Motivation, Test Scores, and Economic Success*

Carmit Segal†
Harvard Business School

November 22, 2006

JOB MARKET PAPER

Abstract

In this paper I investigate through which channels low-stakes test scores relate to economic success. The inferences in the economic literature regarding test scores and their association with economic outcomes are mostly based on tests without performance-based incentives, administered to survey participants. I argue that the lack of performance-based incentives allows for the possibility that higher test scores are caused by non-cognitive skills associated with test-taking motivation, and not necessarily by cognitive skills alone. I suggest that the coding speed test, which is a short and very simple test available for participants in the National Longitudinal Survey of Youth 1979 (NLSY), may serve as a proxy for test-taking motivation. To gather more definite evidence on the motivational component in the coding speed test I conduct a controlled experiment, in which I induce motivation via the provision of incentives. In the experiment, the average performance improved substantially and significantly once incentives were provided. More importantly, I find heterogeneous responses to incentives. Roughly a third of the participants improved their performance significantly in response to performance-based incentives, while the others did not. These two groups have the same test score distributions when incentives were provided, suggesting that some participants are less motivated and invest less effort when no performance-based incentives are provided. These participants however are not less able. I then explore to what extent coding speed test scores relate to economic success. Focusing on male NLSY participants, I show that the coding speed scores are highly correlated with earnings 23 years after NLSY participants took the test even after controlling for usual measures of cognitive skills like the Armed Forces Qualification Test (AFQT) scores. Moreover, I find that while for highly educated workers the association between AFQT scores and earnings is significantly larger that the one between coding speed scores and earnings, for less educated workers these associations are of similar size.

*I would like to thank Ed Lazear, Muriel Niederle, and Al Roth for their encouragement, useful suggestions, and numerous conversations; seminar participants at Harvard University, Stanford University, Wesleyan University, ESA meetings, and NBER Labor Studies Fall Meeting ; and George Baker, Greg Barron, Vinicius Carrasco, Pedro Dal Bo, Liran Einav, Florian Englmaier, Itay Fainmesser, Richard Freeman, Ed Glaeser, Avner Greif, Ben Greiner, Felix Kubler, Steve Leider, Aprajit Mahajan, Tatiana Melguizo, Guy Michaels, Amalia Miller, Joao de Mello, Andreas Ortmann, Luigi Pistaferri, Daniel Tsiddon, Ed Vytlacil, Pierre-Olivier Weill, Toni Wegner, and Nese Yildiz for helpful comments.

†Harvard Business School, Baker Library room 437, Boston MA, 02163. Email: csegal@hbs.edu. Phone: (617)-495-6652. Homepage: http://www.people.hbs.edu/csegal/

1 Introduction

The inferences regarding standardized test scores and their association with economic outcomes, race, and gender, are mostly based on tests administered to survey participants. Usually, no performance-based incentives are supplied in these surveys, and thus there is no a priori reason to believe that survey participants try their best to solve the test. As a result, the issue of effort, or motivation, might be crucial to the interpretation of the empirical findings. In particular, if individuals differ not only in their cognitive ability but also in their test-taking motivation, then when no performance-based incentives are supplied, higher test scores do not necessarily imply higher cognitive ability. Instead, higher test scores may be caused by higher test-taking motivation, or by differences in personality traits associated with it, i.e., by non-cognitive skills. Therefore, it is possible that the associations between higher test scores and future economic success should be attributed, at least in part, to differences in non-cognitive skills, and not solely to differences in cognitive skills. In this paper I investigate the relations between test-taking behavior, test scores, and economic success. To do so, I explore the low-stakes test scores available for the participants in the National Longitudinal Survey of Youth 1979 (NLSY). I identify a test that may serve as a proxy for individuals' test-taking motivation. I use controlled experimental results to show that indeed this test does not measure pure ability, but also motivational components. I then explore its relations with other low-stakes test scores, which have been repeatedly used as a measure of cognitive skills across race/ethnicity, and with economic success.

Although the role of incentives has been emphasized in economic theory, there are surprisingly few empirical papers in economics dealing directly with the provision of incentives and their effect on performance (see for example Lazear 2000, Bandiera, Barankay, and Rasul 2005, 2006a and 2006b). In particular, there are few empirical studies investigating the relationship between motivation and test scores.[1] In experimental settings, however, there is clear evidence that performance on tests is related to incentives. Gneezy and Rustichini (2000) show that while the effect of monetary incentives can be non-monotonic, sufficiently high incentives lead to a significant increase in average test scores. Borghans, Meijers and ter Weel (2006) show that test scores on IQ tests respond to incentives when the time allotted to the test is sufficiently long. Kremer, Miguel, and Thornton (2005) show in a randomized experiment in Kenya that monetary incentives raise test scores on low-stakes tests in primary schools. Angrist and Lavy (2004) find in a randomized experiment that monetary incentives raise performance on high-stakes tests in Israel. Although the psychometric literature seems to work under the assumption that all test takers are well motivated, there is substantial evidence, dating back to the beginning of the last century, that motivation affects performance and that motivation may be related to personality traits (for an excellent summary see Revelle 1993). [2]

[1] Investigating Danish test scores results, McIntosh, Munk, and Chen, (2006) explicitly suggest that the lack of incentives may create problems in interpreting test scores.

[2] Several fields in psychology have investigated the effects on intrinsic and extrinsic motivation on performance. Here the results are less consistent, in part probably due to different selection rules. For a summary on the effects of rewards on

While test scores may indeed measure cognitive skills when test takers are highly motivated to take the test (as would be the case when individuals take pre-employment tests that affect their chances to get a job or the SAT or ACT exams that affect their chances to be accepted to the university of their choice), it is not clear what the same test scores measure when test takers are not highly motivated. Economic theory implies that if costly effort is needed to solve a test, then when no performance-based incentives are supplied test takers invest the lowest level of effort possible to do it. However, survey participants' test scores are rarely equal to zero. A possible explanation is that individuals gain psychic benefits from higher test scores. If individuals with higher cognitive ability have lower costs of effort (or may even find higher test scores more rewarding), higher ability individuals will have higher test scores than lower ability individuals. As a result, even when no performance-based incentives are supplied survey participants' test scores would provide a correct ranking according to their ability. However, if the most able test-takers are not the ones that gain the highest psychic benefits form having higher tests scores, then test scores will not provide correct ranking according to the test takers' ability. In this case, low test scores do not necessarily imply that individuals or groups have low cognitive ability. Moreover, if test-taking motivation relates to personality traits, which are unrelated to cognitive ability, then cognitive ability may not necessarily be the sole cause of the associations between test scores and economic outcomes. One personality trait that seems to be a likely candidate to have an effect on both test-taking motivation and later labor market outcomes is conscientiousness.[3] Understanding what is being measured by low-stakes test scores may increase our knowledge of the mechanisms underlying economic success, and may even allow for better design of policies to help individuals and groups.

To investigate the relationships of motivation and low- and high-stakes test scores to outcomes, ideally one would like to have both low-stakes and high-stakes test scores for the same individuals and the same test. However, to the best of my knowledge, there exists no such data set.[4] Instead, I turn to the NLSY data set to search for a proxy for test-taking motivation and to investigate the relationship between low-stake tests and economic success. The NLSY contains information about a battery of ten tests, namely the Armed Service Vocational Aptitude Battery (ASVAB).[5] Participants in the NLSY were paid $50 honorarium for completing the ASVAB test, but no direct performance-based incentives were provided. Thus, for these participants the ASVAB is a low-stakes test. As such, it is possible that the relationships found between scores on different ASVAB subtests and economic success could be, at least partially, attributed to test-taking motivation and thus to personality traits associated with it, and not to cognitive skills alone.

intrinsic motivation and performance see Eisenberger and Cameron (1996) and Cameron, Banko and Pierce (2001). See Arvey, Strickland, Drauden and Martin (1990), for a short summary of the literature on motivation and pre-employment testing.

[3]While conscientiousness has been repeatedly found to be positively correlated educational attainments (see for example Duckworth, Peterson, Matthews and Kelly 2006) and with labor market outcomes (see for example Judge, Higgins, Thoresen, and Barrick, 1999), it is found to be either uncorrelated to IQ (see for example Ackerman and Heggestad, 1997) or negatively correlated with it (see for Moutafi, Furnham, and Paltiel, 2005).

[4]Even if data with SAT/ACT test scores were available one would face serious selection problems.

[5]The ASVAB test consists of 10 subtests that are described in Table 1.

While the scores on all ASVAB subtests should be affected by test-taking motivation, the effects may be more pronounced and therefore easier to detect in tests which seem to be very simple, and do not require specialized knowledge. A possible candidate is the coding speed test. The task in the coding speed test is to match words with four digit numbers (an example of the test is given in Table 2). To figure out which word matches to which number, test takers need to look at the key, in which the four digit number associated with each word is given. It seems likely that as long as one knows how to read, one possess the knowledge necessary to answer questions correctly on the coding speed test and that concentration and effort would be the main skills contributing to high scores. If this were the case, then when no performance-based incentives are provided the coding speed test scores would provide a perfect measure of individuals' test-taking motivation. Still, the time allotted to the coding speed test is very short (84 questions in 7 minutes), and thus it is possible that the coding speed test scores measure ability, maybe one which relates to speed.

The coding speed test scores have been utilized in the past. Using factor analysis Heckman (1996) and Cawley, Conneely, Heckman, and Vytlacil (1997) have shown that the two speeded tests in the ASVAB (i.e., the coding speed test and the numerical operation test, which includes very simple arithmetic computations) correspond to a different factor than the other subtests in the ASVAB. The authors suggest that the coding speed test scores (as well as the numerical operation test scores) measure "fluid intelligence or problem solving ability" (Heckman 1996, p. 1105). As was argued above, while it is possible that the coding speed test scores measure fluid intelligence when test takers are highly motivated, this is not necessarily the case when no performance-based incentives are supplied.

To gather more conclusive evidence on the nature of the coding speed test, I conducted a controlled experiment. The lab, where it is possible to control and induce motivation via the provision of incentives, is the natural place to investigate what the coding speed test scores measure. The experiment consisted of two parts. In the first part participants were asked to solve the coding speed test twice for a fixed payment. The first version was called "practice test" and the second was called "the test". If test scores differ between these two tests, the difference may be attributed to a change in intrinsic motivation. In the second part they solved a third version of the test for which they were paid according to their performance. As expected, participants' performance improved substantially and significantly between the three tests.[6] Thus, if we were interested in an absolute measure of participants' ability, the experimental results suggest that we are unable to estimate it correctly from tests in which no performance-based incentives were supplied.

If the coding speed test scores measure ability alone, then the provision of incentives may result in improvement in individuals' test scores, nevertheless their relative ranking should not be changing. However, if test taking motivation differs across individuals, then the relative ranking of individuals may change in response to the provision of incentives. In the experiment, I find that different participants respond differently to the change in incentives. Specifically, participants can be divided into two types. While the first type consists of individuals whose performance did not improve with monetary

[6]There is very little evidence of learning by doing in the experiment, if at all, learning is restricted to practice test.

incentives, the performance of all individuals of the second type improved significantly. It turns out that when no monetary incentives are provided the test score distributions of the two types differ, and the test scores of type one first order stochastically dominates the test scores distribution of type two. Thus, type two participants look less able. However, individuals of different types have the same test score distributions when performance-based monetary incentives are provided. Taken together these results suggest that those who perform worse when no monetary performance-based incentives were provided invest less effort and were less motivated.

Thus, the experiment serves as direct evidence that the coding speed test scores do not measure ability alone. Moreover, it shows that some individuals do not try their best when no performance based incentives are provided, though these individuals are not the less able ones. Through a psychological survey, administered in the end of the experiment, I find that male participants who only invest high levels of effort when monetary incentives are provided are less likely to be conscientious. Therefore, if this phenomenon does not depend on the particular test, then all low-takes test scores will measure motivation too. Confirming evidence can be found in an independent experimental study by Borghans et al. (2006), which corroborates that indeed these types exist, and correlates them to additional measures of personality traits and individuals' preferences.[7]

The ASVAB is administered to potential recruits to the armed forces, and serves as the entrance exam to the military as well as a sorting device to different military professions.[8] Therefore, it seems likely that potential recruits want to enlist and thus would be highly motivated when taking the ASVAB. Indeed, potential recruits to the armed forces have higher coding speed test scores than NLSY participants, in particular in the lower part of the test score distribution (see Maier and Sims, 1983, and Maier and Hiatt, 1986).[9] In contrast, potential recruits scored worse than the NLSY participants on all other ASVAB subtests (besides numerical operations). Maier and Hiatt (1986, p.2) suggest that the reason is "because the educational level of the 1980 Youth Population [i.e., participants in the NLSY] is higher than that of military applicants, the group is expected to have higher ASVAB scores". This finding certainly allows for the possibility that the coding speed test scores measure test taking motivation.

I use the NLSY data set to investigate the relationship between coding speed scores and earnings. To control for cognitive ability I use the Armed Forces Qualification Test (AFQT). The AFQT has been extensively used in the past as a measure of cognitive skills and it has been repeatedly found to be highly correlated with future income of NLSY participants (see for example Herrnstein and Murray 1994, Heckman 1996, Neal and Johnson 1996). My main empirical findings are as follows.[10]

[7] A closer look at appendix 2 in Gneezy and Rustichini (2000) seems to suggest that the effect of incentives in the standardized test environment was to move people away (when the incentives were high enough) and toward (when the incentives were low) zero scores on the test, where zero score could be interpreted as zero effort. This effect of incentives on the distributions was noted also by Rydval and Ortmann (2004).

[8] The coding speed test was originally made part of the ASVAB to help sort recruits to clerical positions and to help detect cheating on the AFQT (see Maier and Sims, 1983, p. A10, and Maier and Hiatt, 1986, p. A3).

[9] This is also the case for the numerical operation test, which is the other speeded test on the ASVAB (Maier and Sims, 1983, and Maier and Hiatt, 1986).

[10] The results reported below are for male NLSY participants of the three youngest school-cohorts (i.e., born between

The coding speed test scores are economically and statistically significantly associated with earnings, 23 years after NLSY participants took the coding speed test. After controlling for the AFQT scores, the coding speed scores are both economically and statistically associated with earnings.[11]

The coding speed test scores are important to earnings for workers of all education levels, while the AFQT scores are significantly more important to earnings of workers who at least graduated from college. However, the relative magnitudes suggest that for highly educated workers the AFQT scores plays a significantly larger role in their earnings.

The first set of findings indicates that the coding speed test scores measure skills positively valued in the market. However, these findings do not tell us whether the coding speed test scores measure test-taking motivation or another innate ability, possibly fluid intelligence. While the experimental results suggest that test-taking motivation is important for the coding speed scores, the variance across individuals is not zero even when monetary performance-based incentives are provided. Thus, there may be differences across individuals that may be attributed to speed or fluid intelligence. The second set of findings indicates that if the coding speed test scores measure fluid intelligence, then problem solving ability seems to be more important for low educated workers.

Is it possible that the motivation to take the ASVAB relates to earnings of NLSY participants 23 years after they took the ASVAB test? If the motivation to take the ASVAB relates to personality traits, and if these traits are valued in the market then we may observe the patterns above. Indeed, in the experiment, male participants who invested high effort in solving the test only when monetary performance-based incentives were supplied, were less likely to be conscientious.

In recent years several papers have demonstrated directly and indirectly the importance of non-cognitive skills to economic success (Bowles, Gintis and Osborne 2001; Cawley, Heckman and Vytlacil 2001, Jacob 2002, Carneiro and Heckman 2003, Heckman and Rubinstein 2001, Persico, Postlewaite and Silverman 2004, Borghans, ter Weel and Weinberg 2005, Segal 2005, and Heckman, Stixrud and Urzua 2006). Using the National Educational Longitudinal Survey (NELS) data set I have shown in Segal (2005) almost identical findings. In particular, I used teacher evaluations of student behavior in the 8^{th} grade on five attributes (absenteeism, disruptiveness, inattentiveness, tardiness, and homework completion) to create a pre-market measure of misbehavior. I found that for young men, after controlling for 8^{th} grade test scores, misbehavior in the 8^{th} grade is associated with earnings. Moreover, misbehavior in the 8^{th} grade was associated with lower earnings for young men of all education levels, while 8^{th} grade test scores were associated with earnings only for young men with postsecondary degrees. While misbehavior in the 8^{th} grade is unlikely to be related to fluid intelligence or to speed of reading, it seems very likely that it is associated with conscientiousness and the tendency to follow the rules. Thus, the results in Segal (2005) can serve as an independent indicator that indeed the coding speed test scores measures personality traits, and conscientiousness in particular.

October 1^{st} 1961 and September 30^{th} 1964). See Appendix B for details.

[11] These results are consistent with the evidence in Heckman (1996) and Cawley et al. (1997) that provide evidence that the two speeded tests in the ASVAB are related to wages of NLSY participants even after controlling the AFQT scores.

Numerous studies have documented that African Americans (and to a lesser degree Hispanics) score lower than whites on standardized tests.[12] There is a substantial black-white AFQT gap.[13] The experimental results already caution us from trying to make inferences from test scores distributions to ability distributions unless all test takers are as motivated to take the test. As it turns out, in the NLSY both blacks and Hispanics have much lower coding speed test scores than whites, suggesting that they may have been less motivated than whites to take the ASVAB. In regressions, I find that by themselves the coding speed test scores can explain about one third of the variation in the AFQT scores. Moreover, I find that they can explain up to 40% of the black-white AFQT gaps. This effect is present even after controlling for family and characteristics.

There are several explanations in the literature that are consistent with minorities being less motivated than whites to take tests. In particular, if minorities believe that due to discrimination doing well on any exam will not affect their future prospects in life, then a rational reaction is to put little value on doing well on tests. In addition, the literature on stereotype threat (see Steele and Aronson, 1998) suggests that minorities, and blacks in particular, may do worse on tests that are supposed to measure their abilities. If indeed minorities are less motivated than whites to take tests, all inferences made from tests in which no performance based incentives are supplied regarding the relative abilities of minorities as opposed to whites are most likely to be wrong.

The paper proceeds as follows. Section 2 contains a model of test-taking motivation. Section 3 contains a description of the tests in the NLSY data set. Section 4 presents the indirect evidence from the armed forces. Section 5 presents the experimental results. Section 6 presents the detail analysis of the relations between coding speed test scores and outcomes. Section 7 concludes.

2 The Model

2.1 The Basic Model

In this section I present a model that describes the problem faced by the agents taking a low-stakes test and agents taking a high-stakes test. In this basic model agents differ from one another only by their endowment of skills the test is supposed to measure, denoted by, x. The random variable x has a density $f(x)$.

Test scores are being produced using two inputs: skill and effort, denoted by e. Hence,

$$TS = TS(x, e)$$

where test scores are increasing in skill, x and effort, i.e. $TS_e > 0$, and $TS_x > 0$. Assume further that $TS_{ex} \geq 0$, i.e. a given increase in effort would results in weakly higher test scores for individuals with higher skills.

[12] For the an excellent summary of the trends, in the black-white test score gap throughout the last four decades, and possible explanations, see Neal (2005).

[13] This gap has been used by Herrnstein and Murray (1994) to argue that the reason blacks are doing worse than whites is that they have lower IQ.

There are costs associated with effort, and therefore the production of test scores is costly. The cost function is given by $C(x,e)$, where $C_e > 0$, and $C_{ee} > 0$. Thus, the costs associated with effort are increasing and convex. It is natural to assume that the costs are decreasing in skill, i.e., $C_x < 0$, and that a given increase in effort is weakly less costly for agents with skill, i.e. $C_{xe} \leq 0$.

When agents take a low-stakes test the resulting test scores do not affect their future. Therefore, the benefits obtained from higher test scores in these conditions could only be psychic.

When agents take a high-stakes test their benefits from higher test scores are no longer psychic alone. Higher test scores have an effect on agents' future. Similarly, when monetary performance based incentives are provided (as is the case in the experiment) higher test scores lead to higher monetary compensation. Therefore, agents' utility function includes another component: their benefits from having higher test scores. In what follows I will concentrate on the provision of monetary incentives. [14] If monetary incentives are provided individuals monetary gains are given by $M = A + \phi TS$, where $A \geq 0$ is a constant, and $\phi > 0$ is the amount paid for each correct question. Individuals' utility function is given by $U(TS, M)$ where utility is increasing in monetary incentives and in test scores, i.e., $U_{TS} > 0$ and $U_M > 0$. Moreover, utility is weakly concave in monetary incentives and test scores, i.e., $U_{TS,TS} \leq 0$ and $U_{M,M} \leq 0$. I assume further that agents' utility is weakly concave in test scores, i.e., $\frac{d^2U}{dTS^2} = \left(U_{TS,TS} + 2\phi U_{TS,M} + \phi^2 U_{M,M}\right) \leq 0$.[15] Agents face the following problem:

$$\underset{e}{Max} \;\; U(e,x) - C(e,x).$$

When no performance-based incentives are supplied to the agents, and if agents do not gain any psychic utility from higher test scores, since effort is costly agents will supply the minimal amount of effort. Given that it is always possible not to solve any question on the test, and that not all survey participants get test scores equal to zero, I will assume that for what follows that agents obtain psychic utility from higher test scores.

Proposition 1 *If agents obtain psychic benefits from higher test scores and/or monetary performance-based incentive are provided to them, then the resulting test scores provide a correct ranking according to agents' skill levels. Moreover, if the marginal utility of money is increasing in ϕ then an increase in the intensity of incentives would result in higher test scores and higher effort.*
The Proof is given in Appendix A.

Proposition 1 indicates that when individuals differ only in their skill endowment test scores would provide correct relative ranking of individuals according to their skills.

One of the common usages of test scores is to compare skill distribution across populations. To allow for this possibility we need to introduce different population (or types) to the model. Assume that there are two types of agents in the model, called type 1 and type 2. The two types possibly differ by their skill levels. Denote the skill of type i by x_i, where x_i is a random variable with a density

[14] The results can be easily extended to high-stakes tests situations in which test scores affect agents' future.
[15] This condition is trivially fulfilled if agents' utility function is separable in money and test scores.

function $f(x_i)$, $\underline{h}_i \leq x_i \leq \overline{h}_i$, and $i = 1, 2$. Denote the test scores of type i by TS_i, where TS_i is a random variable with a density function $\tilde{f}_i(TS_{i2})$, $\underline{TS}_i \leq TS_i \leq \overline{TS}_i$, and $i = 1, 2$.

Proposition 2 *If TS_1 first order stochastically dominates TS_2, then x_1 first order stochastically dominates x_2.*
The Proof is given in Appendix A.

Since test scores provide a correct ranking of individuals according to their skills, then we can make inferences about their skill distribution by looking at their test scores distribution. Thus, proposition 2 implies that if we can find two groups such that the test scores of one group first order stochastically dominates the other we can infer that this is the case for their respective skill distributions. This would be the case regardless of the incentives provided to the agents.

2.2 Types with Different Test-Taking Motivation

So far we have assumed that all agents value low-stakes test scores in the same way, i.e. all gain the same benefits from doing well on a low-stakes test. However, it is possible that not all individuals care about low-stakes test in the same manner. To capture the possibility that some individuals might gain less psychic utility than others from doing well on a test, I introduce types to the setting of the basic model described above. Individuals of different types value the test differently, or equivalently they differ in their test-taking motivation. The new setting is the following.

Agents are of different types, denoted by θ. Agents with lower values of θ gain less psychic benefits from test scores, i.e., $U_{TS,\theta} > 0$. The type θ, though, does not affect agents' utility from money, i.e., $U_{M,\theta} = 0$. Assuming as before that agents' utility function is a function of test scores and money, we can write it $U(TS, M; \theta)$. As before, agents are endowed with skill, denoted by, x. The random variable x has a density which might depend on the type, denoted by $f(x; \theta)$.

The assumptions regarding the test score production function and the cost function associated with effort are the same as before. The production function of test scores is given by $TS = TS(x, e)$, where $TS_e > 0$, $TS_x > 0, TS_{ee} \leq 0, T$ and $TS_{ex} \geq 0$. The cost function, as before, is given by $C(x, e)$, where $C_e > 0$, $C_{ee} > 0$, $C_x < 0$,and $C_{xe} \leq 0$. Agents' problem is given by

$$\underset{e}{Max}\ U(e, x; \theta) - C(e, x).$$

Proposition 3 *Conditional on θ, test scores provide a true ranking according to agents' skill levels. If the marginal utility of money is increasing in ϕ then an increase in the incentives would result in higher effort and higher test score for individuals with the same value of θ. Moreover, Agents with the same skill levels, but with higher values of θ invest more effort, and have higher test scores.*
The Proof is given in Appendix A.

Proposing 3 suggests that If indeed it is the case that there are several types that value the test scores differently (or have different test-taking motivation), then generally test scores would not

provide correct ranking according to individuals' skills.[16] However, they do provide correct ranking according to agents' skill for the agents of the same type. The intuition for the first part is as follows. Types with lower values of θ have lower marginal benefits from higher test scores, though they have the same marginal costs. Thus, for a given skill level, they would choose to invest less effort. Since test scores are produced using both skill and effort, and types with lower values of θ systematically invest less effort, the comparison of test scores across types is uninformative with respect to their relative skills. Within the group of test takers having the same type the comparison of test scores is informative with respect to skill. In order to recover the ranking according to skill in the population as a whole, we would need to induce test takers to invest maximum effort levels. This might be achieved by providing incentives to test takers.

In the previous section we saw that when all individuals value a test in the same manner one could compare test scores across population in order to assess skill differences. As we will see next, this is no longer the case if the populations differ in their valuation of the test.

Proposition 4 *Denote the skill of type θ_i by $x(\theta_i)$. $x(\theta_i)$ is a random variable with a density function $f(x;\theta_i)$, and $\underline{h}_i \leq x(\theta_i) \leq \overline{x}_i$, where $i = 1, 2$ and $\theta_1 > \theta_2$. Assume that $TS(x, \phi, \theta_1)$ first order stochastically dominates $TS(x, \phi, \theta_2)$. This does not imply that $x(\theta_1)$ first order stochastically dominates $x(\theta_2)$.*
The Proof is given in Appendix A.

Proposition 4 implies that even if we find two groups of individuals such that the test scores distributions of one group first order stochastically dominates the test scores distribution of the second group in one incentives scheme, this is not necessarily the case under another incentives scheme if individuals differ in their test-taking motivation. In this case, it is feasible that the group with the low values of θ (i.e., low test-taking motivation), may have higher skill levels than the group with high values of θ.

The comparison between propositions 1 and 3 suggests how it would be possible to test whether individuals differ in their motivation to take a test. Proposition 1 implies that if all test takers have the same valuation of the test, then test scores would provide correct relative ranking of individuals according to their skills, regardless of the incentives provided to them. However, if individuals differ in their valuation of the test, then test scores provide correct relative ranking of individuals according to their skills only for individuals with the same test valuations (i.e., the same θ), but not for the population as a whole. Thus, the provision of incentives may change the relative ranking of individuals according to their test scores. Similarly, the comparison between propositions 2 and 4 suggests that first order stochastically domination of test scores of one group over the another may change with the provision of performance-based incentives. Thus, there are at least two ways to investigate whether the coding speed test, when administered without performance-based incentives, provides correct ranking according to participants' skills. The first is to directly investigate whether ranking are changing. Alternatively, try to identify two groups for which different test scores imply different relationships

[16] The test scores would provide a correct ranking according to skill only if there is no overlap between the test scores of the different types.

between their respective abilities.

3 Data

The data set used in section 6 of the paper is the National Longitudinal Survey of Youth 1979 (NLSY). A nationally representative sample of over 12,000 individuals between the ages of 14 to 22 was first surveyed in 1979. The sampled has been re-surveyed annually until 1994, and then biannually. The NLSY includes questions on wide range of topics, including family background, education, work histories, and annual earnings. For present purposes, this source is unusual in combining detailed labor market data with a battery of standardized tests. Since the NLSY is a well-known survey, this section will focus on aspects particular to this paper. Specifically, I described in the detail tests given to the NLSY participants and the tests being used in the analysis. Details regarding the sample restrictions and variable construction can be found in Appendix C.

Throughout the analysis I look at several tests

The ASVAB - The NLSY contains information about the ASVAB. The ASVAB test was administered to the participants of the NLSY between June and October of 1980, as part of a large-scale social research project; The Profile of American Youth 80 (PAY80). The project was sponsored by the U.S. Department of Defense (DoD) with the cooperation of the U.S. Department of Labor (DoL). The purpose of this project was to assess the vocational aptitudes of contemporary American young people and to establish a national norm for the ASVAB, which is the exam prospective enlists need to take before they can be recruited to the armed forces. Participants in the NLSY were paid $50 honorarium for completing the test.[17] However, no direct performance based incentives were given to participants.[18] Thus, for the NLSY participants the ASVAB is a low-stakes test.[19] The ASVAB test consists of 10 subtests that are described in Table 1. The NLSY data set contains the raw scores on all subtest of the ASVAB and these will be used in the analysis.[20]

The AFQT - The scores of four of ASVAB subsets (word knowledge, paragraph comprehension, arithmetic reasoning, and mathematics knowledge) are added to produce the Armed Forces Qualification Test (AFQT) scores. The AFQT scores are the most commonly used test scores in analysis utilizing the NLSY data set (see for example Herrnstein and Murray 1994; Heckman 1996, Neal and Johnson 1996, 1998).

The Coding Speed Test - Throughout the paper, I look at the coding speed subtest of the

[17] "...The decision to pay an honorarium was based on the experience in similar studies, which indicated that an incentive would be needed to get young people to travel up to an hour to a testing center, spend three hours or more taking the test, and then travel home. The honorarium was set at $50. It has been anticipated that the monetary incentive offered for participation in the aptitude profile study would counteract attrition of the NLS sample..." (Profile of American Youth: 1980 National Administration of the ASVAB, 1982, p. 12).

[18] Some indirect incentives may have been provided by promising participants that in a future date they will get their own test scores, which may help them makes plans for their future.

[19] However, for individuals wanting to enlist the AFQT serves as the entrance exam for the armed forces, and thus for these individuals it is a high-stakes test.

[20] For each subtest the NLSY data set contains also a set of standardized test scores. These are the raw test scores in which low test scores are bunched together and then normalized.

ASVAB to find direct and indirect evidence to the relationship between test-taking motivation and test scores. In the coding speed test participants are given a key, which consists of four-digit numerical codes that correspond to different words. Each key includes 10 words and their respective codes, the questions associated with a specific key consist of 7 words taken from the key. In each of the questions, test-takers are asked to find the correct code for a given word from five possible codes. The instructions and an example of the questions asked in the coding speed test are given in Figure 1. NLSY participants took the paper and pencil version of the coding speed test, in which the test lasts for 7 minutes and consists of 84 questions.

The model suggests that individuals with higher ability may invest higher levels of effort trying to solve any test, holding all else constant. In addition, if individuals differ in their test-taking motivation, regardless of their ability, they will invest different levels of effort in solving the test. Thus, a test for which all individuals have the knowledge necessary to correctly answer all questions would serve as a cleaner measure of differences across individuals in test-taking motivation.

It seems likely that everyone that knows how to read has the knowledge to correctly answer questions on the coding speed test. Therefore, the coding speed test scores may serve as a good measure of test taking motivation. However, given that the time allotted to the coding speed test is short, it may be the case that not everyone will be able to achieve a perfect score. Thus, the coding speed test may also measure ability related to speed. This ability may be different than the one that is being measured by the AFQT (for example Heckman (1996) suggests that the coding speed and numerical operation tests measure fluid intelligence or problem solving ability. It is important to note that even though the name coding speed may suggest some complicated reasoning task, test takers do not need to infer the relationships between the words and the numbers, the relationships are given in the key. Thus, this test is not similar to IQ tests in which test takers need to infer relationships between letters/symbols and numbers.

The ASVAB contains another speeded test that may seem a priori appropriate to use as a proxy for motivation – the Numerical Operation test. The numerical operation test includes 50 very simple algebraic questions (i.e., 2+2=?, 16/8=?, etc.) and lasts for 3 minutes. However, it is possible, though unlikely, that some individuals may not have the knowledge necessary to correctly solve the test questions. The more serious concern is that individuals with high mathematical skills will try harder to solve the Numerical Operations test than individuals with low mathematical skills. Thus, the numerical operation test scores may include a larger component of knowledge than the content of its questions may suggest. In addition, about 16% of NLSY participants correctly solved at least 90% of questions on the numerical operation test. In contrast, only 1% of NLSY participants correctly solved at least 90% of the coding speed test questions. Thus, the coding speed test may serve as a better measure since its range is unrestricted. Psychologists investigating motivation (or the lack of it) suggest that its effects are more pronounce the longer the task lasts (see Revelle, 1993). The coding speed test lasts more than twice as much as the numerical operation test, suggesting that if there are effects for test-taking motivation they may be more pronounced when using the coding speed test.

Since the AFQT is composed of 4 tests that together last for 83 minutes, it seems likely that the longer the test is, the more similar test-taking motivation effects will be to the ones in the AFQT.

4 Indirect Evidence from the Armed Forces

If the lack of performance-based incentives results in lower test scores when tests are administered to survey participants, then higher test scores are expected when the same tests are administered to highly motivated population, everything else equal. Moreover, if the coding speed test is a good measure for test-taking motivation the effect should be more pronounce for this test. I take advantage of the fact that the ASVAB is the screening exam for individuals who want to enlist. Unlike the NLSY participants that were not provided with performance-based incentives, applicants to the armed forces have, at least in principal, the incentives to do well on the ASVAB exam. This is far from being a perfect measure, because the two populations are not identical on all dimension besides motivation. In particular, there may be differences in terms of educational attainments and racial composition between the two populations. Nevertheless, the comparison between the NLSY participants and armed forces applicants may serve as an indication whether the effect exists.

As it turns out, there have been substantial problems in comparing the test scores on the coding speed test (and the test scores on the numerical operation test) between the NLSY participants and potential recruits to the armed forces. Specifically, potential recruits had scored higher on the coding speed test (and the numerical operation test) than the NLSY participants. These problems were first discovered and documented by Maier and Sims (1983) when establishing the national score scale for the ASVAB.[21] The authors show that while potential recruits for the armed forces score higher on the speeded test (i.e., coding speed and numerical operations tests) than the NLSY participants who were born before 1/1/1962, they did worse on any other test. The latter part is to be expected since the NLSY participants were more educated than the population of applicants for enlistment (Maier and Hiatt, 1986).[22] The former part was not.[23]

[21] The reason the speeded tests attracted so much attention in the military was that the numerical operation test was part of the AFQT that serves as the "entrance exam" to armed forces.

[22] It is not clear however whether the participants NLSY were scoring as high as could expected given their education level.

[23] However, by 1985 the problem was considered solved by the military (Maier and Sims, 1986). The differences in the scores between the NLSY youth population and potential recruits was attributed to differences in the shape of the answer space (the NLSY participants had to fill a "bubble sheet" while the potential recruits needed to fill "slim rectangles") and in the layout of the answer sheet (the ASVAB answer sheet used by the military corresponded exactly with the layout of the questions). Wegner and Ree (1990) have shown in a large-scale experiment that potential recruits do worse on "NLSY answer sheet" than on the military one. Comparing the two groups of potential enlists Wegner and Ree (1990) have found that the gaps in test scores between the "NLSY answer sheet" and the military one increase with GT test scores (the sum of arithmetic reasoning, word knowledge and paragraph comprehension standardized test scores). This makes sense if one thinks that people with higher GT test scores work faster on average, then their speeded test scores would suffer the most if it takes them longer to record their answers. However, Maier and Sims (1983) have shown that gaps in test scores on the speeded test between the NLSY youth and potential recruits actually decrease with GT. Moreover, the way the experiment was conducted potential recruits did not have incentives to do well on the speeded tests used in the research. Participants in the research were asked to take the coding speed and the numerical operations tests relating to the research before taking the whole ASVAB (including these two subtests). In addition, all individuals that displayed significant change in speeded test scores between the research part and the ASVAB part were deleted

Infact, in 1984, with the introduction of new ASVAB forms, the old problems with the speeded tests got even worse (Maier and Hiatt, 1986), and resulted in the recommendation that the numerical operation subtest would be taken out of the AFQT.[24] Maier and Hiatt (1986) suggest that the gaps on the speeded tests are the results of "test taking strategies" among which they count: "work as fast as possible" and "keep your attention focused on the problem" (Maier and Hiatt, 1986, p.5). They add: "...The extent to which all applicants use the same test-taking strategies in not known. What is known is that the 1980 Youth Population generally did not know or follow these strategies..." (Maier and Hiatt, 1986, pp. 5-6). It seems unlikely that NLSY participants did not know these test taking strategies, however, they may not care enough to follow them.

Maier and Hiatt (1986) provide a conversion between the 1980 Youth Population (i.e., all NLSY participants who were born before 1/1/1963) coding speed test scores[25] and the 1984 applicants for enlistment.[26] The conversion was done by setting the raw scores that had the same cumulative frequency, conditional on measure of ability, equal to one another. The ability measure used was the HST composite scores, which is the sum of arithmetic reasoning, word knowledge and paragraph comprehension and mechanical comprehension standardized test scores (Maier and Hiatt, 1986, Appendix A, pp. A1-A10). Figure 2 presents the coding speed test scores by HTS category for 3 groups of males: NLSY civilian sample born before 1/1/1963, NLSY military sample, and the 1984 military applicant sample. The test scores for the 1984 military sample were calculated using the 1980 Youth Population (civilian and military samples) weights with the coding speed tests scores implied by the conversions presented in Table A-4, Appendix A, pp. A9-A10 in Maier and Hiatt, 1986.[27,28] Figure 2 clearly displays that for every single interval of the HST scores potential recruits have higher test scores than the participants in NLSY civilian sample.[29] Moreover, this is also true for the distributions. Figure 3 presents the cumulative coding speed test scores for the 3 groups (i.e., the NLSY civilian sample born before 1/1/1963, the NLSY military sample and the 1984 military sample).[30] Again, it is clear from the figure that the NLSY civilian population has the lowest test scores, in particular for the lower 80%

from the data (Wegner and Ree, 1990 reported less than 2%).

[24]Indeed in 1989 it was taken out of the AFQT.

[25]Maier and Hiatt (1986) also provide the similar conversion for the numerical operations test scores.

[26]This group is known as the Initial Operational Test and Evaluation 1984 (IOT&E 1984), was tested in October and November of 1984 when the new forms of the ASVAB (forms 11, 12, and 13) were introduced.

[27]Unfortunately, Maier and Hiatt (1986) do not provide the test score distributions for the military sample they have been using. However, given that they provide the conversion between the test scores of the two populations it is possible to reconstruct the military test scores. Even if their conversion has problems, it should still give the correct means for the military IOT&E 1984 sample conditional on the HST intervals they have been using.

[28]There were two decisions to be made while reconstructing the IOT&E 1984 coding speed test scores. The first Maier and Hiatt (1986) do not provide an equivalent to the test scores of zero, however there were 12 NLSY participants that had this test score. Given that we are interested to see whether the potential recruits have higher coding speed test scores I have set the equivalent test score to zero. The second, Maier and Hiatt (1986) report a range for coding speed test score of two (2-10), in the constructing the military coding speed I have again taken the lowest value (2).

[29]This is also the case when looking at the all male NLSY participants who were born before 1/1/1963 (i.e., the ones belonging to either the military or the civilian NLSY samples).

[30]To construct the cumulative distribution of the coding speed test scores I have used the same procedure reported above. To construct the correct relations between the different HST intervals, I used Panel A in Table A1 in Maier and Hiatt (1986), which reports the relative weights for each of the HST intervals for the IOT&E 1984 and the 1980 Youth Population.

of the test scores distribution.[31] If indeed the coding speed test scores measure effort this is exactly what we would have expected if the potential recruits are highly motivated to take the ASVAB while (not all) the NLSY participants may be.

Thus, the comparison between potential recruits to the armed forces and the NLSY participants provides us with the first indication that the coding speed test scores may measure motivation to take the ASVAB test. However, there may be other possible explanations to account for the differences between the NLSY participants and potential recruits (one of which is obviously selection). Therefore, in order to gather direct evidence that indeed motivation plays an important role in determining the coding speed test scores I turn to the results from the controlled experiment.

5 Experimental Evidence

The model implies that to test whether individuals vary in their test-taking motivation one needs to investigate whether there is a change in the relative ranking of individuals according to their test scores under different incentives schemes. Improvement between tests and even narrowing of gaps between groups is feasible even of all individuals have the exact test-taking motivation but differ in their ability. However, relative rank changing is only possible if individuals have different test-taking motivation. In order to look for change relative rank one needs to look at test scores for the same individuals under different incentives schemes. This is easily and naturally done in the lab. In the next section I describe the experiment I conducted to investigate this question and its results.

5.1 Experimental Design

I conducted an experiment in which participants solved a computerized version of the coding speed test. There are 2 treatments in the experiment: main and control. In each of the treatments participants solved three versions of the coding speed test. Each test lasted 10 minutes and included 140 questions.[32]

The experiment was conducted at Harvard using the CLER subject pool and standard recruiting procedures. Overall there were 127 participants in the two treatments: 99 in six sessions for the main treatment (50 men and 49 women) and 28 (14 men and 14 women) in the control.

Each participant received a $10 show-up fee, and an additional $5 for completing the experiment. Participants were told in advance how many parts the experiment will include (two in the main treatment and three in the control), and that one of these parts will be randomly chosen for payment at the end of the experiment. Participants were informed of the tasks they need to perform in each

[31] Unfortunately, Maier and Hiatt (1986) do not provide any summary statistics on the IOT&E 1984 sample, so it is impossible to test whether the two distributions are equal.

[32] The tests were constructed in the following manner. For each test, 200 hundred words were randomly chosen from a list of 240 words, and were then randomly ordered to construct the 20 keys. For each word in the keys a random number between 1000 and 9999 was drawn. Of the 10 words in each key 7 were randomly chosen to be the questions. The possible answers for each question were then randomly drawn (without replacement and excluding the correct answer) from the 9 remaining possible numbers in the key. Then the placement of the correct answer (1-5) was drawn, and the correct code was inserted in this place. All participants saw the same tests, and there is not reason to believe given the construction process describes above that any of them was harder.

part only immediately before performing the task. The specific compensations and order of tasks were as follows. The instructions for the two treatments are given in Section C1 in Appendix C.

Main Treatment

Part 1 – Fixed Payment: Participants were told that their task would be to solve the coding speed test. If task 1 is randomly selected for payment, they receive \$10. They were told that first a practice test will be given.

There are two reasons to have the practice test. The first is as a control for learning. If learning occurs, and it is fast, then it will be contained for the duration of the practice test. In addition, if it turns out that learning is not an issue, then the practice test can serve as another measure of test scores for a test in which subjects may have been less intrinsically motivated than in the \$10 test. In the analysis that follows I refer to these two tests as practice test and \$10 test.

Part 2 – Piece Rate Compensation: Participants were told that they would solve another version of the coding speed test. Below, I refer to this test as the incentives test. They were given a choice between payment based on their performance on the \$10 test and a payment based on their future performance on the incentives test. If task 2 is randomly selected for payment, then their payment is the following. If they choose to be paid according to their past performance (i.e., on the \$10 test) they receive \$10*(the fraction of \$10 test questions solved correctly). If they choose to be paid according to their future performance (i.e., on the incentives test) they receive \$30*(the fraction of incentive test questions solved correctly).

The main purpose of the experiment is to figure out whether the coding speed serves as measure of effort or motivation. If participants are choosing the piece rate, then it can serve as a indication that the incentives scheme is desirable, at least to some degree. If even after choosing the piece rate scheme some participants do not improve their performance, then this may indicate that they invested high levels of effort even without performance based incentives.

Control Treatment

All the parts in the experiment were identical. In each, participants were told that their task would be to solve the coding speed test. There were told that if the current part is randomly selected for payment, they receive \$10.

Survey: After participants finished the two parts they were ask to answer a demographic survey, and a personality survey, designed to detect the "big 5" constructs.[33] The Personality survey and its instructions are given in Section C2 in Appendix C.

The Testing Program

It is worthwhile in this stage to discuss the some features of the testing program that would come handy in the analysis of the experimental data. Figure C1 in Appendix C depicts a typical screen of the test. The key and the answers are on the left hand side of the screen, while the answer sheet (an electronic version of a "bubble sheet") is on the right hand side of the screen. To answer a question participants need to press the on one of the radio buttons associated with this question. The test

[33] I discuss the "big 5" constructs in detail below.

was design to match as closely as possible the paper and pencil version of the coding speed test. Thus, participants were able to freely move between different sets of keys and answers. Specifically, in order to see the next key and the answers associated with that key, participants need to press the "Continue" button. Similarly, they can see previous sets of keys and answers by pressing the "Go Back" button. Moreover, participants were able to freely move between all the answers on the answer sheet by pressing the "Next" and "Previous" buttons. The program records all the answers given when any of these buttons has been pressed. This design of the testing programs makes it possible to detect guessing. Specifically, if participants provided answers to questions that they did not yet see, they have to be guessing. In addition the program records all answers given in a 30-second interval. Thus, if participants were guessing, it is possible to identify the 30-second intervals in which they have been doing so. Moreover, for each participant it is possible to know how many questions they have answered correctly in up to twenty 30-second periods.[34]

5.2 Basic Experimental Results

All the 99 participants in the experiment chose the piece rate compensation scheme in part 2 of the experiment. Thus, the results reported below include all participants. Table 2 reports the means and the standard deviations of performance for the three tests. In the first 3 columns, performance is measured by the number of total correct answers on the test. In the last 3 columns, the measure of performance is the number of correct answers per 30-second period in the periods before the first guess.[35] Participants' performance have improved significantly between the tests. Between the practice and \$10 tests participants correctly solved on average 13.8 more questions, which is a significant improvement in performance (a one-sided t-test allowing for different variances yields $p < 0.001$). The improvement between the practice and \$10 tests is also seen in the number of correct answers in the 30-second periods before the first guess. Between the first two tests participants improved significantly by 0.82 correct answers per 30 second (a one-sided t-test allowing for different variances yields $p < 0.001$). Between the \$10 and incentives tests participants improved even further, and correctly solved on average 8.2 more questions. Again, this is a significant improvement in performance; a one-sided t-test allowing for different variances yields $p = 0.003$. Participants also correctly solve significantly more questions per 30-second in the incentives test than in the \$10 test. On average, between the two tests participants improved by 0.32 correct answers per 30 second (a t-test allowing for different variances yields $p < 0.001$).

In addition, when looking at the total number of correct answers participants provided in each test we see a pattern emerging for the variance. The variance in test scores is the largest for the \$10 test. It increases by 77% in comparison to the incentives test and by 54% in comparison to the practice test (a two-sided F-test yields $p = 0.033$ for equality of variances between the \$10 and incentives tests

[34] Not all individuals chose to guess in the last periods. See Table C1 in Appendix C for details.

[35] The periods after guessing provide a very noisy measure of participants' performance. In particular, it is impossible to figure out how many questions one answered correctly in these periods since some of the questions were already guessed correctly. Therefore, only the periods before guessing are used.

and p = 0.005 for equality of variances between the $10 and the practice tests).

Figure 4 presents the cumulative distributions of the total test scores for the three tests and Figure 5 reports the corresponding cumulative distributions of the number of correct answers in the 30-second periods before participants started guessing. The patterns seen for the means are also present for the distributions. In particular, there is an improvement in the distributions from the practice test to the $10 test. A matched sign rank test yields $p < 0.001$ for the test scores as a whole, and a Mann-Whitney for the number of correct answers per 30-second period yields $p < 0.001$. Similarly, there is an improvement in the distributions from the $10 test to the incentives test (a matched sign rank test yields $p < 0.001$ for the total test scores, and a Mann-Whitney test for the number of correct answers per 30-second period yields $p < 0.001$).

The improvement in performance may be in response to the incentives scheme. However, it is possible that there is something in the nature of coding speed test that necessitates improvement when individuals takes the test repeatedly. The most obvious possibility is learning. To investigate whether learning by doing, i.e., whether participants become more efficient in solving the coding speed test as they solve more questions, can explain the improvement between the tests we are interested only in periods in which participants tried to solve the test, i.e., they were not guessing.[36] Thus, we look at the 30-second periods before the first guess.

Table 3 presents the coefficients from regressions of the number of correct answers in each test in the 30-second periods before guessing on period number and period number squared controlling for individual fix effects.[37] Evidence for learning exists only for the practice test, and there the relation between the number of correct answers and time is concave; after about 7 minutes, the number of correct answers is decreasing with time. Thus, the largest increase in test scores per 30-seconds is 0.35 per 30-seconds. In comparison, the average increase in performance between the practice and $10 tests is 0.82 per 30-seconds. Thus, learning may account for less than half of this increase. In both the $10 and incentives tests there is a decrease in the number of correct answers per 30-second period, and no indication of improvement at all.[38] These results suggest that learning within the tests is not a real problem, and in particular in the $10 and incentives tests. Participants do not seem to experience improvement within the tests instead they seem to improve between the tests.

The improvement between the three tests may happen every time participants take the coding speed test repeatedly. It may be attributed to learning by assimilation, i.e., participants come up with better strategies to solve the test, which would occur whenever participants take the test repeatedly.[39]

[36] Guessing could certainly contribute to one's total test scores. However, the possible increase in test scores resulting from guessing does not indicate that one learned how to do the task better. On the contrary, even without understanding the task one can always guess. Thus, the decision to guess is treated here as another strategy to get higher scores on the test and not as part of learning by doing.

[37] The error term in these regressions is the result of the measurement error introduced by the 30-second intervals.

[38] This decrease in the number of correct answers is common finding in the psychological literature (see for example Revelle, 1993).

[39] The testing program allowed participants to contemplate on possible strategies to solve the test by allowing each participant to start the test whenever he or she felt ready to do so. Nevertheless, it is possible that one needs to experienced the test first hand.

To investigate this question I have run the control treatment, in which participants solved the coding speed test three times for a fixed payment. The results of the control treatment are different than the ones in the main treatment. Specifically, in the mean performances (standard deviations) were 90.3 (21.1), 93.64 (26.5), and 88.5 (31.7) for the first, second and third test, respectively. Thus, participants have actually experienced an insignificant decrease of 5.1 correct answers on average between the second and the third time they took the test (a one-sided t-test allowing for different variances yields $p = 0.26$) instead of a significant increase of 8.2 in the main treatment. Moreover, between the first two tests there was an insignificant improvement of 3.3 correct answers on average (a one-sided t-test allowing for different variances yields $p = 0.3$) instead of a significant increase of 14 correct answers on average in the main treatment.[40] In addition, while the distribution of test scores is the same in the first test for two treatments (Mann-Whitney test yields $p = 0.90$), it differs for the last two tests (Mann-Whitney test yield $p = 0.04$ for the second test and $p < 0.001$ for the third test). Thus, I conclude that improvement between the tests should be attributed to the effect of incentives.

5.3 Change in Relative Ranking

We have seen that participants significantly improved their performance between the three tests in response to the incentives provided to them. This improvement in performance indicates that when no incentives are supplied, we underestimate individuals' capabilities. If we would like to know how many coding speed questions individuals can correctly answer in 10 minutes, the test scores in both the practice and the $10 tests underestimate this number. The reason is that on average participants invest less than the maximum effort in trying to solve tests when no incentives are supplied. Even if we cannot get a correct measure of individuals' capabilities, it still possible that the coding speed test scores provide a correct ranking of individuals according to these capabilities. The improvement in performance documented above does not indicate itself that the coding speed test scores do not provide a correct ranking of individuals according to their ability when no performance-based incentives are supplied to participants. Nor does it indicate that the coding speed test measures test-taking motivation, regardless of participants' abilities. In order to investigate these questions we need to look at the ranking of individuals.

If indeed individuals do not differ in their test-taking motivation, regardless of their ability, and test scores are produced using effort and ability, then the provision of incentives will not change the relative ranking of individuals. Equivalently, two participants that had the same test scores in one test will have the same test scores in another test. Thus, there are at least two ways to investigate whether the coding speed test, when administered without performance-based incentives, provides

[40] In addition I ran a simulation using the total test scores participants got in the main experiment. I randomly drawn, with replacement, a group of 14 men and 14 women and calculated the mean improvement in their test scores between every two consecutive tests. The exercise was repeated a 1,000,000 times. The probability that participants in the main experiment would experience on average an increase between the practice and the $10 tests smaller than 4 is 0.0002. The probability that they would experience on average a decrease in tests cores between the $10 and incentives test of 5 or more is less than 0.0001. This serves as another indication that the improvement in the main experiment is due to the provision of incentives and not to learning occurring between tests.

correct ranking according to participants' ability.

5.3.1 Do Different Tests Allow for Comparison Between Participants' Ability?

If coding speed test scores would have provided a correct ranking of individuals according to their ability then two individuals who had the same test scores on a given test have the same ability. Thus, they should have the same test scores on any other test, regardless of the incentives. This measure does not capture all the change in rankings. In particular, if two individuals did not look the same in one test, it is possible that the relative ranking had changed but they still do not look the same on another test. Pairs like this would not be capture by this measure, and therefore it serves as a lower bound for the amount of rank changing.

To construct this measure, we face the problem of what does it mean to have the same test scores. Do total test scores separated by one correct answer different or not? Here the performance on the 30-second periods before participants start guessing is more natural, as we can use a statistical definition to determine whether two mean performance are the same or not.[41] The criterion adopted was the following. For all possible pairs of participants (4656)[42] I use a t-test allowing for differences variances to test whether their mean performance is different in the 5% significant level in any two tests. I then count number of pairs for which the mean performance is significantly different in one test but not in the other. Of the 4656 possible participants' pairs 56.7% (2642 pairs) have performance which is not significantly different in either the \$10 or the incentives tests or both. Of those 2642 pairs, 51.1% (1349 pairs) have significantly different performance on the \$10 test but not on the incentives test and vice verse. Similarly, of the 4656 possible participants' pairs 60.6% (2823 pairs) have performance which is not significantly different in either the practice or the \$10 tests or both. Of those 2823 pairs, 53.7% (1517 pairs) have significantly different performance on the \$10 test but not on the practice test and vice verse. [43]

Given that the 5% significant level has been used we would have expected at most 10% of all participants that have the same mean performance in one test to have different mean performance in another. I find that at least 51% participants to have the same mean performance in one test but not in another, i.e., five times as much. Therefore, we have the first indication that the coding speed test does not provide the correct ranking according to participants' ability. Even in a very restrictive measure of rank changes, for a non-negligent fraction of participants the coding speed test scores do not provide correct indication regarding their relative abilities.

[41] Without adopting a criterion for test scores differences, it is impossible to use the total test scores to answer this question.

[42] For two participants, one in the \$10 test and one in the incentives test, who have started guessing in the first and the second periods respectively, it is impossible to construct this measure of performance.

[43] Of the 4656 possible participants' pairs 62.5% (2909 pairs) have performance which is not significantly different in either the practice or the incentives tests or both. Of those 2909 pairs, 54.1% (1574 pairs) have significantly different performance on the practice test but not on the incentives test and vice verse.

5.3.2 Does the Relative Ranking Change?

To further provide evidence on the change in ranking, I examine the change in ranking according to total test scores. To do so, I assign within each test a higher rank to participants with lower test scores. Thus, all the participants with the highest test scores are assigned the rank 1, all the participants with the second to highest test scores are assigned the rank 2, and so on and so forth. This is a very conservative measure of relative rank change. In particular, if a participant improved her test scores between any two tests and now she is one of the participants with the highest test scores, only her rank will be changing. The maximum rank varied across the tests. The maximum rank is 56 in the \$10 test, 51 in the practice test and 49 in the incentives test. It is now possible to ask whether the relative ranking is changing between the tests. There are several possible ways to investigate relative rank changes. The first is just to look at the percentage of participants who have changed their relative rank. However, it is not clear whether small differences in test scores actually indicate to real differences in ability. Thus, we may want to look at the percentage of participants who have changed their ranking above a certain threshold. In addition, it is possible to look at the distributions of the absolute rank change to get a sense for how big the relative changes in ranking are.

Table 4a reports the percentage of participants who experienced an absolute change in relative ranking bigger than 4, i.e., a change that move them outside a decile of ranks centered around their own ranking, the mean of the absolute change in relative rank, and the maximum absolute change in ranks. As can be clearly seen in the table, substantial fraction of participants changed their relative ranking between the tests. In particular, at least 53% of participants changed their relative ranking by more than 4 ranks or less than -4 ranks. In addition, the average participant experience an absolute change her relative ranking by more than 6 ranks between any two tests.

Are there other any other factors that may create the change in relative ranking? There are three obvious sources that may create noise that in turn may create rank changes. The first has to do with having a good/bad day, the second has to do with the test-questions themselves, and the third is guessing. Since the whole experiment lasted for about an hour, it seems that participants that were having a good or a bad day would have had it for the duration of the experiment. Therefore, having a good/bad day cannot generate the observed change in relative ranks. Usually in tests that measure knowledge a possible source of noise is the specific questions asked. In particular, test-takers may get higher test scores in one test and not in another since some questions on a test may relate to a specific firsthand knowledge that they may have (for example, if a test taker spent the summer in England, she may know that London is its capital, even though she cannot answer any other question about capitals of other countries). However, this is not the case for the coding speed test, as all the questions require identical knowledge (recognize words and numbers).

Participants being lucky or unlucky in their guesses may cause some of the changes in relative ranking. To investigate this possibility I look at participants' performance in the 30-second periods

before they started guessing. However, participants guessing behavior varies. Specifically, both the period in which they started guessing (if at all) and the length of time spent guessing vary. Thus, it is impossible just to compare participants test scores by just excluding the periods in which they have been guessing. Therefore, in order to compare participants' performance taking away the random component introduced by guessing I use the average number of questions correctly solved in the 30-second periods before they started guessing.[44] To make the ranking comparable to one reported in Table 4a, I then assume that participants would have had this average performance for the duration of the test, i.e., for 10 minutes or for twenty 30-second periods had they not started guessing. This is a very strong assumption. In terms of test-taking motivation, it means that test-taking motivation has no effect on participants' boredom and fatigue as the test progresses. However, psychologists investigating motivation claim that lack of motivation causes boredom and fatigue that increase in the length of the task (Revelle, 1993). To further make the resulting test scores comparable to the original ones, I then truncate the test scores at 140, and I allow the test scores to take only integer values.[45] Ranks are assigned given this the new set of test scores. As before, all the participants with the highest test scores are assigned the rank 1, all the participants with the second to highest test scores are assigned the rank 2, and so on and so forth. The maximum rank is 53 in the \$10 test, 46 in the practice test and 49 in the incentives test.

Table 4b reports the percentage of participants who experienced an absolute change in relative ranking bigger than 4, the mean of the absolute change in relative rank, and the maximum absolute change in ranks. Even once the guesses are taken out of the test scores, substantial fraction of participants changed their relative ranking between the tests. In particular, at least 48.5% of participants changed their relative ranking by more than 4 ranks or less than -4 ranks. In addition, the average participant experience an absolute change her relative ranking by more than 5 ranks between any two tests. Even though the numbers are somewhat different neither the percentage participants experiencing an absolute rank change bigger than 4 nor the mean of the absolute rank change are different across the two ways defining performance. In addition, the distributions of the absolute change in ranks are not statistically different (Mann-Whitney tests yield $p = 0.33$ for the absolute rank change between the \$10 and incentives tests, $p = 0.97$ for the absolute rank change between the practice and \$10 tests, and $p = 0.53$ for the absolute rank change between the practice and incentives tests). Thus, the change in ranking between the different tests does stem from random error associated with the guessing.

It is possible to use the 30-second periods before participants start guessing to investigate whether the rank changing reported above stem from another source of noise. Specifically, it possible to resample from the 30-seconds periods before participants started guessing to create potential test

[44] As was mentioned above, for two participants it is impossible to construct this measure of performance.

[45] The magnitudes of the absolute rank changes as well as the percentage of participants who experienced an absolute rank change bigger than 4 is much larger than the numbers reported in Table 4a when the ranking is being determined by participants average performance in the 30-second periods before they started guessing. This is the results of the fact that the average performance provides a finer partition of participants to different ranks as it is not the case that multiplying it by 20 gives an integer less than, or equal to, 140.

scores in a given test, and hence potential relative ranking, and see whether noise can explain the change in relative ranking between the tests. Doing this we implicitly assume that the noise is being drawn from some distribution in every period. Here some cautious is warranted. First, the division of the test to 30-second periods itself creates measurement error. Even if all participants solve the test in a constant pace, this pace does not necessarily coincide with the 30-second periods. Therefore, in order to reduce this measurement error, I look at 1-minute periods and 2-minute periods. Second, we want to distinguish possible noise from lack of test-taking motivation. The latter may explain rank changing within the first two tests, in particular if its effects are more pronounced in different point in time during the test. However, lack of motivation should not explain rank changing for the incentives test.[46] Therefore, I look at rank changing implied by what individuals did in the incentives test to test whether the rank changing reported above could stem from this particular noise. As long as the noise generating process is the same in all the tests (or that the noise is the most detrimental to test scores in the incentives test) this could serve as a good bound for the effects of noise.

To simulate the rank changing within the incentives test for each participants, I resample twice from the periods before she started guessing. To reduce the measurement error I re-sample either ten 1-minute periods or five 2-minutes periods. I add the number of correct answers in each period to construct two sets of test scores for each participant. I then construct for all participants two sets of relative ranking, where the highest test scores are ranked 1, the second highest test scores are ranked 2, and so on, and compute the absolute rank changes between the two sets of scores. To compare to the rank change between the practice and \$10 tests, and the \$10 and incentives tests, I repeat the process above now drawing also from the practice and the \$10 tests. I then repeat the whole procedure 10,000. Figure 6 depicts the CDF's of absolute rank change obtained from simulations using the 1-minute and the 2-minute periods. It is clear from the figures that the amount of rank changes between the tests is much higher than the amount of rank changes within the tests. In particular, we can ask what fraction of absolute rank changes is bigger than x, where x = 1,2..., for each rank change simulation, and in what percentage of the simulations this fraction was bigger within the incentives test than between two tests. For the 1-minute periods simulations, in at most 6% of the simulations the fraction of absolute rank changes bigger than one is bigger within the incentives test than between the tests. The percentage is reduced to less than 1% when looking at fraction of absolute rank changes bigger than three. For the 2-minute periods simulations, this fraction is less than 1% already when looking at absolute rank changes bigger or equal to one. While there are some rank changes implied by the simulations within the incentives test, the magnitudes are significantly smaller than the rank changes between the \$10 and incentives tests and between the practice and \$10 tests. Therefore it seems likely that the rank changing between the tests cannot be attributed noise.

The last two sections provide evidence that the coding speed test does not provide the correct ranking according to participants' ability. If it were, there should have been no rank changes between

[46] Although, one needs to keep in mind that if participants started guessing after periods in which they their pace fell, not all time periods are interchangeable, which may create additional rank changing.

the different tests (outside of noise). Moreover, the change in ranking between the practice and the $10 tests suggests that the change in ranking cannot only stem from individual differences in the valuation of money. In particular, in terms of monetary incentives nothing has been changing between the practice and $10 tests. In both cases monetary incentives were flat. However, participants did change their behavior. We have seen that just repeating the test three times or learning within the tests cannot account for the increase in test scores between these two tests. Instead, some participants may be trying harder when the test is called "The Test" while others do not. Thus, it has to be the case that some participants value the test scores higher in the $10 test than in the practice test. This by itself implies that the test scores are correlated with an increase in effort unrelated to participants' ability (which is the same in both tests).

5.4 Do Different People React Differently to Incentives?

In the last sections we have seen that the relative ranking of individuals changes between the tests. This suggests that individuals differ in their test-taking motivation. In particular, it implies that less motivated individuals respond the most to performance-based monetary incentives. To investigate this question, I investigate individuals' performance in the different tests . The measure of performance I use is the mean number of correct answers in the 30-second periods before participants started guessing in each test.[47] Between the $10 and the incentives tests 37 participants out of 99 significantly improved their own performance,[48] while the other 62 participants did not (of these 62 participants only two experienced a significant decline in their performance). I divide participants into two groups according to whether or not their own performance significantly improved between the $10 and the incentives tests.[49] This is a measure of individual performance across the tests. While we know that individuals in one group significantly improved their performance, the relations between the total test scores distributions of the two groups across the different tests cannot be defined theoretically.

Figure 7 presents the cumulative distribution of the total test scores of the two groups in the different tests. Panel A present the total test scores for the incentives test. The figure suggests that the once monetary incentives are supplied there is no difference between the two distributions of total test scores. To test for stochastic dominance I follow McFadden (1989). Neither the hypothesis that the tests scores distribution of the participants who significantly improved between the $10 and incentives

[47]Three individuals started guessing very early on in the $10 and incentives tests. Thus looking at the periods before guessing is not informative to test whether they significantly improved between the tests. However, all 3 had multiple periods in the $10 test in which they did not try to solve any question on the test (one guessed the whole test in the first 2 minutes and then ended the test). None of the three experienced in the incentives test, in the periods before they have started guessing, any period in which they did not try to answer questions on the test, or even a period in which they correctly answered no questions on the test. Thus, they all have been classified as experiencing an improvement between the $10 and the incentives tests. The results reported below remain qualitatively and quantitatively the same if I exclude these individuals from the analysis.

[48]The criterion used was significant level of 0.05 or less using one sided t-tests allowing for different variances. As a robustness check I used a significant level of 0.1, and the results reported below remain qualitatively the same.

[49]The 2 participants whose test scores significantly decrease between the two tests are assigned to the group that did not significantly improved. However, the results reported below remain qualitatively and quantitatively the same if I exclude these individuals from the analysis or assign them to the other group.

tests first order stochastically dominates the distribution of the rest of participant can be rejected (p = 0.420) nor the opposite one (p = 0.757). Similar picture arises when we look at the maximum test scores each participant achieved in the experiment (Panel B), which are the best estimate we have of participants' ability. Again, there is no difference in the distributions test scores between those participants who significantly improved between the \$10 and incentives tests and the ones who did not. Neither the hypothesis that the tests scores distribution of the highly responsive participants first order stochastically dominates the distribution of the rest of participant can be rejected (p = 0.266) nor the opposite one (p = 0.839). Thus, the these two panels suggest that both groups have the same underlying ability. A different picture arises when looking at the the total test scores of participants in the practice and \$10 tests (Panels C and D, respectively). Given that the two groups do not differ in their test score distributions in the incentives test and the way participants were selected into the two groups, this result is to be expected. Nevertheless, in accordance with the rank changes we have seen before, we find evidence for rank changes here too. Specifically, The hypothesis that the test scores distribution of the participants who significantly improved between the \$10 and incentives tests first order stochastically dominates the test scores distributions of the rest of the participants can be rejected for both the practice test (p = 0.025) and the \$10 tests (p = 0.002). However, for both tests the opposite one cannot (p = 0.967 and p = 1 for the practice and the \$10 tests, respectively).

These results imply that the participants who significantly improved between the \$10 and incentives tests were not doing their best to solve the coding speed test when no performance-based monetary incentives were provided, while the other group of participants did. Moreover, once performance-based incentives were provided the two groups do not differ in their test scores. Taken together these results suggest that it is not the case that the more able individuals are the ones who are doing better when performance-based incentives are not provided. The differences in the mean performance on the \$10 test are striking. While the first group of participants correctly solved on average 93.4 questions, the second group correctly solved on average 110.6 questions (a t-test yields p < 0.001). This difference is as big as the standard divination across participants in the incentives test. In contrast the difference between the groups in correct answers to the incentives test is 2 (111.1 for the first group and 113.2 for the second, a t-test yields p = 0.57).

Looking at the pattern of improvement between the practice and the \$10 test, each of the two groups of participants identified above can be further divided into two groups. Of the 62 participants who did not significantly improved between the \$10 and incentives tests, 42 (67.7%) have significantly improved between the practice and \$10 tests. Of the 37 participants who significantly improved between the \$10 and incentives tests, 17 (45.9%) have significantly improved between the practice and \$10 tests. A fisher exact test for the equality of the distributions yields p = 0.037 (a chi-squared test for the equality of the distributions yields p = 0.033).[50] This suggest that while some participants

[50]In addition, between the practice and the \$10 tests 59 participants have significantly improved their performance, while between the \$10 and incentives tests only 37 participants did so. A fisher exact test for the equality of the distributions yields p = 0.002 (a chi-squared test for the equality of the distributions yields p = 0.003). This suggests that the improvement between the tests cannot be attributed to noise that is generated by the same process in all tests,

tried their best already in the practice test (20 participants overall, their mean performance in the practice test is 96.25), others needed to to hear that the test is important (i.e., "The ($10) test") in order to try their best (42 participants overall), and yet others needed performance-based monetary incentives (37 participants) to try their best.[51]

The basic experimental results indicate that the variance of the test scores is the largest in the $10 test. The results reported above can provide an explanation. While in the incentives test participants invest high levels of effort and in the practice test most participants invest little effort, in the $10 test the difference in effort levels invested by participants are the largest. While some participants increase their effort levels substantially others did not. As a result the variance in test scores increases.

It seems that when no performance-based monetary incentives were provided there is a group of participants that have chosen to work lass hard than they could. Just looking at their test scores in the $10 test or the practice tests one would label them at low ability individuals. However, once performance-based monetary incentives are supplied, it turns out that they have the same ability distribution as their fellow participants who choose to work hard in the first place. Therefore, we have another indication that the coding speed test scores also measure motivation.

5.4.1 What Do the Coding Speed Test Scores Measure?

In the last section we saw that some participants chose to invest high levels of effort in trying to solve the coding speed test even without performance-based monetary incentives, while others did not. In this section I investigate what traits are correlated with effort choices participants made.

At the end of the experiment participants were asked to fill the "Big 5" questionnaire. The questionnaire includes 50 questions (10 for each of the constructs).[52]. To create the five constructs the answers to the "Big 5" questionnaire were added[53] and then normalized to have mean zero and standard deviation 1 among participants with valid data.[54] The normalization was done separately for men and women, since men and women significantly differ in their answers to the "Big 5" questionnaire. The instructions and the questions are given in Appendix C.[55]

The "Big 5" theory is part of research effort in psychology dating back to the 1930's trying to find the most important ways in which individuals differ from one another. In the last few decades a consensus has arisen among empirical psychologists that few dimensions are enough to describe most

as we would have expected that the fraction of individuals improving between any two tests would stay the same.

[51]While 17 of the participants who significantly improved their performance between the $10 and incentives tests also responded to the cue "The ($10) test", they still only did their best when performance-based monetary incentives were supplied.

[52]The survey was taken from http://ipip.ori.org/newQform50b5.htm

[53]When the question was phrased in a negative manner (for example, "Feel little concern for others") the answer were subtracted.

[54]Participants with valid data are participants who answered all 50 questions in the "Big 5" questionnaire (91 participants). In addition, participants reported their SAT scores (96 participants). Overall data is available for 88 participants, 42 men and 46 women.

[55]The personality test was administered without incentives, mainly since it is unclear how to provide incentives for such a test. If participants just randomly chosen their answers then the resulting test scores would not be very informative about participants personalities.

personality traits. The common classification of personality traits is to five dimensions, which are referred to as the big five.[56] These five factors are usually referred to as Extraversion (E), Agreeableness (A), Conscientiousness (C), Neuroticism (N), and Openness to experience (O). The definitions below follow Roberts, Robin, Caspi, and Trzeniewski (2003). Extroversion refers to the differences across individuals in the tendency to be social active, and assertive. Agreeableness refers to traits that reflect differences across individuals in the tendency to be trusting, modest, altruistic and warm. Conscientiousness reflects the tendency to be rule following, task- and goal-directed, planful, and self controlled. Neuroticism, or its converse emotional stability, contracts the experience of anxiety, worry, anger, and depression with even-temperedness. Openness to experience reflects individual differences in the tendency of individuals to be open to new ideas, complex, original, and creative. The literature regarding the big five relates these traits to various aspects of individuals' life. In particular interest is the relationship found between these constructs and various measures of economic success. As it turns out, conscientiousness is the trait that is most consistently relates to job performance, affective job seeking behavior, retention, and attendance at work (Judge, Higgins, Thoresen, and Barrick, 1999).

We have seen that some participants chose to invest high levels of effort in trying to solve the coding speed test even when no performance based incentives were supplied while others did not. Using participants' answers to the "Big 5" questionnaire it is possible to relate participants' decision to their personality traits. For male participants, the decision of whether or not to invest high effort in the tests without performance-based incentives can be related to their level of conscientiousness. Specifically, the average conscientiousness level of male participants who invested little effort in trying to solve the \$10 test (i.e., those who significantly improved their performance between the \$10 and incentives tests) is -0.32 while the average conscientiousness level of male participants who already invested high effort in trying to solve the \$10 test is 0.29 (a one sided t-test allowing for differences in variances between the two groups yields p = 0.023). No other personality construct is related to individuals' type for either men or women. Moreover, the self reported SAT scores were not related to individuals' effort investment behavior either. The average SAT scores of male participants who invested little effort in trying to solve the \$10 test is 1,429 while the average conscientiousness level of male participants who already invested high effort in trying to solve the \$10 test is 1,457 (a one sided t-test allowing for differences in variances between the two groups yields p = 0.29). Probit regressions in which the dependent variable is one's type (where 1 denote the individuals who invest high effort levels only when performance-based monetary incentives were supplied support this conclusion. The independent variables in this regression are the five "Big 5" constructs and the self reported SAT scores. For men participants, the only significant variable in the regression is conscientiousness level, and 1 standard deviation increase in conscientiousness is associated with a decrease of 17.2% (the p-value associated with this coefficient is 0.045) in the likelihood that a male participant, who has the mean level of personality traits and mean SAT scores, will only invest high effort in solving the coding

[56] For excellent summaries on the history of this research tradition in general, and on the big five theory in particular, see Digman (1990) and McCrae and John (1992).

speed test when monetary incentives are supplied. For women the decision of whether to invest high levels of effort cannot be explain by the either the "Big 5" constructs or the SAT scores.

While for women I found no personality trait that can predict whether they choose to invest effort in solving the test even without performance-based incentives, one's gender seems to predict this choice. Specifically, women are more likely to invest effort even without performance-based incentives. Of the 49 women participating in the experiment, only 14 (28.6%) improved their performance significantly between the \$10 and the incentives test. In contrast, out of the 50 men participating in the experiment 23 (46%) improved their performance significantly between the \$10 and the incentives test. A Chi-squared test for the equality of the distributions yields $p = 0.073$.

Therefore, for male participants the decision of whether or not to invest effort in the \$10 test relates to their conscientiousness level, i.e., their tendency to follow the rules. Moreover, women seem to be more likely to choose to invest higher levels of effort in solving the even without performance-based incentives.

6 Outcomes

In the previous sections we have seen that while some individuals are highly motivated to take the coding speed test even when no performance-based incentive are supplied (i.e., they invest high levels of effort) others are not. Moreover, the experimental results indicate that the unmotivated individuals are not the less able. If it is always the case that some individuals invest high levels of effort in trying to solve a test, regardless of the incentives provided, while others only invest high levels of effort when performance based incentives are provided, then all low-stakes test scores would measure test taking motivation. Thus, whatever individuals differences a test measures when all test takers try their best to solve the test, when no performance-based incentive are supplied the resulting test scores no longer serve as a good measure of these differences. Instead, on average, low motivated test taker will have low test scores while high motivated test takers will have high test scores. However, if this is the case, then the associations between low-stakes test scores and economic success become puzzling. In particular, it is now possible that some of these associations should be attributed to differences in test taking-motivation, or to differences in personality traits associated with it (i.e., to non-cognitive skills), and not cognitive skills.

To be able to separate test-taking motivation from what a test is supposed to measure one need to know either individuals' test-taking motivation or their ability. Of course, the latter is what the test was supposed to measure to begin with. In light of the experimental results, it seems likely that the coding speed test scores can serve as a proxy individuals' test-taking motivation. However, even in the experiment it was still the case that the variance of test scores between individuals was not reduced to zero once performance-based monetary incentives were provided. Therefore, the experimental results cannot rule out the possibility that differences in coding speed test scores may be attributed to difference in speed or in problem solving ability. Nevertheless, investigating the relation between the

coding speed test scores and outcome may provide us with suggestive evidence.

To shed light on the relations between low-stakes test scores and earnings I turn to the NLSY data set. The NLSY participants took the ASVAB test in the summer of 1980. As was mentioned above, no performance-based incentives were supplied to the participants, thus the ASVAB is a low-stakes test for these individuals.

In this section, I present evidence that the coding speed test scores have meaningful economic association with economic success. Here I will consider two outcomes, the scores on the AFQT and earnings. Tables 5a and 5b present the basic characteristics of the outcome variables to be used in this section for men and women respectively. Tables 5a and 5b present the means of the key variables, breaking them down by a coding speed dummy for men and women respectively.[57] The coding speed dummy is set to zero for all individuals whose coding speed test scores were lower than the mean, and is set to one otherwise.[58]

The story that will be told in more details below shows up in the simple means. For example those who had low coding speed test scores are more likely to be blacks or Hispanics, and they had low AFQT scores, as well. More than two decades after they took the ASVAB test, those who had low coding speed test scores have lower education and they are less likely to be employed in 2003. Moreover, conditional on being employed, those who had low coding speed test scores earn on average 35% less than those who had high coding speed test scores, and are working less. Those who work report working fewer weeks, but the same amount of total hours.

The coding speed test scores seem to be correlated with economic outcomes. However, the simple statistics presented above do not take other factors into account. In particular, one cannot, at this point, infer whether the coding speed test scores are associated with labor market outcomes once more conventional measures of cognitive ability and educational attainment are accounted for. In the following sections I investigate these questions in detail.

6.1 Relationships between the Coding Speed Test Scores and Earnings

The experimental results indicate that while some individuals invest a lot of effort in solving a test even when no performance-based incentives are supplied others do not. Suggesting that low stakes test scores measure test-taking motivation and ability. However it is not necessary that test-taking motivation has by itself any effect on earnings. If indeed the coding speed test scores measure test-taking motivation, we may have the two following models in mind. In the first one, earnings are only a function of cognitive ability and maybe some other characteristics of the individual. To proxy for

[57] As was discussed in the data section, the AFQT and the coding speed scores have been adjusted for school-year cohort, where a school year-cohort includes all the individuals that were born between October 1^{st} of one calendar year and September 30^{th} of the next calendar year. The residuals from the regressions of AFQT and the coding speed scores on school-year cohort indicators were then normalized to have a weighted mean zero and standard deviation one, using the ASVAB sampling weights. In the regressions that follow the sample is restricted to include the three youngest cohorts, i.e., all individuals born between October 1^{st} 1961 and September 30^{th} 1964.

[58] The fraction of individuals with coding-speed test scores lower than the mean is 47% for men and 45% for women.

cognitive ability I use the AFQT scores. [59] If the AFQT scores are a function of cognitive ability and test-taking motivation and if the true earnings model does not include test-taking motivation, controlling in earning regressions for the coding speed test scores should increase the estimated associations between AFQT scores and earnings. In addition, the coefficient on the coding speed test scores in the earnings regressions should be negative. The intuition is simple. By itself test-taking motivation actually masks the relation between the underlying cognitive skills that the AFQT is supposed to measure. To see that think about two individuals with the same AFQT scores that have different levels of motivation while taking the test. The one who is more motivated work harder, but only managed to correctly solve as many questions as her fellow participant who worked less hard. Thus, we should infer that the participant who worked less hard have higher cognitive ability.[60]

However, it is also possible that in the true earnings models not only cognitive ability affect earnings but also some personality traits (non-cognitive skills, like conscientiousness) are highly valued in the labor market. If test-taking motivation is related to these personality traits, then part of the relations between low-stakes test scores (the AFQT scores in our case) and earnings stem from the fact that low-stakes test scores also measure these non cognitive traits. In this case, if the coding speed test scores can serve as a proxy for these personality traits, then once they are controlled for, the association between the low-stakes test scores (i.e., the AFQT) and earnings decreases.[61]

In this section I present the earnings regressions results for men only. Several papers had cautioned against inferences made from earning regressions of women to offered wages due to severe selection problems. Neal (2004) have shown that within women while non-working white women tend to be mothers who are support by their spouse, non-working black women tend to be single mother receiving government aid. Mulligan and Rubinstein (2005) have suggested that selection is important determinant in women wages. Thus, the problem of women earnings is beyond the scope if this paper. However, the results for women are very similar to the results for men, for completeness the basic regressions results for women are presented in Table D1 in Appendix D.

Tables 6 presents the basic earning regressions for men. The dependent variable is log of earnings in 2003 of civilian workers not enrolled in school. Column 1 presents the basic log earnings regressions controlling only for age and race dummies. Column 2 adds to the regressions the AFQT scores. In accordance with the literature (see for example Neal and Johnson 1996, 1998) the AFQT scores themselves are highly correlated with earning, suggesting that they measure a trait which is highly evaluated in the market. Column 3 adds to the basic regression in column 1 the coding speed test scores instead of the AFQT scores. The coding speed test scores are also highly correlated themselves with the earnings. Thus, a one standard deviation increase in the coding speed scores corresponds to an increase of about 27.8% in yearly income. Here we have the first indication that the coding speed

[59] This is a common use of the AFQT scores (see for example, Herrnstein and Murray 1994, Heckman 1996).

[60] A similar argument can be made if the coding speed test scores measure speed (in particular if one has in mind reading speed) if it were the case that the time allotted to the AFQT is not long enough to enable of individuals to try to solve all AFQT questions.

[61] Of course, if the coding speed test scores measure fluid intelligence, and if fluid intelligence is valued in the market, we would also expect to find that the coding speed test scores are significantly associated with earning.

test scores measure some trait that is positively priced in the labor market.

So far we have seen that both the coding speed and the AFQT scores measure some traits or skills positively evaluated in the labor market. However, the regressions results in columns 2 and 3 tells us nothing about these skills. It is possible that the correlation reported in column 3 between the coding speed test scores and earnings stem from the fact that the coding speed test scores measure a skill that is positively correlated with the skills measured by the AFQT scores. In this case once we control for the AFQT and the coding speed scores the effect of the latter should disappear.

Column 4 in Table 6 addresses this possibility. It is clear from the table that the coefficients on both the coding speed scores and the AFQT scores are positive and highly significant (the F-test whether the two are jointly equal to zero yields $p < 0.0001$). After controlling for AFQT scores, one standard deviation in coding speed scores is associated with an increase 9.6% in earnings for men. Moreover, controlling for the coding speed scores decreases the association between the AFQT scores and earnings by 16.3%.

So far the specifications were very sparse, controlling only for a few variables. Of particular interest is whether the relationships between both test scores relate to educational attainments. Column 5 adds to the regression years of schooling completed by 2003. Once years of schooling completed are controlled for, the coefficient on the coding speed test scores are reduced by 30%. However, the association between the coding speed test scores and earnings are still economically and statistically significant; one standard deviation increase in coding speed test scores is associated with an increase of 6.6% in earnings. The association between the AFQT scores and earnings are reduced by even a larger amount - 66%. Nevertheless, the association between the AFQT scores and earnings are still economically and statistically significant. Moreover, even though the point estimates on the associations between AFQT scores and earnings are larger than the point estimate of the coding speed test scores and earnings, statistically they are not different than one another (F-test for the equality of the two coefficients yields $p = 0.27$). Indicating that once educational attainments are controlled for, the coding speed test scores are as important to earnings as the AFQT scores.

Therefore the coding speed test scores are significantly associated with earnings both economically and statistically. They are associated with earnings by themselves and even after controlling for conventional measures of cognitive skills like the AFQT scores and educational attainments. This suggests that the coding speed test scores measure skills which are positively priced in the labor market. Moreover, once the coding speed test scores are controlled for, the association between the AFQT scores and earnings is reduced. This implies that the AFQT scores contain a component which is also being measured by the coding speed scores. The experimental results suggest that since for the NLSY participants the AFQT scores are low-stakes test scores they would also measure test-taking motivation. Thus, test-taking motivation is the likely component to be measured by both the coding speed test scores and the AFQT scores.

6.1.1 For Whom Do the Coding Speed Test Scores Matter the Most?

The question now remains: how it is possible that the motivation to take the ASVAB relates to earnings (and educational attainments) 23 years after individuals took the test? One possibility is that test-taking motivation relates to personality traits. When taking the ASVAB test people who were trying hard, regardless of their ability, may have done so since they have been told that this is what is expected out of them (after all this is what they have been paid to do). Therefore, the traits that come to mind have to do with doing one's work even without being regularly monitored or being conscientious.[62]

In this section I investigate the relations between the coding speed test scores and individuals characteristics. I start by looking at the relationships between income and AFQT and coding speed scores for individuals of different educational attainments and different occupations in the NLSY. These regressions may be informative in assessing what skills the coding speed scores may measure.

If the AFQT scores measure cognitive ability we may expect that these test scores will be more important for highly educated individuals. If the coding speed test scores measure problem solving ability, or fluid intelligence as was suggested by Heckman (1996) and Cawley et al. (1997) then it also seem likely that they would be more important for individuals who are highly educated, and thus most likely work in jobs that require these kind of skills. The last 2 columns of Table 6 investigate this issue. In column 6 I allow for the coefficients on the AFQT scores to vary between those who at least graduate from college and those who did not. The regression results are clear; the association between AFQT scores and earnings is much stronger for those individuals who got at least bachelor degree than for the ones who did not (F-test for the equality of the two coefficients yields p = 0.026). This is not a result of controlling for the coding scores, as can clearly be seen in column 7.[63] In contrast, the coding speed test scores are related to earnings for individuals of all education levels. In parallel specification in which the coefficients on coding speed test scores were allowed to vary between those who at least graduate from college and those who did not, the two coefficients on the coding speed test scores were practically identical (F-test for the equality of the two coefficients yielded p = 0.998). Thus, for individuals of all education levels, one standard deviation in the coding speed scores is associated with 7.1% increase in earnings.

Though the coding speed test scores seem to measure a trait which is important to all workers, the relative magnitudes are quite different across education groups. For individuals who at least graduated from college, the AFQT scores are almost 4 times as important to earnings as are the coding speed scores (F-test for the equality of the coefficient yields p = 0.011). In contrast, for individuals with lower educational levels, the AFQT and the coding speed scores are as important to earnings (F-test for the equality of the coefficient yields p = 0.641). Comparing columns 6 and 7 we see that controlling for the coding speed test scores reduces the associations between the AFQT scores and earnings mainly

[62] In the experiment male participants worked hard without performance-based incentives if they were more conscientious.

[63] F-test for the equality of the two coefficients yields p = 0.035.

for the less educated workers, for whom the association between the AFQT scores and earnings is reduced by 30%. This may suggest that the skills that are being measured by the coding speed test are relatively less important for earnings of highly educated people.

Another way to investigate the problem is to look at workers in different occupations. An estimation of an occupational choice model is beyond the scope of this paper. Instead, I look at wages of individuals of different occupation. Here I look at two extreme examples, production workers, working with machines, and managers and professional. Table 7 describe the results from regressions where the dependent variable in the log of wages in 2004 for the job the workers reported.[64,65] The first two columns depict the regressions results for production workers with at most high school diploma and the last 2 columns present the regressions results for mangers and professional with at least an Associate of Arts degree. As can be clearly seen in Table 7, for production workers coding speed scores are the only test scores that influence their earnings. In contrast, for managers and professional only the AFQT scores seem to influence their wages. It seems reasonable that assume that production workers are require to do what is mostly a repetitive job, which is usually not very mentally demanding. A production worker that can be trusted to do his job even without being constantly monitored, and that is dependable (i.e., come on time and is not frequently absent) may be more valuable than one that have great mental skills.

Is it possible that for managers and professionals the personality traits measured by the coding speed test scores do not matter? As far as the coding speed test scores measure docility it may be the case that this is not the most important trait for managers and professionals, maybe just the contrary. This however does not mean that personality traits are not important for managers. In particular, in the regressions for managers and professionals the explanatory power of the both the AFQT and coding speed test scores is very low, in particular in comparison to the respective regressions for production workers. This may suggest that at least for managers and professionals we are missing some crucial explanatory variables.

6.2 The Minorities-White AFQT Gap (The Bell-Curve Revisited)

The simple means presented in Tables 5a and 5b suggest that minorities are over represented among individuals with low coding speed test scores. The experimental results suggest that the coding speed scores relate to test-taking motivation and personality traits, conscientiousness in particular. Test-taking motivation may help explaining the minorities-white AFQT gap. In this section I investigate this option in detail.

The first column in Tables 8a and 8b present the coefficients from regressions of the cohort-adjusted AFQT scores on dummies for black and Hispanic for men and women respectively. I find black-white AFQT gaps of about 1.1 standard deviations, and somewhat smaller AFQT gaps between Hispanics

[64]Since the occupation is only reported for jobs held in 2004, I use here the respective wage in 2004 for job number 1. The sample was restricted to include all civilian workers reporting positive wages in 2004 on job number 1, for whom dada on schooling in 2004 is available and were not enrolled in school in 2004.

[65]For completion, the basic regression for ln(wage2004) results are reported in Table D2 in Appendix D.

and whites. In the second columns of Tables 8a and 8b the coding speed test scores are added to the regressions. The reduction in the minorities-white AFQT gaps is substantial. The black-white AFQT gap has decrease by 39% for men and by 34% for women. The Hispanic-white AFQT gap has decreased by 29% for men and by 24% for women. Moreover, by adding the coding speed test scores to the regressions it is possible to explain 45% in the variation in AFQT scores for men, and 36% of the variation in AFQT scores for women. In comparison, column 3 in Tables 8a and 8b add to the regressions in column 1 years of schooling completed by 1980 instead of the coding speed test scores. The difference in is notable. In comparison to column 1, the minorities-whites AFQT gap is almost intact. The AFQT gap between blacks and whites is reduced by at most 7% (for men), and the gaps between Hispanics and whites is reduced by at most 15% (again for men).

The coding speed test scores can explain as much as 40% of the minorities-white test score gap. The only other variables, mentioned in the literature (see for example Phillips, Brooks-Gunn, Duncan and Crane 1998), that can explain as much of the minorities-white test scores gaps are family background variables. It is possible then that the coding speed scores are just "a one variable" summary of family background characteristics, and there lies their explanatory power. Tables 9a and 9b investigate this question in details for men and women, respectively. The first two columns of the tables repeat the regression results in the first two columns of Tables 8a and 8b for the restricted sample.[66] It is clear that there is no significant change between the two samples. In column 3 only family background characteristics are used as explanatory variables.[67] The family background characteristics are related to AFQT scores, and in the direction reported in the literature before (see Neal and Johnson, 1996). Thus, higher educated parents, working in (probably) high paying jobs, fewer siblings, and reading material at home are statistically and economically significant in predicting the AFQT scores. Family background variables explain as much of the black-white AFQT gap as the coding speed test does (compare the coefficients on the race dummies in columns 3 and 2 in the two tables). They do an even better job at explaining the Hispanic-white AFQT gap, explaining as much as 63% of the gap for men and 50% of the gap for women. However, column 4 in both tables, in which the coding speed scores are added to the regressions, indicate that coding speed still has explanatory power. Adding to the regressions the coding speed scores decreases the remaining black-white AFQT gap by another 30% and the Hispanic-white gap by another 8% for men and another 20% for women. Moreover, the coefficients on the coding speed scores are statistically and economically significant in all specifications for men and women, and the variation in the coding speed scores can explain as much variation in the

[66] For 8 men and 7 women there is no valid data in 1979 on the number of siblings or whom did they live with at age 14, these individuals were dropped from the regressions.

[67] The family background characteristics used in the Tables are almost identical to the one used by Neal and Johnson (1996) to explore the relationships between AFQT and family characteristics. In part, the use of the same variables as in Neal and Johnson (1996) is to demonstrate that the year-cohort adjustment done to the AFQT and the coding speed scores have no significant bearing on the results. There are two additional variables included here, participants' age in 1980 and an indicator equals to one if participants did not live with both his biological parents at age 14. For women, the latter is related to AFQT scores. Age is added to the regressions since the normalization used for the AFQT and the coding speed scores does not correspond one to one to participants' year of birth. This variable is always insignificant in the regressions.

AFQT scores as all the family variables combined.

It is possible that school characteristics may explain the association between the coding speed test scores and the AFQT ones. Columns 5 to 7 in Tables 9a and 9b add to the regressions school variables. The variables describing school characteristics are taken from the school survey in the NLSY. Since many schools did not respond the sample size is substantially decreased. Moreover, the racial/ethnic composition of the sample is somewhat change, the restricted sample includes 20% more black men, 25% more black women, 17% more Hispanic men, and 13% more Hispanic women, thus the results reported for the restricted sample may not represent the unrestricted one. Column 5 in the two tables presents the basic minorities-whites AFQT gaps for the restricted sample. The gaps are of comparable size to the ones for the whole sample. In column 6 family and school characteristics are added. School characteristics have the expected effect on AFQT scores. Lower student/teacher ratio, less dropouts and teacher turnover are associated with an increase the AFQT scores. Combined, school and family characteristics explain 50% of the black-white AFQT gap for men and 40% for women. For Hispanics men, these variables explain 90% of the AFQT gap for men, and render it insignificant.[68] For Hispanic women, they explain 65% of the AFQT gap. In column 7 the coding speed test scores are added to the regressions. For blacks, they help reducing the remaining AFQT gap by about 30%, and for Hispanic women they help reducing the remaining gap by 12%. However, for Hispanic men the coding speed test scores actually increase the AFQT gap with whites. The coefficients on family and school characteristics are decreasing when the coding speed scores are added, suggesting a positive relation between these variables and the coding speed scores. Moreover, the coding speed scores are still highly significant, both economically and statistically.

6.2.1 The Relations between the Coding Speed Scores and Family and School Characteristics

The regressions results reported in Tables 9a and 9b suggest that the coding speed test scores are related to family and school characteristics. In particular, the magnitudes on all the family and school characteristics are reduced once the coding speed test scores are controlled for. To see it directly, Tables 10a and 10b report the coefficients from regressions of family background characteristics on the coding speed test scores for men and women, respectively. The first two columns of the tables present the regressions of family background characteristics on the coding speed test scores. Columns 3 and 4 repeat these regressions for the restricted sample for which school information is available. In column 5 the school characteristics are added to the family explanatory variables. There are few striking differences between Tables 9a and 9b and Tables 10a and 10b. While family and school characteristics can explain about a third of the variation in the AFQT scores, the same variables explain only about a half of that amount of variation in the coding speed test scores. Moreover, the

[68] This reduction to economically and statistically insignificance of the Hispanic-white AFQT gap is mainly the result of adding the school characteristics to the regressions. However, school characteristics alone do not reduce the Hispanic-white AFQT gap to nil. Thus, it is has to do with the interaction between the two. Neal and Johnson (1996) get the same results for their NLSY sample (ibid. Table 5, p. 888)

relationships between these variables and coding speed test scores are economically and statistically less significant, in particular for women. Nevertheless, family and school characteristics are related to the coding speed test scores. In terms of the minorities-whites coding speed gaps we see similar patterns to the ones we saw for the AFQT scores. For blacks family background characteristics reduces the coding speed gap by 42% for men and 25% for women. For Hispanic men, they explain the entire coding speed gap, and for Hispanic women they explain between 50% and 55% of the gap. School characteristics reduce the remaining black-white coding speed gap by 14% for men and 9% for women. However, for Hispanics they actually reverse the gap for men, and reduce the remaining gap by 35% and making it insignificant for women.

The findings reported above indicate that the coding speed test scores can explain substantial fraction of the minorities-white AFQT gap. Even though they correlated with family and school characteristics, their relation with the AFQT scores is quite separated. Moreover, the coding speed test scores can explain substantial fraction of the variation in the AFQT scores. If the coding speed test scores measure test-taking motivation, then minorities seem to value the test less and hence the resulting AFQT scores may not provide correct ranking according to ability across racial/ethnic groups.[69]

It is possible that the results reported above suggest that minorities, and blacks in particular, are less motivated to take tests than whites. There are few explanations suggested in the sociology and psychology literature that may explain why minorities would be more likely to invest less effort than whites on low-stakes tests.[70] Higher levels of anxiety during exams (due maybe to stereotype threat[71]), or the fear that doing well on exams would be considered as "acting white", might serve as a negative non-monetary incentive that would cause blacks to invest less effort in solving questions on these tests. Economic theory provides some explanations too. In particular, a rational reaction to either actual (statistical or taste based) or perceived discrimination may induce minorities to invest little effort in exams. Thus, minorities may not believe that doing well on an exam is going to affect their future prospects in life. In the last section we saw that the coding speed test scores capture personality traits, and the tendency to follow the rules in particular. Unlike cognitive ability this tendency will be hard to detect in pre-employment testing (one would assume that conditional on wanting to get the job, job applicants would actually lie regarding these tendencies).[72] If these tendencies could be acquired

[69]Even if that it were the case that the coding speed test scores measure reading pace we would not change this conclusion. In this case the results reported above suggest that minorities read slower than whites. However, it cannot be that reading slowly can account for all the minorities-whites AFQT gaps, since otherwise the coding speed test scores would have explained all the variation in the AFQT scores. Thus, if minorities read more slowly than whites, it does not mean that they have lower cognitive ability than whites. In this case, allotting more time to the test may help lower the minorities-whites gaps substantially. However, the AFQT is supposedly composed of "power tests", i.e., tests with generous time limits. Thus, it seems very implausible that the relations reported above could stem from minorities reading much slower than whites.

[70]Some of these explanations could also account to the reason why minorities would invest less effort in high-stakes exams.

[71]For the stereotype threat, see Steele and Aronson, 1998.

[72]Of course one needs to assume here that minority job applicants actually understand that these traits are important to employers and that the lack of them may reduce their chances to get the job. In reality it is not clear that this is

through some investment behavior, then these tendencies may serve as the unobservable trait that is in the base of all statistical discrimination models.

7 Discussion and Conclusions

To investigate the importance of non-cognitive skills for earnings, past research has utilized various measures and proxies for non-cognitive skills. Rather than looking for other proxies for non-cognitive skills, the focus of this paper is on the non-cognitive component of test scores available in surveys, which have been used to date as the main measure of cognitive skills. When tests are administered to survey participants, usually no performance-based incentives are supplied. The lack of incentives allows for the possibility that personality traits, i.e., non-cognitive skills, have an affect on the resulting test scores. In particular, if some individuals respond to the lack of incentives by not trying their best, as economic theory predicts, while others do, then the resulting test scores will include a non-cognitive component, if the individuals who try their best are not the most able ones.

Using experiments and the NLSY survey data this paper investigate the relations between low-stakes test scores and economic success. In a controlled experiment, I show that the coding speed test scores are highly responsive to monetary incentives. In particular, not only did participants in the experiment improve their coding speed scores substantially in response to incentives, their relative ranking changed. While some participants chose to invest high levels of effort even without performance-based incentives, others did not. However, the participants who decided to invest less effort in solving the coding speed test without performance-based monetary incentives were not less able than their fellow participants. Moreover, I find that for male participants in the experiment, being conscientious is positively associated with investing more effort in the test even without performance-based incentives. In addition, female participants were also more likely to invest high effort in the test even without performance-based monetary incentives.

I then explore the associations between coding speed test scores and earnings for participants in the NLSY. I find that after controlling for the AFQT scores, one standard deviation increase in coding speed test scores is associated with 9.6% increase in 2003 earnings of male workers. This suggest that the coding speed test scores measure skills that are positively valued in the market. The coding speed test scores are important to earnings of all workers. Moreover, they are as important as AFQT scores to earnings of workers who had at most an Associate of Arts degree. In contrast, for worker who at least graduated from college, the association between AFQT scores and earnings are significantly larger than the associations between coding speed scores and earnings. Investigating wages of workers in different occupations, I find that while the coding speed scores are significantly and positively associated with the 2004 wages of production workers with at most a high school diploma, the AFQT

indeed the case. Autor and Scarborough (2006) suggest that there is a big minorities-white test score gaps on "Big 5" personality tests. The raw gaps reported by the authors are 0.19 and 0.12 standard deviations for blacks and Hispanics, respectively (Autor and Scarborough (2006) Table 3, p. 44). Though these gaps are much smaller than the ones on the coding speed test or the AFQT they are still pretty big given that all job applicants should know that they need to lie on these tests.

scores are not. In contrast, for managers and professionals with at least an Associate of Arts degree the situation is reversed: having higher AFQT scores is significantly associated with having higher wages in 2004, while the having high coding speed scores seem to play a minute role, if at all.

In addition, I show that the coding speed test scores themselves can explain about 30% of the variation in the AFQT scores, and up to 40% of the minorities-whites AFQT gaps. This effect is separate from the effect of family and school characteristics. Nevertheless, the coding speed test scores themselves seem to be correlated with family and school characteristics, though these variables explain a relatively small fraction of their variation.

If it is always the case that some individuals choose to invest high effort in solving a test without performance-based incentives while others do not, and this decision is not based on individuals' cognitive ability, then all low-stakes test scores will also measure participants' test-taking motivation. In an independent study Borghans et al. (2006) have shown that even when solving IQ tests some individuals invest more effort than others when no performance-based incentives were supplied. Moreover, this study too found that the decision whether to invest effort or not in solving a test relates to individuals' characteristics and preferences. Taken together, they may suggest that all low-stakes tests also measure motivation to take the test, and through it personality traits of test takers. Therefore, it is possible that some of the association found in survey data between tests and economic success should be attributed to non-cognitive skills.

While the experimental results suggest that test-taking motivation is important for the coding speed scores, the variance across individuals is not zero even when performance-based incentives are provided. Thus, there may be differences across individuals that may be attributed to speed or fluid intelligence. If this is the case, then the findings in this paper indicate that problem solving ability seems to be more important for low educated workers, and production workers in particular.

However, if the coding speed test scores measure test-taking motivation, the results presented above suggest that it is possible that at least in part the very robust associations, found in surveys, between low-stakes test scores and economic success stem from non-cognitive skills, like conscientiousness. Specifically, individuals who do well on tests do well later in life not only because they have higher cognitive skills, but also because they have personality traits that are highly valued in the market. Moreover, the existence of substantial minorities-whites coding speed gaps may suggest that there are considerable minorities-whites gaps in test-taking motivation. If this is indeed the case, then the well documented black-white and Hispanic-white test scores gap do not necessarily mean that blacks and Hispanics have lower cognitive ability than whites.

References

[1] Angrist, J.D., and Lavy, V. (2004): "The Effect of High Stakes High School Achievement Awards: Evidence from a School-Centered Randomized Trial," IZA Discussion Papers 1146, Institute for the Study of Labor (IZA).

[2] Arvey, R.D., Strickland, W., Drauden, G. and Martin, C., (1990): "Motivational Components of test Taking", Personnel Psychology, 43, pp. 695-716.

[3] Autor, D.H. and Scarborough, D., (2004): "Will Job Testing Harm Minority Workers?," NBER Working Paper No. w10763, (2004).

[4] Bandiera, O., Barankay, I., and Rasul, I. (2005):"Social Preferences and the Response to Incentives: Evidence from Personnel Data," Quarterly Journal of Economics, 120, pp. 917-62.

[5] Bandiera, O., Barankay, I., and Rasul, I. (2006a):"The Evolution of Cooperative Norms: Evidence From a Natural Field Experiment," Berkeley Electronic Journals in Economic Policy and Analysis: Advances (special issue on field experiments edited by John List), 6, pp.1-28.

[6] Bandiera, O., Barankay, I., and Rasul, I. (2006a):"Incentives for Managers and Inequality Among Workers: Evidence From a Firm Level Experiment," forthcoming, Quarterly Journal of Economics.

[7] Borghans, L., ter Weel, B., and Weinberg, B. A. (2005): "People people: Social capital and the labor-market outcomes of underrepresented groups", IZA DP No. 1494.

[8] Borghans, L., Meijers H., and ter Weel, B. (2006): "The Role of Noncognitive Skills in Explaining Cognitive Test Scores", Unpublished manuscript.

[9] Bowles, S., Gintis, H., and Osborne, M. (2001a): "The Determinants of Earnings: A Behavioral Approach," Journal of Economic Literature 39 (4) pp. 1137-1176.

[10] Cameron, J., Banko, K.M., and Pierce, W.D. (2001): "Pervasive Negative Effects of Rewards on Intrinsic Motivation: The Myth Continues", The Behavior Analyst, 24, pp. 1-44.

[11] Carneiro, P. and Heckman, J. (2003): "Human Capital Policy," in Heckman, J. and Krueger, A., Inequality in America: What Role for Human Capital Policy?, MIT Press.

[12] Cascio, E.U., and Lewis, E.G. (2006):"Schooling and the Armed Forces Qualifying Test,", Journal of Human Resources, 41(2), pp. 294-318.

[13] John Cawley, Karen Conneely, James J. Heckman and Ed Vytlacil (1997): "Cognitive Ability, Wages, and Meritocracy,' in Devlin, S. E. Fienberg, D. Resnick and K. Roeder, (eds), Intelligence Genes, and Success: Scientists Respond to the Bell Curve, 179-192, (Copernicus:Springer-Verlag, 1997).

[14] Cawley J., Heckman, J., and Vytlacil, E. (2001): "Three Observations on Wages and Measured Cognitive Ability," Labour Economics 8, pp. 419-442.

[15] Chan, D., Schmitt, N., Deshon, R.P., Clause, C.S. and Delbridge K. (1997): "Reaction to Cognitive Ability Tests: The Relationship Between Race, Test Performance, Face Validity Perception, and Test-Taking Motivation", Journal of Applied Psychology, 82, pp. 300-310.

[16] Digman, J. M. (1990): "Personality structure: Emergence of the five-factor model,"Annual Review of Psychology. Vol 41, pp. 417-440.

[17] Duckworth, A.L., Peterson, C., Matthews, M.D., and Kelly, D.R. (2006): "Grit: Perseverance and Passion for Long-Term Goals,", Mimeo University of University of Pennsylvania.

[18] Eisenberger, R. and Cameron, J. (1996): "The Detrimental Effects of Reward: Myth or Reality?", American Psychologist, 51, pp. 1153-1166.

[19] Frankel, M.R. and McWilliams, H. A. (1981): Profile of American Youth: 1980 National Administration of the ASVAB, (1982), Office of the Assistant Secretary of Defense.

[20] Gneezy, U. and Rustichini A. (2000): "Pay Enough or Don't Pay At All", QJE, pp. 791-810.

[21] Hansen, K. T., Heckman, J., and Mullen, K. J. (2004): "The effect of schooling and ability on achievement test scores," Journal of Econometrics, 121(1), pp. 39-98.

[22] Heckman, J. (1996): ""Lessons From The Bell Curve," Journal of Political Economy, 103(5), pp. 1091-1120.

[23] Heckman, J. and Rubinstein, Y. (2001): "The Importance of Noncognitive Skills: Lessons from the GED Testing Program," American Economics Review 91 (2), pp. 145-149.

[24] Heckman, J., Stixrud, J., and Urzua, S. (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior, " Journal of Labor Economics, 24(3), pp. 411-482.

[25] Herrnstein, R. J. and Murray, C. (1994), The Bell Curve: Intelligence and Class Structure in American Life, The Free Press, New York.

[26] Jacob, B. A. (2002): "Where the Boys Aren't: Non-cognitive Skills, Returns to School and the Gender Gap in Higher Education," Economics of Education Review, 21(6), pp. 589-98.

[27] Johnson, W. R. and Neal, D. A. (1996): "The Role of Premarket Factors in Black-White Wage Differences," Journal of political Economy, pp. 869-895.

[28] Johnson, W. R. and Neal, D. A. (1998): "Basic Skills and the Black-White Earnings Gap," in Jencks, C. and Phillips, M. (ed.), The Brookings Institute, Washington D.C.

[29] Judge, T. A., Higgins, C. A., Thoreson, C. J., and Barrick, M. R. (1999): "The Big Five personality traits, general mental ability, and career success across the life span," Personnel Psychology, 52, 621-652.

[30] Kremer, M., Miguel, E., and Thornton, R. (2005): "Incentives to Learn," NBER Working Paper No. 10971.

[31] Lazear, E.P. (2000): "Performance Pay and Productivity,"American Economic Review, 90(5), pp. 1346-1361.

[32] McCrae R.R., John, O.P. (1992): "An introduction to the five-factor model and its applications", Journal of Personality, 60(2), pp. 175-215.

[33] Maier, M.H., and Sims,W.H. (1983): "The Appropriateness for Military Applicants of the ASVAB Subtests and Score Scale in the New 1980 Reference Population," CNA, Memorandum 83-3102, Unclassified.

[34] Maier, M.H., and Sims,W.H. (1986): "The ASVAB Score Scales: 1980 and World War II," CNA Report 116.

[35] Maier, M.H., and Hiatt, C.M. (1986): "Evaluating the Appropriateness of the Numerical Operations and Math Knowledge Subtests in the AFQT," CRM 86-228, Unclassified.

[36] McFadden, D. (1989): "Testing for Stochastic Dominance," in in T. Fomby and T.K. Seo (eds.), "Studies in the Economics of Uncertainty," pp. 113-134, Springer: New York, 1989.

[37] McIntosh, J., Munk, M.D., and Chen Y. (2006):"What Do Test Scores Really Measure?," Unpublished manuscript.

[38] Moore, W., Pedlow, S. and Wolter, K. (1999): "Profile of American Youth 1997 (PAY97) – Technical Sampling Report," National Opinion Research Center, Chicago.

[39] Mulligan, C.B. and Rubinstein, Y. (2005): "Selection, Investment, and Women's Relative Wages Since 1975," NBER Working Papers 11159.

[40] Neal, D. (2004):"The Measured Black-White Wage Gap Among Women Is Too Small," Journal of Political Economy, 112, pp. S1-S28.

[41] Neal, D. (2005): "Why Has Black-White Skill Convergence Stopped?,"National Bureau of Economic Research, Inc, NBER Working Papers: 11090.

[42] Persico, N., Postlewaite, A., and Silverman, D. (2004): "The ffect of Adolescent Experience on Labor Market Outcomes: The Case of Height," The Journal of Political Economy, 112(5), pp. 1019-53.

[43] Ree, M.J. and Wegner, T.G. (1990):"Correcting differences in answer sheets for the 1980 Armed Services Vocational Aptitude Battery population," Military Psychology, 2(3), pp. 157-169.

[44] Revelle, W. (1993): "Individual Differences in personality and motivation: Non-cognitive' determinants of cognitive performance" in Baddeley, A. and Weiskrantz, L. (Eds.) "Attention: selection, awareness and control: A tribute to Donald Broadbent", , Oxford University Press, pp. 346-373.

[45] Roberts, B. W., Caspi, A. and Moffitt, T. E. (2003): "Work experiences and personality development in young adulthood", Journal of Personality & Social Psychology. Vol 84(3), pp. 582-593.

[46] Roberts, B. W., Robins, R. W., Caspi, A. and Trzesniewski, K. (2003): "Personality trait development in adulthood", in Mortimer, J. and Shanahan, M. (Eds.), Handbook of the life course, New York: Plenum Press.

[47] Rydval, O. and Ortmann, A. (2004): "Why Has Black-White Skill Convergence Stopped?"How Financial Incentives and Cognitive Abilities Affect Task Performance in Laboratory Settings: An Illustration ,"Economics Letters 85, pp. 315–320.

[48] Segal, C. (2005): "Misbehavior, Education, and Labor Market Outcomes," Unpublished Paper, Stanford University.

[49] Steele, C.M. and Aronson, J. (1998): "Stereotype treat and the test performance of academically successful African American", In Jencks, C. and Phillips, M. (eds.), "The Black-White Test Score Gap," The Brookings Institute, Washington DC, pp. 401-427.

Table 1: The ASVAB Subtests

Subtest	Minutes	Questions	Description
General Science	11	25	Measures knowledge of physical and biological sciences
Arithmetic Reasoning	35	30	Measures ability to solve arithmetic word problems
Word Knowledge	11	35	Measures ability to select the correct meaning of words presented in context, and identify synonyms
Paragraph Comprehension	13	15	Measures ability to obtain information from written material
Numerical Operations	3	50	Measures ability to perform arithmetic computation (speeded)
Coding Speed	7	84	Measures ability to use a key in assigning code numbers to words
Auto and Shop Information	11	25	Measures knowledge of automobiles, tools, and shop terminology and practices
Mathematics Knowledge	24	25	Measures knowledge of high school mathematics principles
Mechanical Comprehension	19	25	Measures knowledge of mechanical and physical principles, and ability to visualize how illustrated objects work
Electronics Information	9	20	Tests knowledge of electricity and electronics

The Coding Speed Subtest – Instructions and Sample Questions

The Coding Speed Test contains 84 items to see how quickly and accurately you can find a number in a table. At the top of each section is a number table or "key." The key is a group of words with a code number for each word. Each item in the test is a word taken from the key at the top of that page. From among the possible answers listed for each item, find the one that is the correct code number for that word.

Example:

Key

bargain …8385 game …6456 knife…7150
chin ...8930 house ...2859 music ...1117
sunshine ...7489 point ...4703 owner ...6227
sofa …9645

Answers

	A	B	C	D	E
1. game	6456	7150	8385	8930	9645
2. knife	1117	6456	7150	7489	8385
3. bargain	2859	6227	7489	8385	9645
4. chin	2859	4703	8385	8930	9645
5. house	1117	2859	6227	7150	7489
6. sofa	7150	7489	8385	8930	9645
7. owner	4703	6227	6456	7150	8930

Figure 1: The Coding Speed Test – Instructions and Sample Questions

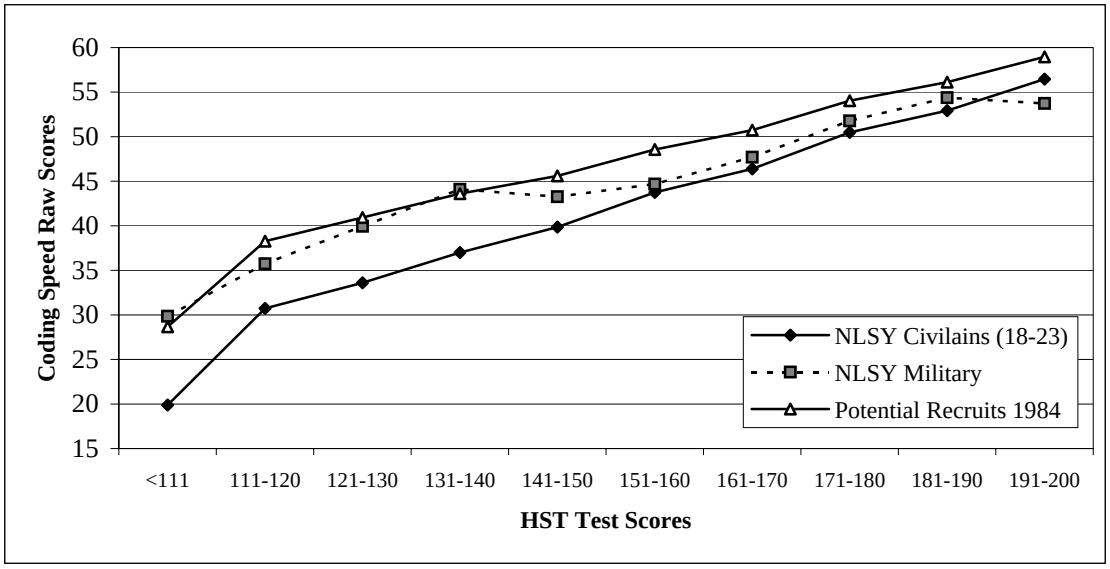

Figure 2: Raw Coding Speed Test Scores by HST Scores for NLSY participants and Potential Recruits to the Armed Forces - Men

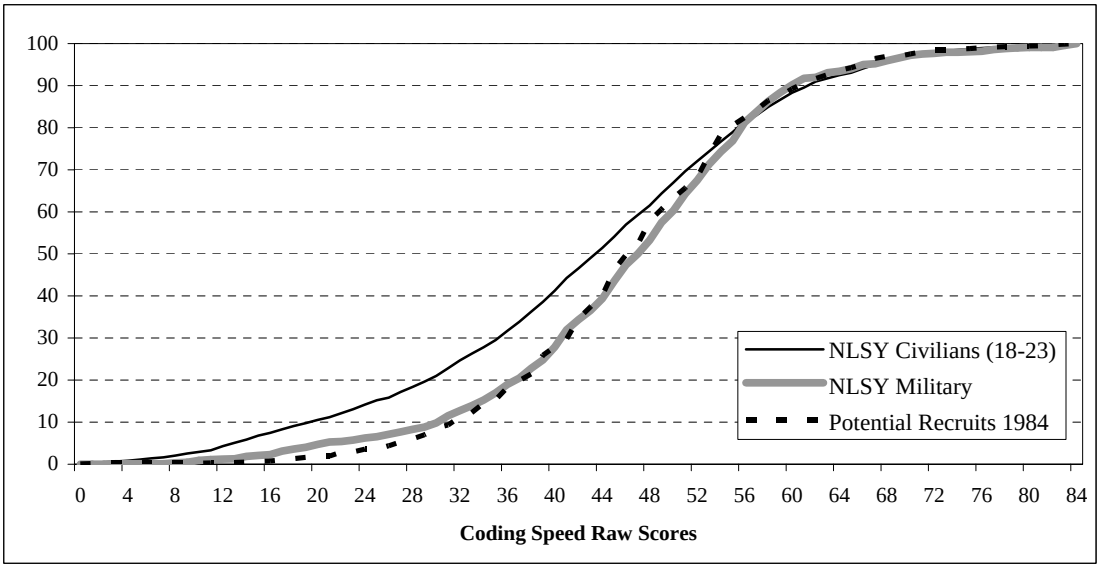

Figure 3: CDFs of Raw Coding Speed Test Scores for NLSY participants and Potential Recruits to the Armed Forces - Men

Table 2: Mean and Standard Deviation of Participants' Performance in the Experiment

	Number of Correct Answers in the Test			Number of Correct Answers in 30-Second Periods Before First Guess		
	Practice Test	$10 Test	Incentives Test	Practice Test	$10 Test	Incentives Test
Mean	90.4	104.2	112.4	4.47	5.29	5.61
Standard Deviation	18.6	23.1	17.3	1.51	1.53	1.52
Observations	99	99	99	1864	1785	1768

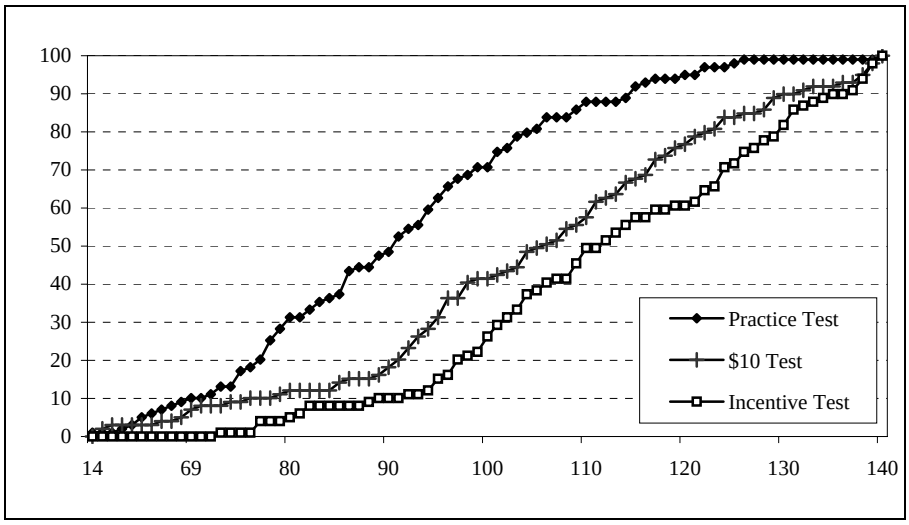

Figure 4: CDFs of Total Test Scores by Incentives Scheme

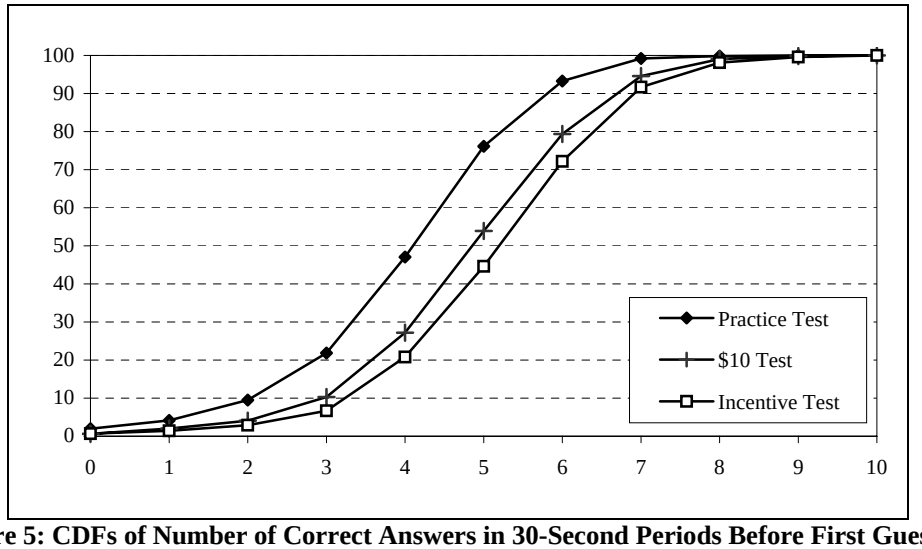

Figure 5: CDFs of Number of Correct Answers in 30-Second Periods Before First Guess by Incentives Scheme

Table 3: Relations between Number of Correct Answers in 30-Seconds Periods Before Start Guessing and Period Number

	Practice Test	Practice Test	$10 Test	$10 Test	Incentives Test	Incentives Test
Period	0.021	0.057	-0.013	-0.019	-0.030	-0.028
	(0.005)***	(0.021)***	(0.005)***	(0.020)	(0.005)***	(0.022)
$Period^2$		-0.002		0.0003		-0.0001
		(0.001)*		(0.001)		(0.001)
Observations	1863	1863	1784	1784	1767	1767
R^2	0.39	0.39	0.49	0.49	0.41	0.41

Standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Notes:

1. Individuals who start guessing within the first minute of the test were excluded. All 30-second periods after individuals start guessing were excluded too.
2. All regressions include individual fix-effects.

Table 4a: Change in Relative Ranking Between Tests Using o Total Test Scores

	% Participant that Changed Ranking by more than |4|	Mean Absolute Rank Change	Maximum Absolute Rank Change
Between Incentives Test and			
$10 Test	54.5	6.04	25
Practice Test	53.5	6.38	32
Between $10 Test and			
Practice Test	58.6	6.9	31
Observations	99	99	99

Notes: Highest rank is 1. All individuals with the highest test scores were assigned rank of 1, all the individuals with the second to highest test scores were assigned rank 2, etc.
For the practice test the test scores varies from 14 to 140, the ranks between 1 and 51. For the $10 test the test scores varies from 21 to 140, the ranks between 1 and 56. For the incentives test the test scores varies from 73 to 140, the ranks between 1 and 49.

Table 4b: Change in Relative Ranking Between Tests Using Test Scores based on Participants' Average Performance Before they Started Guessing

	% Participant that Changed Ranking by more than |4|	Mean Absolute Rank Change	Maximum Absolute Rank Change	**Obs.**
Between Incentives Test and				
$10 Test	48.5	5.45	21	97
Practice Test	62.2	6.62	28	98
Between $10 Test and				
Practice Test	62.2	6.66	28	97

Notes: Highest rank is 1. All individuals with the highest test scores were assigned rank of 1, all the individuals with the second to highest test scores were assigned rank 2, etc.
The test scores were constructed using the data from the 30-second periods before participants started guessing, see text for details. The maximum ranks are 46 for the practice test, 53 for the $10 test, and 49 for the incentives test.

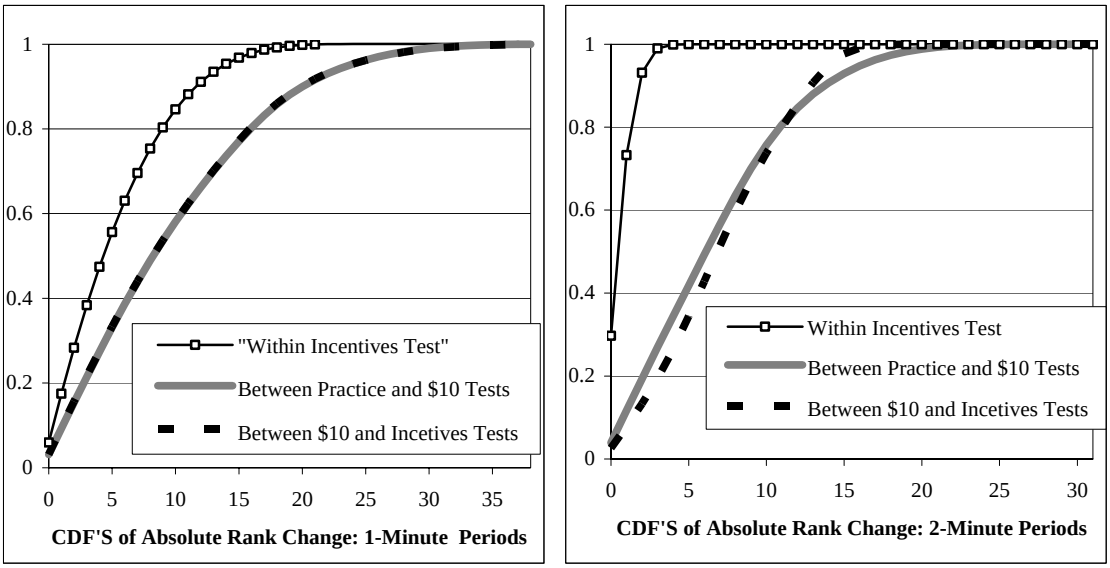

Figure 6 – CDF's of Absolute Rank Change Simulation Results (see text for details)

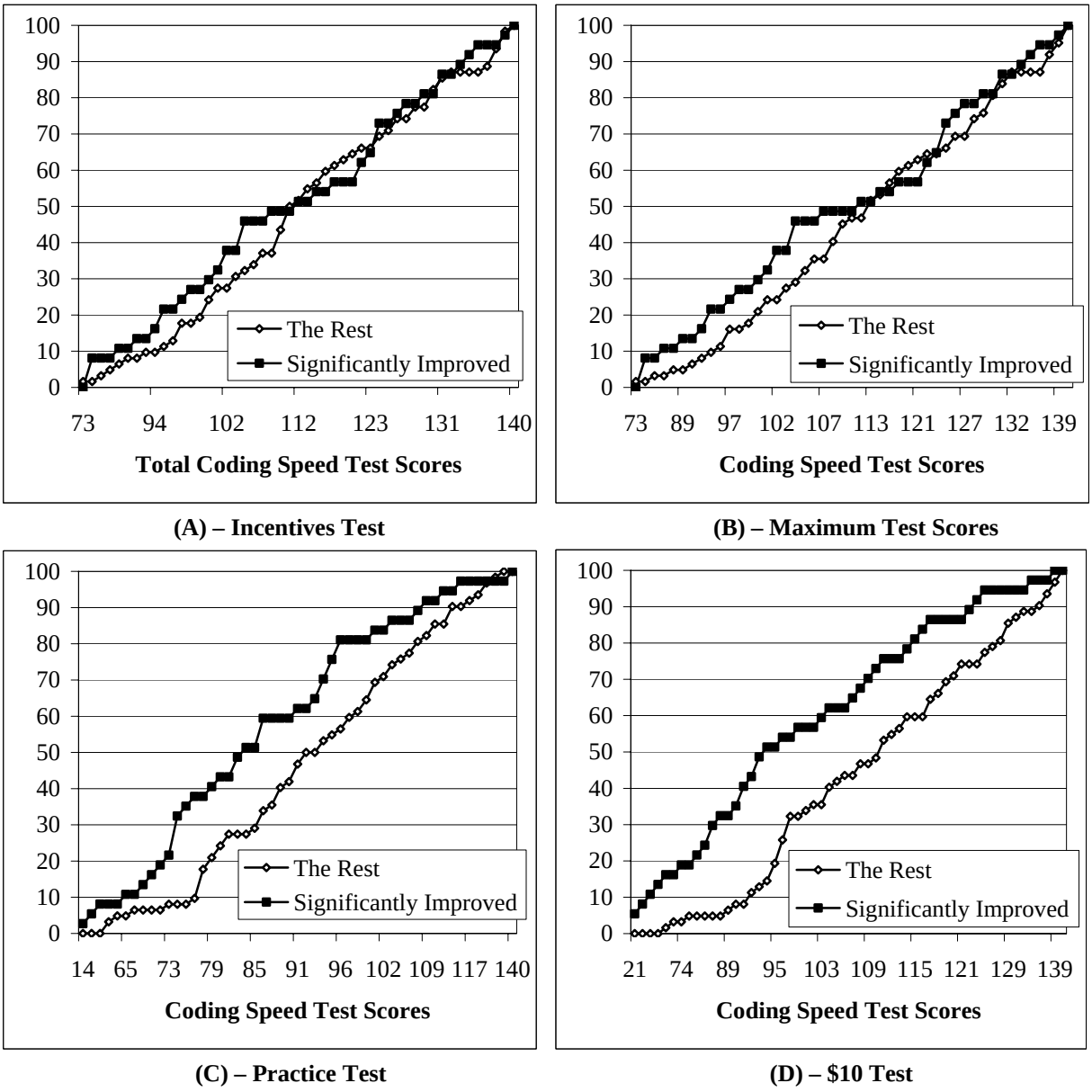

Figure 7: CDFs of Coding Speed Test Scores by Compensation Scheme and Participants Improvement in Performance between the $10 and the Incentives Tests. Panel (A) – Incentives Test, Panel (B) - Maximum Test Scores Achieved in the Experiment, Panel (C) – Practice Test, and Panel (D) –$10 Test . Significantly Improved are the Participants who Significantly Improved their Performance between the $10 and Incentives Tests. The Rest are the Complementary Group of Participants.

Table 5A: Mean and Standard Deviation of Key Outcome Variables for Men by Cohort-Adjusted Coding Speed Test Scores[1,2]

Low Coding Speed Test Scores – Individuals with Coding Speed Test Sores Below the Mean
High Coding Speed Test Scores – Individuals with Coding Speed Test Sores Above the Mean

Variable	Low Coding Speed Test Scores		High Coding Speed Test Scores			Number of observation
	Mean	Standard Deviation	Mean	Standard Deviation	Difference	
% Black	22.9		7.6			1969
% Hispanic	8.3		5.2			1969
AFQT Scores	-0.55	0.91	0.49	0.80	1.04***	1969
Years of Schooling Completed by 2004	12.3	1.97	14.0	2.41	1.7***	1484
% Working for Pay in 2003	85.8			93.7		1427
Conditional on Working in 2003						
Income 2003	$43,596	$35,069	$67,894	$56,932	$24,298***	1187
Weeks Worked 2003	48.7	13	50.4	10.2	1.7***	1187
Hours worked 2003	2253	761	2315	649	62	1187
Wage 2004	$23.1	$45	$37.1	$120.7	$14**	1187

Notes: 1. All numbers are weighted by the appropriate sampling weights.

2. The numbers are calculated for men who where born between October 1st 1961 and September 30th 1964, who have competed the ASVAB test and were not given "Spanish Instructions Cards". Individuals belonging to the poor white over-sample were excluded from the analysis.
3. Individuals who were define as working in 2003 are civilians with valid ASVAB scores, who were nor enrolled in school in 2003, and reported positive earnings.

Table 5B: Mean and Standard Deviation of Key Outcome Variables for Women by Cohort-Adjusted Coding Speed Test Scores[1,2]

Low Coding Speed Test Scores – Individuals with Coding Speed Test Sores Below the Mean
High Coding Speed Test Scores – Individuals with Coding Speed Test Sores Above the Mean

Variable	Low Coding Speed Test Scores		High Coding Speed Test Scores			Number of observation
	Mean	Standard Deviation	Mean	Standard Deviation	Difference	
% Black	24.7		6.9			1879
% Hispanic	7.9		5.1			1879
AFQT Scores	-0.48	0.93	0.39	0.88	0.87***	1879
Years of Schooling Completed by 2004	12.6	1.97	13.9	2.41	1.3***	1536
% Working for Pay 2003	79.1			81.9		1466
Conditional on Working in 2003						
Income 2003	$26,550	$18,495	$37,952	$33,474	$11,401***	1126
Weeks Worked 2003	46.6	12.1	48.4	9.1	1.8**	1126
Hours worked 2003	1824	774	1871	788	47	1126
Wage 2004	$15.7	$15.7	$23.1	$76.1	$7.3***	1126

Notes: 1. All numbers are weighted by the appropriate sampling weights.

2. The numbers are calculated for women who where born between October 1st 1961 and September 30th 1964, who have competed the ASVAB test and were not given "Spanish Instructions Cards". Individuals belonging to the poor white over-sample were excluded from the analysis.
3. Individuals who were define as working in 2003 are civilians with valid ASVAB scores, who were nor enrolled in school in 2003, and reported positive earnings.

Table 6: Earnings Men

Dependent Variable: Log of Earnings 2003

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Black	-0.555	-0.212	-0.378	-0.202	-0.286	-0.276	-0.287
	(0.066)***	(0.069)***	(0.066)***	(0.069)***	(0.069)***	(0.070)***	(0.070)***
Hispanic	-0.331	-0.120	-0.245	-0.123	-0.150	-0.151	-0.150
	(0.076)***	(0.072)*	(0.071)***	(0.071)*	(0.069)**	(0.069)**	(0.069)**
AFQT Scores		0.332		0.278	0.123		
		(0.030)***		(0.033)***	(0.040)***		
Coding Speed Scores			0.245	0.092	0.064	0.069	
			(0.026)***	(0.027)***	(0.026)**	(0.026)***	
AFQT – College Graduates or More						0.262	0.289
						(0.067)***	(0.066)***
AFQT – Less than College Degree						0.095	0.132
						(0.044)**	(0.042)***
Years of Schooling Completed 2003					0.104	0.084	0.088
					(0.014)***	(0.016)***	(0.015)***
Age in 2003	0.026	0.015	0.020	0.015	0.025	0.023	0.024
	(0.031)	(0.029)	(0.030)	(0.029)	(0.029)	(0.029)	(0.029)
Constant	9.681	10.042	9.892	10.062	8.251	8.588	8.494
	(1.257)***	(1.191)***	(1.221)***	(1.188)***	(1.223)***	(1.186)***	(1.187)***
Observations	1187	1187	1187	1187	1187	1187	1187
R-squared	0.05	0.18	0.12	0.18	0.23	0.24	0.23

Robust standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Notes:

1. The sample includes men who where born between October 1st 1961 and September 30th 1964, who have competed the ASVAB test and were not given "Spanish Instructions Cards", and did not belong to the over-sample. The sample was restricted further to include only civilian who reported positive earnings in 2003, were not enrolled in school in 2003 for whom data on schooling is available.
2. AFQT and coding speed scores are school-year cohort adjusted (see Appendix B for details).

Table 7: The Relationships between AFQT and Coding Speed Scores and Wages in 2004 for Men of Different Occupations

Dependent Variable: Log of wages 2004

	Production Workers with at Most High School Diploma		Managers and Professionals with at least Associate of Art Degree	
	(1)	(2)	(3)	(4)
Black	-0.270	-0.301	-0.042	-0.141
	(0.104)**	(0.104)***	(0.114)	(0.112)
Hispanic	-0.202	-0.210	0.028	0.039
	(0.115)*	(0.118)*	(0.104)	(0.092)
AFQT Scores	-0.043	-0.080	0.172	0.102
	(0.087)	(0.085)	(0.060)***	(0.056)*
Coding Speed Scores	0.110	0.106	0.021	-0.009
	(0.061)*	(0.056)*	(0.067)	(0.063)
Years of Schooling Completed 2004		0.113		0.092
		(0.032)***		(0.035)***
Age in 2004	0.001	0.026	-0.041	-0.041
	(0.049)	(0.053)	(0.065)	(0.065)
Constant	5.827	-44.618	87.036	86.578
	(95.561)	(102.659)	(126.829)	(126.716)
Observations	98	98	181	181
R-squared	0.14	0.20	0.04	0.08

Robust standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Notes:

1. The sample includes men who where born between October 1st 1961 and September 30th 1964, who have competed the ASVAB test and were not given "Spanish Instructions Cards", and did not belong to the over-sample. The sample was restricted further to include only civilian who reported positive wages in 2004, were not enrolled in school in 2004 for whom data on schooling is available.
2. AFQT and coding speed scores are school-year cohort adjusted (see Appendix B for details).
3. The occupation data is the occupation in job number 1 in the 2004 survey using the Census 2000 occupational categories. Production workers are workers who reported a job in the Production and Operating Workers or in the Setter, Operators, and Tenders category. Mangers are workers who reported occupation in the Executive, Administrative, and Managerial Occupation categories or in the Management Related Occupations category. Professionals are workers who reported occupation in the Mathematical and Computer Scientists category, or in the Engineers, Architects, and Surveyors category, or in the Physical Scientists category, or in the Social Scientists and Related Workers category or reported being Lawyers or Judges, Magistrates, and Other Judicial Workers.

Table 8A: Relationship Between AFQT Scores and Coding Speed Scores – Men
Dependent Variable: AFQT Scores

	(1)	(2)	(3)	(4)
Black	-1.108	-0.674	-1.036	-0.679
	(0.046)***	(0.043)***	(0.043)***	(0.041)***
Hispanic	-0.762	-0.544	-0.652	-0.504
	(0.063)***	(0.051)***	(0.060)***	(0.049)***
Coding Speed Test Scores		0.545		0.487
		(0.020)***		(0.021)***
Years of Schooling Completed 1980			0.363	0.207
			(0.025)***	(0.023)***
Age 1980	0.017	0.018	-0.327	-0.179
	(0.026)	(0.021)	(0.035)***	(0.030)***
Constant	-0.074	-0.165	2.247	1.166
	(0.448)	(0.367)	(0.456)***	(0.391)***
Observations	1969	1969	1969	1969
R-squared	0.17	0.45	0.28	0.48

Robust standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%
Notes:

1. The sample includes men who where born between October 1st 1961 and September 30th 1964, who have competed the ASVAB test and were not given "Spanish Instructions Cards", and did not belong to the poor white over-sample.
2. AFQT and coding speed scores are school-year cohort adjusted (see Appendix B for details).

Table 8B: Relationship Between AFQT Scores and Coding Speed Scores – Women
Dependent Variable: AFQT Scores

	(1)	(2)	(3)	(4)
Black	-1.133	-0.744	-1.104	-0.763
	(0.047)***	(0.048)***	(0.045)***	(0.046)***
Hispanic	-0.831	-0.637	-0.739	-0.599
	(0.058)***	(0.054)***	(0.056)***	(0.053)***
Coding Speed Test Scores		0.444		0.401
		(0.023)***		(0.024)***
Years of Schooling Completed 1980			0.280	0.174
			(0.026)***	(0.023)***
Age 1980	0.002	0.005	-0.282	-0.172
	(0.027)	(0.024)	(0.038)***	(0.033)***
Constant	0.191	0.069	2.248	1.358
	(0.472)	(0.418)	(0.498)***	(0.442)***
Observations	1879	1879	1879	1879
R-squared	0.19	0.36	0.25	0.39

Robust standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%
Notes:

1. The sample includes women who where born between September October 1st 1961 and September 30th 1964, who have competed the ASVAB test and were not given "Spanish Instructions Cards", and did not belong to the poor white over-sample.
2. AFQT and coding speed scores are school-year cohort adjusted (see Appendix B for details).

Table 9A: Relationship Between AFQT Scores and Coding Speed Scores and Family Background Characteristics- Men

Dependent Variable: AFQT Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Black	-1.109	-0.677	-0.675	-0.469	-0.989	-0.497	-0.346
	(0.046)***	(0.043)***	(0.051)***	(0.045)***	(0.066)***	(0.074)***	(0.066)***
Hispanic	-0.762	-0.543	-0.289	-0.268	-0.559	-0.068	-0.150
	(0.063)***	(0.051)***	(0.061)***	(0.051)***	(0.093)***	(0.087)	(0.074)**
Coding Speed Test Scores		0.544		0.443			0.423
		(0.020)***		(0.020)***			(0.029)***
Mother High School Grad.			0.308	0.225		0.254	0.178
			(0.057)***	(0.050)***		(0.078)***	(0.071)**
Mother College Grad.			0.415	0.325		0.300	0.212
			(0.093)***	(0.081)***		(0.121)**	(0.113)*
Father High School Grad.			0.208	0.118		0.204	0.132
			(0.058)***	(0.051)**		(0.078)***	(0.069)*
Father College Grad			0.558	0.399		0.575	0.424
			(0.082)***	(0.072)***		(0.103)***	(0.096)***
Mother Professional			0.207	0.207		0.155	0.198
			(0.083)**	(0.080)***		(0.118)	(0.116)*
Father Professional			0.133	0.092		0.034	0.059
			(0.078)*	(0.064)		(0.105)	(0.087)
Did Not Live with Both			-0.028	-0.010		-0.078	-0.053
Biological Parents at Age 14			(0.054)	(0.044)		(0.075)	(0.061)
Number of Siblings			-0.044	-0.027		-0.041	-0.029
			(0.009)***	(0.008)***		(0.013)***	(0.012)**
No Reading Materials at Age			-0.346	-0.223		-0.398	-0.242
14			(0.081)***	(0.073)***		(0.112)***	(0.097)**
Numerous Reading Materials			0.314	0.210		0.313	0.231
at Age 14			(0.049)***	(0.043)***		(0.067)***	(0.059)***
Student/Teacher Ratio						-0.019	-0.012
						(0.007)**	(0.007)*
Disadvantage Student Ratio						-0.002	-0.001
						(0.002)	(0.001)
Dropout Rate						-0.004	-0.003
						(0.001)***	(0.001)**
Teacher Turnover Rate						-0.010	-0.006
						(0.004)**	(0.003)*
Age 1980	0.019	0.018	0.030	0.027	0.022	0.026	0.017
	(0.026)	(0.021)	(0.023)	(0.020)	(0.036)	(0.033)	(0.028)
Constant	-0.110	-0.180	-0.817	-0.683	-0.203	-0.225	-0.187
	(0.448)	(0.368)	(0.410)**	(0.346)**	(0.631)	(0.601)	(0.511)
Observations	1961	1961	1961	1961	1027	1027	1027
R-squared	0.18	0.45	0.38	0.54	0.12	0.35	0.50

Robust standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Notes:

1. The sample includes individuals who where born between October 1st 1961 and September 30th 1964, who have competed the ASVAB test and were not given "Spanish Instructions Cards", and did not belong to the over-sample. The sample is restricted further to include only individuals for whom data on the variables used was not missing (in specifications 5-7 individuals with missing school data were excluded).
2. AFQT and coding speed scores are school-year cohort adjusted (see Appendix B for details).
3. Specifications 3-7 also includes dummy whether the information regarding parents' educational attainments are missing. The dummy variables for reading materials at age 14 are constructed from information about magazines, newspapers, and library cards in the home. "Numerous" means all of the above, "No" means none of the above.

Table 9B: Relationship Between AFQT Scores and Coding Speed Scores and Family Background Characteristics- Women

Dependent Variable: AFQT Scores

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Black	-1.132	-0.742	-0.738	-0.501	-1.148	-0.699	-0.457
	(0.047)***	(0.048)***	(0.052)***	(0.048)***	(0.065)***	(0.074)***	(0.071)***
Hispanic	-0.831	-0.636	-0.422	-0.343	-0.781	-0.271	-0.228
	(0.058)***	(0.054)***	(0.059)***	(0.053)***	(0.082)***	(0.082)***	(0.075)***
Coding Speed Test Scores		0.444		0.361			0.383
		(0.023)***		(0.023)***			(0.031)***
Mother High School Grad.			0.246	0.194		0.243	0.170
			(0.059)***	(0.053)***		(0.077)***	(0.071)**
Mother College Grad.			0.617	0.512		0.559	0.507
			(0.107)***	(0.096)***		(0.149)***	(0.127)***
Father High School Grad.			0.129	0.125		0.117	0.146
			(0.060)**	(0.055)**		(0.078)	(0.070)**
Father College Grad			0.470	0.436		0.423	0.378
			(0.093)***	(0.085)***		(0.130)***	(0.111)***
Mother Professional			0.079	0.084		0.137	0.073
			(0.093)	(0.084)		(0.121)	(0.111)
Father Professional			0.287	0.213		0.286	0.254
			(0.083)***	(0.077)***		(0.114)**	(0.100)**
Did Not Live with Both			-0.107	-0.095		-0.016	-0.010
Biological Parents at Age 14			(0.055)*	(0.050)*		(0.015)	(0.013)
Number of Siblings			-0.024	-0.011		-0.053	-0.057
			(0.010)**	(0.009)		(0.077)	(0.069)
No Reading Materials at Age			-0.290	-0.201		-0.300	-0.202
14			(0.073)***	(0.059)***		(0.097)***	(0.087)**
Numerous Reading Materials			0.242	0.198		0.199	0.160
At Age 14			(0.052)***	(0.047)***		(0.070)***	(0.064)**
Student/Teacher Ratio						-0.010	-0.006
						(0.004)**	(0.003)*
Disadvantage Student Ratio						-0.004	-0.003
						(0.001)***	(0.001)**
Dropout Rate						-0.002	-0.001
						(0.001)*	(0.001)
Teacher Turnover Rate						-0.007	-0.005
						(0.004)*	(0.004)
Age 1980	0.002	0.005	0.013	0.015	-0.018	-0.010	-0.007
	(0.027)	(0.024)	(0.024)	(0.021)	(0.036)	(0.031)	(0.028)
Constant	0.183	0.063	-0.504	-0.500	0.593	0.333	0.134
	(0.473)	(0.418)	(0.422)	(0.375)	(0.620)	(0.557)	(0.490)
Observations	1872	1872	1872	1872	1004	1004	1004
R-squared	0.19	0.36	0.35	0.47	0.17	0.37	0.48

Robust standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Notes: see notes to Table 9A

Table 10A: Relationship Between Coding Speed Scores and Family Background Characteristics-Men

Dependent Variable: Coding Speed Test Scores

	(1)	(2)	(3)	(4)	(5)
Black	-0.795	-0.466	-0.709	-0.411	-0.356
	(0.051)***	(0.058)***	(0.073)***	(0.083)***	(0.085)***
Hispanic	-0.402	-0.046	-0.177	0.116	0.195
	(0.064)***	(0.068)	(0.088)**	(0.090)	(0.093)**
Mother High School Grad.		0.186		0.197	0.179
		(0.065)***		(0.090)**	(0.091)**
Mother College Grad.		0.203		0.225	0.208
		(0.123)*		(0.172)	(0.166)
Father High School Grad.		0.204		0.172	0.170
		(0.062)***		(0.085)**	(0.085)**
Father College Grad		0.360		0.382	0.358
		(0.099)***		(0.134)***	(0.130)***
Mother Professional		-0.001		-0.087	-0.102
		(0.120)		(0.177)	(0.179)
Father Professional		0.092		-0.050	-0.057
		(0.098)		(0.146)	(0.145)
Did Not Live with Both Parents at Age 14		-0.041		-0.074	-0.057
		(0.059)		(0.085)	(0.084)
Number of Siblings		-0.038		-0.028	-0.030
		(0.010)***		(0.014)**	(0.014)**
No Reading Materials at Age 14		-0.279		-0.394	-0.368
		(0.102)***		(0.134)***	(0.135)***
Numerous Reading Materials at Age 14		0.236		0.213	0.194
		(0.054)***		(0.075)***	(0.075)***
Student/Teacher Ratio					-0.016
					(0.008)*
Disadvantage Student Ratio					-0.001
					(0.002)
Dropout Rate					-0.003
					(0.001)**
Teacher Turnover Rate					-0.009
					(0.005)*
Age 1980	0.001	0.008	0.021	0.029	0.021
	(0.028)	(0.027)	(0.038)	(0.037)	(0.037)
Constant	0.128	-0.304	-0.265	-0.686	-0.090
	(0.476)	(0.467)	(0.661)	(0.655)	(0.676)
Observations	1961	1961	1027	1027	1027
R-squared	0.08	0.18	0.06	0.15	0.16

Robust standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Notes:

1. The sample includes individuals who where born between October 1st 1961 and September 30th 1964, who have competed the ASVAB test and were not given "Spanish Instructions Cards", and did not belong to the over-sample. The sample is restricted further to include only individuals for whom data on the variables used was not missing (in specifications 3-7 individuals with missing school data were excluded).
2. AFQT and coding speed scores are school-year cohort adjusted (see Appendix B for details).
3. Specifications 2-5 also includes dummy whether the information regarding parents' educational attainments are missing. The dummy variables for reading materials at age 14 are constructed from information about magazines, newspapers, and library cards in the home. "Numerous" means all of the above, "No" means none of the above.

Table 10B: Relationship Between Coding Speed Scores and Family Background Characteristics-Women

Dependent Variable: Coding Speed Test Scores

	(1)	(2)	(3)	(4)	(5)
Black	-0.879	-0.654	-0.865	-0.670	-0.632
	(0.055)***	(0.062)***	(0.075)***	(0.085)***	(0.085)***
Hispanic	-0.439	-0.219	-0.396	-0.179	-0.114
	(0.060)***	(0.067)***	(0.076)***	(0.085)**	(0.088)
Mother High School Grad.		0.143		0.187	0.192
		(0.065)**		(0.081)**	(0.080)**
Mother College Grad.		0.289		0.131	0.137
		(0.123)**		(0.149)	(0.148)
Father High School Grad.		0.012		-0.059	-0.075
		(0.066)		(0.082)	(0.083)
Father College Grad		0.095		0.137	0.116
		(0.112)		(0.135)	(0.136)
Mother Professional		-0.012		0.185	0.166
		(0.098)		(0.121)	(0.120)
Father Professional		0.204		0.100	0.083
		(0.101)**		(0.120)	(0.117)
Did Not Live with Both Parents at Age 14		-0.033		0.007	0.012
		(0.063)		(0.083)	(0.083)
Number of Siblings		-0.036		-0.018	-0.017
		(0.011)***		(0.015)	(0.015)
No Reading Materials at Age 14		-0.246		-0.325	-0.257
		(0.109)**		(0.146)**	(0.137)*
Numerous Reading Materials at Age 14		0.122		0.116	0.103
		(0.056)**		(0.071)	(0.070)
Student/Teacher Ratio					-0.011
					(0.008)
Disadvantage Student Ratio					-0.001
					(0.002)
Dropout Rate					-0.003
					(0.001)*
Teacher Turnover Rate					-0.006
					(0.004)
Age 1980	-0.006	0.000	-0.010	0.001	-0.008
	(0.027)	(0.027)	(0.034)	(0.033)	(0.033)
Constant	0.269	0.052	0.375	-0.018	0.519
	(0.464)	(0.471)	(0.579)	(0.584)	(0.604)
Observations	1872	1872	1004	1004	1004
R-squared	0.10	0.15	0.09	0.14	0.16

Robust standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Notes: See notes to Table 10A

Appendix A - Theoretical Appendix

Proof Proposition 1:

When performance based incentive are provided and/or agents obtain psychic benefits from higher test scores, then the optimal level of effort, e^*, solves:

$$\frac{\partial TS(x,e^*)}{\partial e}\left(U_{TS}+U_M\phi\right)-C_e(x,e^*)=0. \tag{1}$$

The second order conditions are given by:

$$D\equiv\frac{\partial^2 TS}{\partial e^2}(U_{TS}+\phi U_M)+\left(\frac{\partial TS}{\partial e}\right)^2\left(U_{TS,TS}+2\phi U_{TS,M}+\phi^2 U_{M,M}\right)-C_{ee}. \tag{2}$$

Under the assumptions made above, a sufficient condition to ensure that $D<0$, and that the solution is indeed a maximum, is that $TS_{ee}\leq 0$.

At the optimal level of effort, e^*, the relations between the test scores and skill are given by

$$\frac{dTS}{dx}=TS_x+TS_e\frac{de^*}{dx}$$

In order to figure out how the optimal level of effort, e^*, depends on skill differentiate (2a) with respect to x to get:

$$\frac{de^*}{dx}=-\frac{1}{D}\left[TS_{ex}(U_{TS}+\phi U_M)+TS_eTS_x(U_{TS,TS}+2\phi U_{TS,M}+\phi^2U_{M,M})-C_{ex}\right]. \tag{3}$$

Using equation (3) and (4) we get

$$\frac{dTS}{dx}=\frac{1}{D}\left[TS_xTS_{ee}(U_{TS}+\phi U_M)+TS_eC_{ex}\right]>0$$

Under the assumptions made above $\frac{dTS}{dx}$ is positive. Thus, the test scores will always increase with skill, regardless of the relations between effort and skill.

To see that an increase in the incentives, i.e. an increase in ϕ, will result in an increase in the optimal level of effort, differentiate (1) with respect to ϕ, to get that $\frac{de^*}{d\phi}=-\frac{TS_e}{D}[U_M+TS(U_{TS,M}+\phi U_{M,M})$. Under the assumption that the marginal benefits from money are increasing in ϕ (i.e., $U_M+TS(U_{TS,M}+\phi U_{M,M})>0$) it is clear that $\frac{de^*}{d\phi}$ and as a result $\frac{dTS}{d\phi}=TS_e\frac{de^*}{d\phi}$ is positive, i.e. an increase in the incentives will result in an increase in effort, and a corresponding increase in test scores.

Proof Proposition 2:

Note that the only difference the two types is possibly in the distribution of skill. Since TS_1 first

order stochastically dominates TS_2 then

$$\int_{TS=\underline{TS}_1}^{x} \tilde{f}_1(TS_1)dTS \leq \int_{TS=\underline{TS}_2}^{x} \tilde{f}_2(TS_2)dTS \tag{4}$$

for all x. Thus, it is the case that $\underline{TS}_1 \geq \underline{TS}_2$ and $\overline{TS}_1 \geq \overline{TS}_2$.

From Proposition 1, we know that test scores are monophonically increasing function of skill, i.e. $\frac{dTS}{dx} > 0$, $\forall x$. Thus, $x_i = TS^{-1}(TS_i)$, where $\underline{h}_i = TS^{-1}(\underline{TS}_i)$, and $x_i = TS^{-1}(\overline{TS}_i)$, $i = 1, 2$. Hence we can rewrite (4) as,

$$\int_{x=\underline{h}_1}^{TS_1^{-1}(x)} f_1(x_1)\frac{dTS}{dx_1}dx_1 \leq \int_{x=\underline{h}_2}^{TS_2^{-1}(x)} f_2(x_2)\frac{dTS}{dx_2}dx_2 \tag{5}$$

where $f_i(x_i) = \tilde{f}_i(TS_i\,(x_i))$ and $i = 1, 2$.

By Proposition 1, the test scores provide a correct ranking according to skill for all agents, i.e. there is one to one mapping between test scores and skill, regardless of type. Thus, if $TS(x_2) = \widetilde{TS} = TS(x_1)$, Proposition 1 implies that $x_2 = x_1$. Hence, $\int_{x=\underline{h}_i}^{TS_i^{-1}(x)} f_i(x_i)\frac{dTS}{dx_i}dx_i = \int_{x=\underline{h}_i}^{TS^{-1}(x)} f_i(x)\frac{dTS}{dx}dx$, where $i = 1, 2$. Similarly, since $\underline{TS}_1 \geq \underline{TS}_2$ and $\overline{TS}_1 \geq \overline{TS}_2$ then $\underline{h}_1 \geq \underline{h}_2$ and $\overline{x}_1 \geq \overline{x}_2$. Thus we can rewrite (6) as

$$\int_{x=\underline{h}_2}^{TS^{-1}(x)} f_1(x)\frac{dTS}{dx}dx - \int_{x=\underline{h}_2}^{TS^{-1}(x)} f_2(x)\frac{dTS}{dx}dx = \int_{x=\underline{h}_2}^{TS^{-1}(x)} \left[f_1(x) - f_2(x)\right]\frac{dTS}{dx}dx \leq 0.$$

Let R be the lowest value of $\frac{dTS}{dx}$ in the range, i.e. $R \leq \frac{dTS}{dx}$ for all x. Then,

$$R\int_{x=\underline{h}_2}^{TS^{-1}(x)} \left[f_1(x) - f_2(x)\right]dx \leq \int_{x=\underline{h}_2}^{TS^{-1}(x)} \left[f_1(x) - f_2(x)\right]\frac{dTS}{dx}dx \leq 0$$

Since $\frac{dTS}{dx} > 0$ for all x, $R > 0$, $\int_{x=\underline{h}_2}^{TS^{-1}(x)} \left[f_1(x) - f_2(x)\right]dx \leq 0$. Hence, x_1 first order stochastically dominates x_2.

Proof Proposition 3:

The first order conditions are now given by

$$\frac{\partial TS(x, e^*)}{\partial e}\left(U_{TS}(\theta) + U_M\phi\right) - C_e(x, e^*) = 0 \tag{1a}$$

The second order conditions are now given by $D \equiv \frac{\partial^2 TS}{\partial e^2}(U_{TS} + \phi U_M) + \left(\frac{\partial TS}{\partial e}\right)^2(U_{TS,TS} + 2\phi U_{TS,M} + \phi^2 U_{M,M}) - C_{ee}$. Again the assumption made above ate sufficient to ensure that $D < 0$, and that the solution is indeed a maximum.

The first part of the proof is identical to proof 2. The only difference is that now U_{TS} is a function of θ (and as a result so is e^*). Thus, test scores would provide ranking of individuals that have the same θ.

In order to figure out how the optimal level of effort, e^*, depends the type, θ, differentiate (1a) with respect to θ to get $\frac{de^*}{d\theta} = -\frac{1}{D}TS_e U_{TS,\theta}$ which is positive, and hence $\frac{dTS}{d\theta} = \frac{\partial TS(x,e^*)}{\partial e}\frac{de^*}{d\theta}$ is positive too.

Proof Proposition 4:

If $x(\theta_1)$ first order stochastically dominates $x(\theta_2)$, then $\int_{x=\underline{h}_2}^{x} f(x,\theta_1)dx \leq \int_{x=\underline{h}_1}^{x} f(x,\theta_2)dx$ for all possible values of x. Proposition 3 states that $\frac{dTS}{d\theta} > 0$. Thus, the condition $\underline{TS}_1 \geq \underline{TS}_2$ does not guarantee that $\underline{h}_1 \geq \underline{h}_2$, and similarly $\overline{TS}_1 \geq \overline{TS}_2$ does not imply that $\overline{x}_1 \geq \overline{x}_2$. Moreover, if $\underline{TS}_1 = \underline{TS}_2$ then Proposition 3 implies that $\underline{h}_1 < \underline{h}_2$, and similarly if $\overline{TS}_1 = \overline{TS}_2$ then $\overline{x}_1 < \overline{x}_2$. Hence, if either $\underline{h}_1 < \underline{h}_2$, or $\overline{x}_1 < \overline{x}_2$ there will be some values of x for which $\int_{x=\underline{h}_2}^{x} f(x,\theta_1)dx > \int_{x=\underline{h}_1}^{x} f(x,\theta_2)dx$. Thus, it is not true that $x(\theta_1)$ first order stochastically dominates $x(\theta_2)$.

Therefore, without making some assumptions on the skill distributions, which is what we wanted to recover, even in the case where $TS(x,\phi,\theta_1)$ first order stochastically dominates $TS(x,\phi,\theta_2)$ we cannot guarantee that $x(\theta_1)$ first order stochastically dominates $x(\theta_2)$.[1]

Appendix B: Sample Restrictions and Variable Construction

Participants in the NLSY took the ASVAB exam in the summer of 1980 when they were 16-23 years old. The difference in ages and in particular in educational attainments affect the results of the AFQT (see for example see Hansen, Heckman and Mullen 2004, Cascio and Lewis 2006). In addition, it is possible that older individuals may be more mature, which in turn may affect their test scores. As the test scores increase with age, in order to compare between test scores of individuals of different age groups some adjustment of the test scores is needed. In the empirical specification that follows I adjust the test scores variables (both the AFQT and the coding speed scores) by school-year cohorts, where a school year-cohort includes all the individuals that were born between October 1^{st} of one calendar year and September 30^{th} of the next calendar year. School-year cohorts may represent better the effect of education on test scores while ensuring that individuals of a given cohort are on average a year older than the individuals of the preceding cohort. It is a well documented fact that the NLSY sample includes too few participants born after September 30^{th} 1964 in comparison to the general population (NLSY79 User guide pp.18-19). Thus, the sample of individuals who were born after September 30^{th} 1964 is a non-random non representative sample of the population. Thus, an additional benefit in using school-year cohort is that it excludes participants who were born after September 30^{th} 1964 from the analysis.

[1] Note also that in additions to the assumption that about the support of the ability distribution, we would need to assume that $\frac{d^2TS}{dhd\theta} \geq 0$ in order to get stochastic dominance in ability levels.

The sample is restricted further to include only individuals that have valid test scores. Thus, the base sample includes 11,625 individuals who have completed the ASVAB test.[2] To try and avoid some problems of endogeniety, in particular that either test scores or test taking motivation may be affected by experience in the labor market the sample was restricted to include only the three youngest school-year cohorts, i.e., participants born between September 1^{st} 1961 and August 31^{st} 1964.[3] In addition, since the group of poor whites was not resurveyed after 1994, it was excluded from the analysis.

The residuals from regressions of AFQT and the coding speed scores on school-year cohort indicators for the restricted sample, described above, were normalized to have weighted mean zero and standard deviation one, where the weights being used are the ASVAB sampling weights. Since women have significantly higher coding speed test scores than men, the adjustment was done separately for men and women.

When looking at earnings I use years of schooling completed. This variable was constructed using both the data on years of schooling completed as of May 1^{st} of the survey year and the data on the highest degree ever received in the following manner. For all individuals that reported that they have not received a high school diploma the actual year of schooling reported is being used. Individuals who reported receiving a high school diploma were assigned 12 years of schooling. For all individuals who reported completing at least a year of post secondary degree but not receiving any post secondary degree 13 years of completed schooling were assigned.[4] Those that reported receiving an Associate of Arts degree were assigned 14 years of schooling. Participants that reported receiving BA or BS degrees were assigned 16 years of schooling. Those who reported finishing professional school, MS or MA were assigned 18 years of schooling, and those who reported receiving a Ph.D. were assigned 20 years of schooling.

In the Section 6 of the paper I use the income data from 2003. Here, the sample was restricted to include only the civilians who reported positive earning in 2003, who were not enrolled in school for whom data on schilling is available.

To investigate the relations of AFQT and coding speed tests scores to earnings for individuals of different occupations I use the wage and the occupation reported for job number 1 in 2004. The sample was restricted to include all civilian workers reporting positive wages in 2004 on job number 1, for whom dada on schooling in 2004 is available and were not enrolled in school in 2004.

The occupation data is the occupation in job number 1 in the 2004 survey using the Census 2000 occupational categories. Production workers are workers who reported a job in the Production and

[2]The participants who got the "Spanish instruction cards" were excluded from the analysis.

[3]This sample includes some individuals who reported that they have completed 12 years of schooling by May 1^{st} 1980 (63 men out of 1963, and 97 women out of 1897), and one man who completed 13 years of schooling. The results reported in this section remains qualitatively and quantitatively the same if the sample is restricted to include only the two youngest cohorts (i.e., all individuals born between October 1^{st} 1962 and September 30^{th} 1964) or all individuals born before January 1^{st} 1961.

[4]The NLSY variable reporting years of schooling completed as of May 1^{st} of the survey year assign 16 years of completed schooling to all individuals who received BA or BS. However, individuals with 17 years of schooling may be those who continue to graduate school, or those who still did not achieve their AA, BA or BS. Thus, to maintain that those with more years of competed schooling actually have higher educational attainments this coding was chosen.

Operating Workers or in the Setter, Operators, and Tenders category. Mangers are workers who reported occupation in the Executive, Administrative, and Managerial Occupation categories or in the Management Related Occupations category. Professionals are workers who reported occupation in the Mathematical and Computer Scientists category, or in the Engineers, Architects, and Surveyors category, or in the Physical Scientists category, or in the Social Scientists and Related Workers category or reported being Lawyers or Judges, Magistrates, and Other Judicial Workers.

In all the regressions results reported below I use sampling weights. When looking at the AFQT as a dependent variable I use the provided ASVAB sampling weights and when looking at earnings in 2003 or wages in 2004 I use the 2004 cross-sectional weights.

Appendix C - Experiment

C.1 Instructions

C.1.1 Instruction for the Main Treatment

WELCOME

In the experiment today you will be asked to complete two different parts. At the end of the experiment you will receive $5 for having completed the experiment. In addition, we will randomly select one of the parts and pay you. Once you have completed the two parts we determine which part counts for payment by drawing a number between 1 and 2. The method we use to determine your earnings varies across parts. Before each part we will describe in detail how your payment is determined.

Your total earnings from the experiment are the sum of your payment for the randomly selected part, your $5 payment for completing the experiment, and a $10 show up fee. At the end of the experiment you will be asked to come to the side room where you will be paid in private.

Part 1

For the first part of the experiment you will be asked to solve one test named Coding Speed. In this test you will find a "key", which is a group of words with a code number for each word. Each item in the test is a word taken from the key at the top of that page. From the possible answers listed for each item, you need to find the one that is the correct code number for that word. Your job is to read each question carefully and decide which of the answers given is correct. Be sure to work as quickly and as accurately as you can. Your score on the test will be based on the number of answers you mark correctly. There is no guessing penalty on the test. That means if you answer a question wrong, it will not hurt you (it will just not help you). That is why it is always in your best interest to answer every question.

I will show you a demonstration of the test software and explain how to use it. To familiarize you with the test, you will be first given a practice test.

If Part 1 is the one randomly selected for payment, then you receive $10.

Please do not talk with one another for the duration of the experiment. If you have any questions, please raise your hand.

ARE THERE ANY QUESTIONS BEFORE WE BEGIN?

Part 2

As in the previous part, this part of the experiment includes one test. The test is another version of the Coding Speed test you just took. However, you now have to choose which payment scheme you want for this part. You can choose to be paid either a fixed amount of money or according to your future performance on the test in this part.

If Part 2 is the one randomly selected for payment, then your earnings for this part are determined as follows. If you choose fixed payment then you will be paid according to how well you did on test 1 in Part 1. You will be paid $10*(fraction of test questions in Part 1 correctly answered). Thus for example, if in test 1 in Part 1 you correctly answered 70 questions, i.e., you correctly answered half of test questions, your payment will be $5. If you choose to be paid according to your future performance on test 2 in Part 2, then your earnings are $30*(fraction of test 2 questions in Part 2 you correctly answer). Thus for example, if in test 2 in Part 2 you correctly answer 70 questions, i.e., you correctly answer half of test questions, your payment will be $15.

The next computer screen will tell you the fraction of test 1 questions you correctly answered, and will tell you what your fixed payment will be. It will then ask you to choose to be paid either your fixed payment or to be paid according to your future performance on test 2 in Part 2.

Please do not talk with one another for the duration of the experiment. If you have any questions, please raise your hand.

ARE THERE ANY QUESTIONS BEFORE WE BEGIN?

C.1.2 Instruction for the Control Treatment

WELCOME

In the experiment today you will be asked to complete three different parts. At the end of the experiment you will receive $5 for having completed the experiment. In addition, we will randomly select one of the parts and pay you. Once you have completed the three parts we determine which part counts for payment by drawing a number between 1 and 3. Before each part we will describe in detail how your payment is determined.

Your total earnings from the experiment are the sum of your payment for the randomly selected part, your $5 payment for completing the experiment, and a $10 show up fee. At the end of the experiment you will be asked to come to the side room where you will be paid in private.

Part 1

For the first part of the experiment you will be asked to solve one test named Coding Speed. In this test you will find a "key", which is a group of words with a code number for each word. Each item in the test is a word taken from the key at the top of that page. From the possible answers listed for each item, you need to find the one that is the correct code number for that word. Your job is to read each question carefully and decide which of the answers given is correct. Be sure to work as quickly and as accurately as you can. Your score on the test will be based on the number of answers you mark correctly. There is no guessing penalty on the test. That means if you answer a question wrong, it will not hurt you (it will just not help you). That is why it is always in your best interest to answer every question.

I will show you a demonstration of the test software and explain how to use it.

If Part 1 is the one randomly selected for payment, then you receive $10.

Please do not talk with one another for the duration of the experiment. If you have any questions, please raise your hand.

ARE THERE ANY QUESTIONS BEFORE WE BEGIN?

Part 2

As in the previous part, this part of the experiment includes one test. The test is another version of the Coding Speed test you just took.

If Part 2 is the one randomly selected for payment, then you receive $10.

Please do not talk with one another for the duration of the experiment. If you have any questions, please raise your hand.

ARE THERE ANY QUESTIONS BEFORE WE BEGIN?

Part 3

As in the previous part, this part of the experiment includes one test. The test is another version of the Coding Speed test you just took.

If Part 3 is the one randomly selected for payment, then you receive $10.

Please do not talk with one another for the duration of the experiment. If you have any questions, please raise your hand.

ARE THERE ANY QUESTIONS BEFORE WE BEGIN?

C.2 Big 5 Questionnaire

Instructions

On the following pages, there are phrases describing people's behaviors. Please use the rating scale below to describe how accurately each statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age. So that you can describe yourself in an honest manner, your responses will be kept in absolute confidence. Please read each statement carefully, and then fill in the bubble that corresponds to the number on the scale.

Response Options

1: Very Inaccurate
2: Moderately Inaccurate
3: Neither Inaccurate nor Accurate
4: Moderately Accurate
5: Very Accurate

Questionnaire Format for Administering the 50 Big-Five Factor Markers Am the life of the party.
Feel little concern for others.
Am always prepared.
Get stressed out easily.
Have a rich vocabulary.
Don't talk a lot.
Am interested in people.
Leave my belongings around.
Am relaxed most of the time.
Have difficulty understanding abstract ideas.
Feel comfortable around people.
Insult people.
Pay attention to details.
Worry about things.
Have a vivid imagination.
Keep in the background.
Sympathize with others' feelings.
Make a mess of things.
Seldom feel blue.
Am not interested in abstract ideas.
Start conversations.
Am not interested in other people's problems.
Get chores done right away.
Am easily disturbed. Have excellent ideas.
Have little to say.
Have a soft heart.
Often forget to put things back in their proper place.
Get upset easily.
Do not have a good imagination.
Talk to a lot of different people at parties.
Am not really interested in others.

Like order.
Change my mood a lot.
Am quick to understand things.
Don't like to draw attention to myself.
Take time out for others.
Shirk my duties.
Have frequent mood swings.
Use difficult words.
Don't mind being the center of attention.
Feel others' emotions.
Follow a schedule.
Get irritated easily.
Spend time reflecting on things.
Am quiet around strangers.
Make people feel at ease.
Am exacting in my work.
Often feel blue.
Am full of ideas.

Table C1: Statistics Time of First Guess Conditional on Guessing

	Practice Test	$10 Test	Incentive Test
Mean	18.5	18	18.1
Standard Deviation	2.8	4.1	3.6
Minimum	2	1	1
Maximum	20	20	20
Median	19	19	19
% Guessing in last two periods	72.3	74.2	71.4
Observations	47	62	70

Notes:

1. The numbers in the Table are given in terms of periods, each periods includes 30 seconds.
2. Only the practice test is different from the other two tests in terms of guessing.

Figure C1: Snapshot of Testing Program's Screen

Appendix D

Table D1: Earnings in 2003 - Women
Dependent Variable: $ln(Earnings)$

	(1)	(2)	(3)	(4)	(5)
Black	-0.075	0.224	0.104	0.266	0.157
	(0.074)	(0.086)***	(0.086)	(0.091)***	(0.093)*
Hispanic	0.056	0.289	0.136	0.288	0.235
	(0.082)	(0.085)***	(0.082)*	(0.084)***	(0.083)***
AFQT Scores		0.271		0.215	0.075
		(0.048)***		(0.050)	(0.040)***
Coding Speed Scores			0.210	0.121	0.114
			(0.046)***	(0.047)***	(0.046)**
Years of Schooling Completed 2003					0.104
					(0.022)***
Age in 2003	-0.014	-0.007	-0.004	-0.003	-0.008
	(0.044)	(0.043)	(0.043)	(0.043)	(0.042)
Constant	10.570	10.250	10.139	10.067	8.869
	(1.773)***	(1.729)***	(1.748)***	(1.724)***	(1.723)***
Observations	1076	1076	1076	1076	1076
$\mathbf{R}^2$	0.00	0.05	0.03	0.06	0.09

Robust standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%
Notes:

1. The sample includes women who where born between October 1^st 1961 and September 30^th 1964, who have competed the ASVAB test and were not given "Spanish Instructions Cards", and did not belong to the over-sample. The sample was restricted further to include only civilian who reported positive earnings in 2003, were not enrolled in school in 2003 for whom data on schooling is available.
2. AFQT and coding speed scores are school-year cohort adjusted (see Appendix B for details).

Table D2: Wages in 2004 - Men
Dependent Variable: $ln(Wage)$

	(1)	(2)	(3)	(4)	(5)
Black	-0.418	-0.172	-0.284	-0.163	-0.191
	(0.067)***	(0.066)***	(0.066)***	(0.066)**	(0.068)***
Hispanic	-0.234	-0.086	-0.173	-0.090	-0.097
	(0.091)**	(0.093)	(0.093)*	(0.094)	(0.094)
AFQT Scores		0.233		0.182	0.135
		(0.034)***		(0.040)***	(0.038)***
Coding Speed Scores			0.188	0.087	0.077
			(0.032)***	(0.037)**	(0.039)**
Years of Schooling Completed 2004					0.032
					(0.023)
Age in 2004	-0.004	-0.012	-0.006	-0.011	-0.008
	(0.031)	(0.030)	(0.030)	(0.030)	(0.030)
Constant	15.318	31.156	18.981	29.390	22.813
	(59.544)	(57.739)	(58.486)	(57.708)	(58.244)
Observations	1273	1273	1273	1273	1273
$\mathbf{R}^2$	0.02	0.08	0.06	0.08	0.09

Robust standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%
Notes:

1. The sample includes men who where born between October 1^st 1961 and September 30^th 1964, who have competed the ASVAB test and were not given "Spanish Instructions Cards", and did not belong to the over-sample. The sample was restricted further to include only civilian who reported positive wages in 2004, were not enrolled in school in 2004 for whom data on schooling is available.
2. AFQT and coding speed scores are school-year cohort adjusted (see Appendix B for details).