162*** (1.117)	0.005
<i>Male</i>	0.117 (0.623)	0.851
<i>Extraversion</i>	-0.042 (0.336)	0.900
<i>Emotional Resilience</i>	-0.842** (0.375)	0.025
<i>Self-Control</i>	-0.699 (0.472)	0.138
<i>Independence</i>	-0.094 (0.373)	0.801
<i>Practicality</i>	0.658*	0.063
Observations:	77	
LR $\chi^2(7)$ :	26.79	
Pseudo-R <sup>2</sup> :	0.2513	
* Indicates Significance at 10% confidence level ** Indicates Significance at 5% confidence level *** Indicates Significance at 1% confidence level		

Figure 4. Distribution of Attempts by Treatment in Elective Section

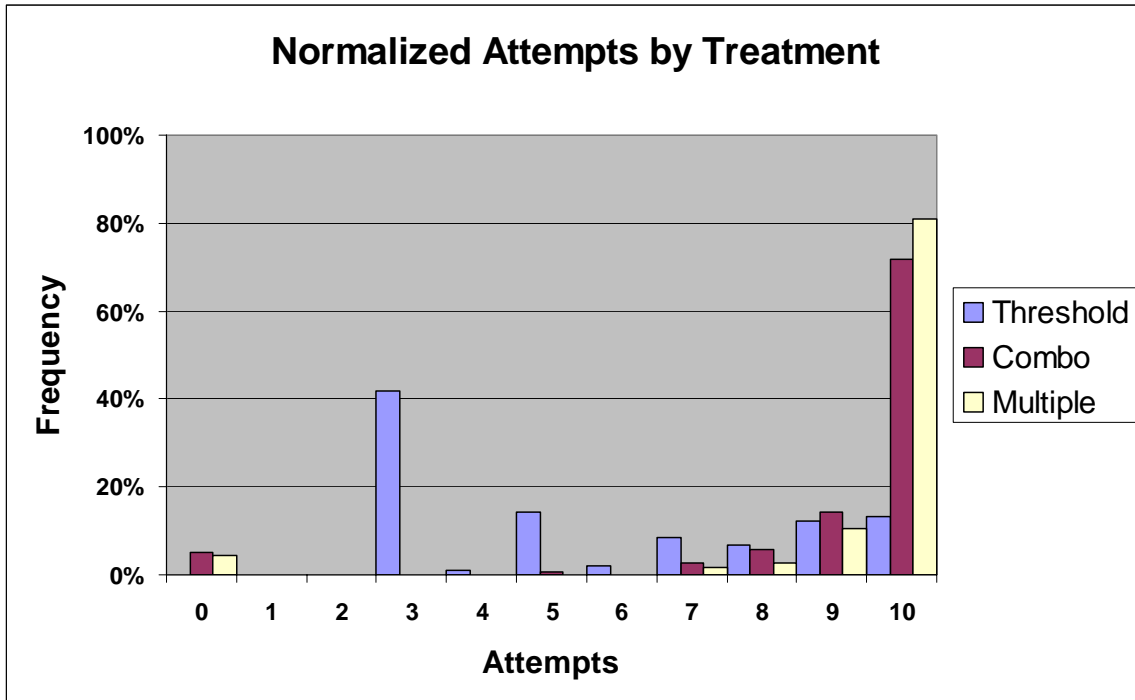
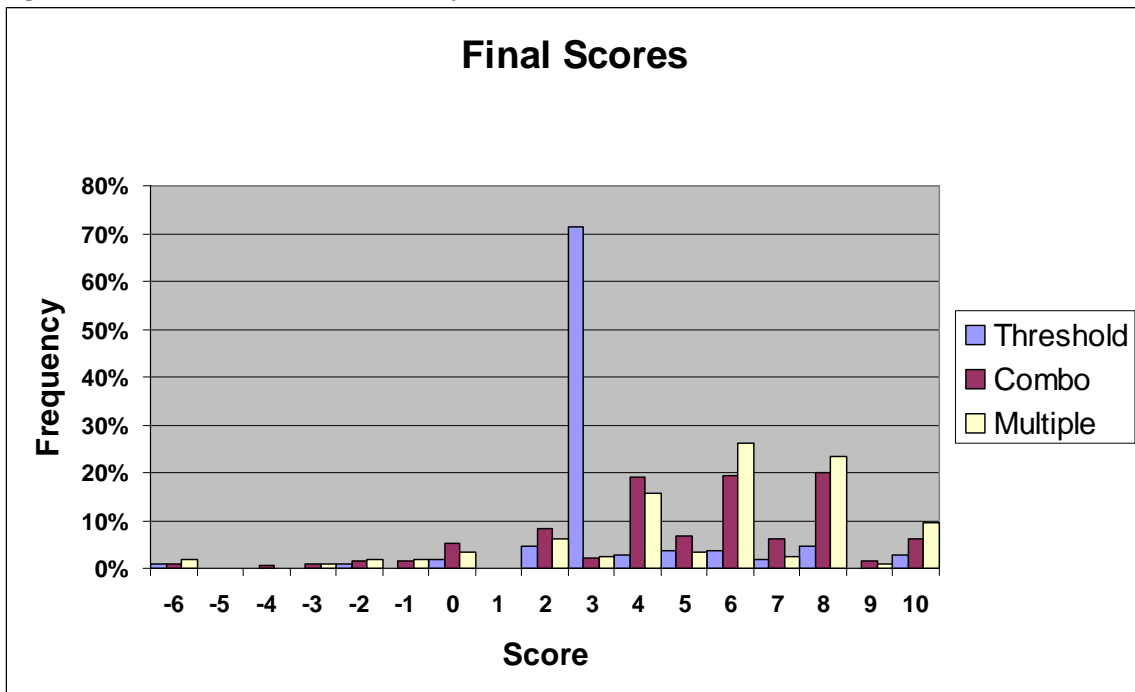


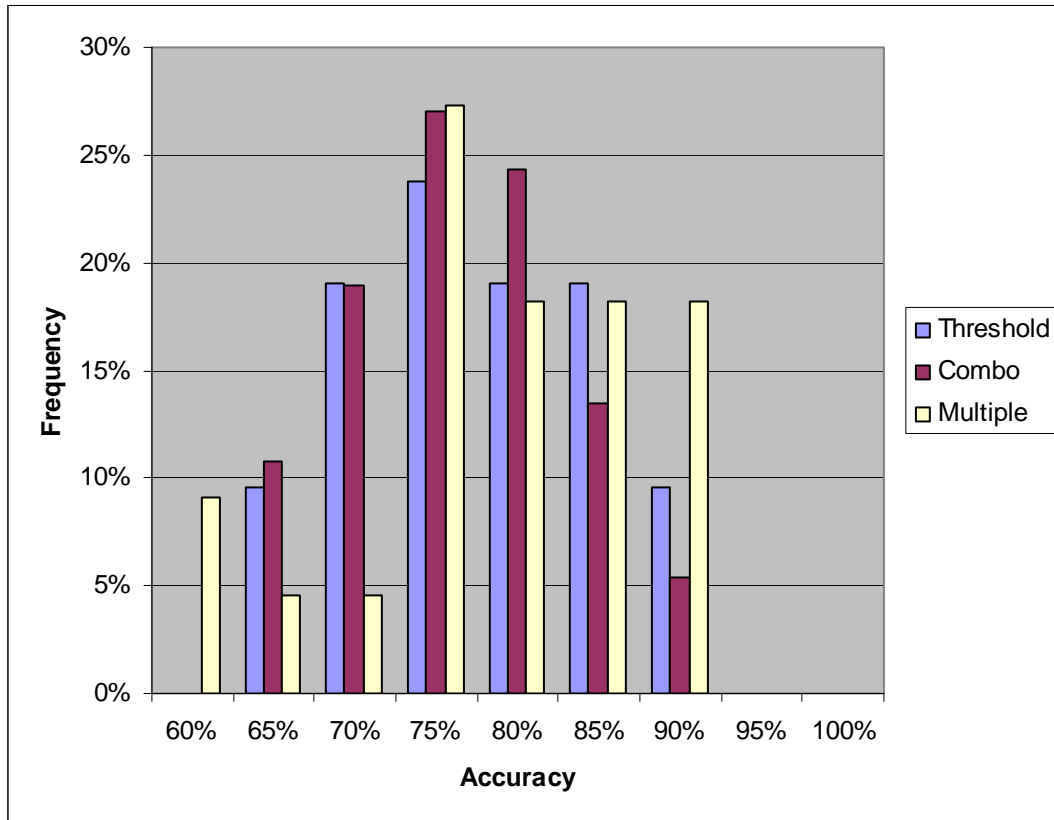
Figure 5. Distribution of Final Scores by Treatment in Elective Section



**Table 7. Results of Regression on Effort in Elective Section**

Variable	Coefficient (Std. Error)		P-Statistic		
$\alpha$	9.59 <sup>***</sup> (0.39)		0.000		
<i>Threshold</i>	-1.07 (0.8)		0.182		
<i>IR * Threshold</i>	4.93 <sup>***</sup> (0.84)		0.000		
<i>Combo</i>	-0.17 (0.47)		0.712		
<i>Accuracy<sub>it</sub></i>	0.24 (0.47)		0.608		
<i>Threshold * Accuracy<sub>it</sub></i>	0.5 (0.98)		0.612		
<i>IR * Threshold * Accuracy<sub>it</sub></i>	-11.04 <sup>***</sup> (0.99)		0.000		
<i>Combo * Accuracy<sub>it</sub></i>	0.02 (0.57)		0.966		
Observations:	400	R <sup>2</sup> :	0.8940	Wald $\chi^2$	1840.52
* Indicates Significance at 10% confidence level ** Indicates Significance at 5% confidence level *** Indicates Significance at 1% confidence level					

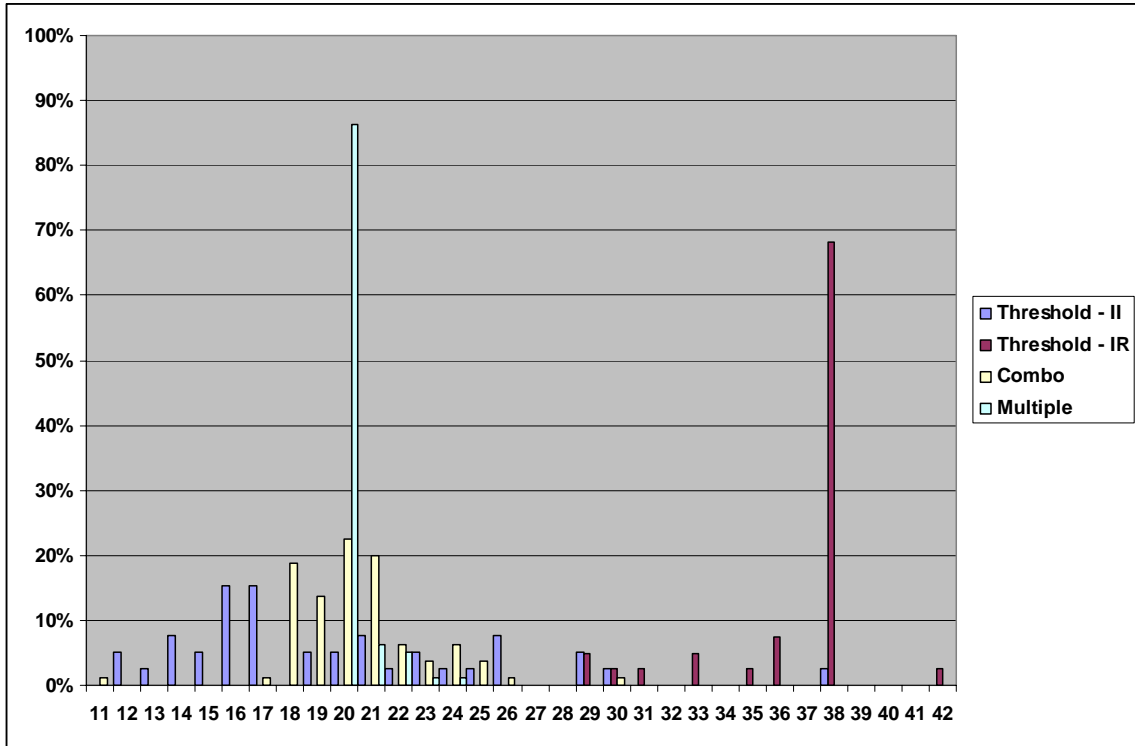
**Figure 6. Distributions of Skill (Measured by Accuracy) by Contract Selection**



**Table 8. Results from Logistic Regression Models of Contract Selection**

Independent Variable	Threshold Model	Combo Model	Multiple Model
	Coefficient (Std. Error)	Coefficient (Std. Error)	Coefficient (Std. Error)
$\alpha$	-6.44 (4.3)	9.92** (4.4)	-1.36 (9.7)
<i>Crossover</i>	0.36* (0.2)	-0.43** (0.2)	0.09 (0.48)
<i>Accuracy</i>	4.86 (5.01)	-8.37* (4.9)	-3.94 (9.41)
<i>Norm Earnings</i>	0.01** (0.004)	0.01*** (0.004)	0.05** (0.02)
<i>Male</i>	-0.15 (0.76)	-0.42 (0.85)	2.36 (1.98)
<i>Extraversion</i>	0.23 (0.42)	0.83 (0.51)	-0.56 (1.04)
<i>Emotional Resilience</i>	-0.44 (0.45)	0.95* (0.52)	0.11 (1.32)
<i>Self Control</i>	0.92 (0.58)	-1.31* (0.69)	1.66 (1.69)
<i>Independence</i>	0.09 (0.46)	-0.36 (0.5)	0.41 (1.37)
<i>Practicality</i>	-0.9** (0.43)	1.27** (0.53)	-0.16 (0.8)
Observations	65	65	65
Pseudo R2	0.2474	0.4385	0.7384
LR $\chi^2(9)$	18.97	39.51	50.01
* Indicates Significance at 10% confidence level ** Indicates Significance at 5% confidence level *** Indicates Significance at 1% confidence level			

**Figure 7. Distributions of Average Cost Per Unit – Compulsory Sections**



**Table 9. Mean ACPU and ACPU Differences across Treatments and Assignment Methods**

Assignment Method	Treatment	Mean ACPU	Difference In Means (t-statistic)								
			Assignment Method								
			Compulsory				Elective				
			Threshold (II)	Threshold (IR)	Combo	Multiple	Threshold (II)	Threshold (IR)	Combo	Multiple	
Compulsory	Threshold (II)	19.71	-								
	Threshold (IR)	37.72	18.01 <sup>§</sup> (15.46)	-							
	Combo	20.39	0.67 (0.7)	-17.34 <sup>§</sup> (-22.7)	-						
	Multiple	20.25	0.53 (0.58)	-17.48 <sup>§</sup> (-24.47)	-0.14 (-0.48)	-					
Elective	Threshold (II)	19.05	-0.66 (-0.35)	-18.68 <sup>§</sup> (-10.28)	-1.34 (-0.79)	-1.2 (-0.72)	-				
	Threshold (IR)	40.3	20.59 <sup>§</sup> (6.94)	2.58 (0.89)	19.92 <sup>§</sup> (7.03)	20.06 <sup>§</sup> (7.11)	21.26 <sup>§</sup> (6.48)	-			
	Combo	24.65	4.94 (1.34)	-13.07 <sup>‡</sup> (-3.58)	4.26 (1.19)	4.40 (1.23)	5.60 (1.42)	-15.65 <sup>‡</sup> (-3.43)	-		
	Multiple	21.51	1.8 (1.16)	-16.22 <sup>§</sup> (-11.32)	1.12 (0.88)	1.26 (1.01)	2.46 (1.18)	-18.8 <sup>§</sup> (-6.1)	-3.14 (-0.83)	-	
<sup>‡</sup> Indicates significance at 1% level <sup>§</sup> Indicates significance at 0.1% level											

**Table 10. Sample Violations of Expected Utility Theory.**

“L” and “H” correspond to choosing the low and high variance lottery from the pair respectively.

Lottery Pair	Minor Oscillation	Oscillation	Major Oscillation	Reversal	Irrational Risk Aversion
1	L	L	L	H	L
2	L	L	H	H	L
3	L	L	L	H	L
4	L	L	H	H	L
5	H	H	L	H	L
6	L	H	H	L	L
7	H	L	L	L	L
8	H	L	H	L	L
9	H	H	L	L	L
10	H	H	H	L	L