Motivation, Test Scores, and Economic Success*

Carmit Segal†
Universitat Pompeu Fabra

First Version: November 2006
This Version: October 2008

Abstract

This paper argues that low-stakes test scores, available in surveys, may be partially determined by test-taking motivation, which is associated with personality traits but not with cognitive ability. Therefore, such test score distributions may not be informative regarding cognitive ability distributions. Moreover, correlations, found in survey data, between high test scores and economic success may be partially caused by favorable personality traits. To demonstrate these points, I use the coding speed test that was administered without incentives to National Longitudinal Survey of Youth 1979 (NLSY) participants. I suggest that due to its simplicity its scores may especially depend on individuals' test-taking motivation. I show that controlling for conventional measures of cognitive skills, the coding speed scores are correlated with future earnings of male NLSY participants. Moreover, the coding speed scores of highly motivated, though less educated, population (potential enlists to the armed forces) are higher than NLSY participants' scores. I then use controlled experiments to show that when no performance-based incentives are provided, participants' characteristics, but not their cognitive skills, affect effort invested in the coding speed test. Thus, participants with the same ability (measured by their scores on an incentivized test) have significantly different scores on tests without performance-based incentives.

JEL Codes: J24, J31, C91

Key Words: Test Scores, Motivation, Cognitive Skills, Non-Cognitive Skills, Earnings

*I would like to thank Ed Lazear, Muriel Niederle, and Al Roth for their encouragement, useful suggestions, and numerous conversations; seminar participants at Harvard, Stanford, Wesleyan, Case Western Reserve, Unversitat Pompeu Fabra, University of Zurich, LSE, Tel Aviv University, FEDEA, ESA International and North American 2006 Meetings, NBER Labor Studies Fall 2006 Meeting, IZA Behavioral and Organizational Economics 2007 Workshop, and SOLE 2008 Annual Meeting; and George Baker, Greg Barron, Vinicius Carrasco, Pedro Dal Bo, Liran Einav, Florian Englmaier, Itay Fainmesser, Richard Freeman, Patricia Funk, Ed Glaeser, Avner Greif, Ben Greiner, Nagore Iriberri, Felix Kubler, Steve Leider, Aprajit Mahajan, Tatiana Melguizo, Guy Michaels, Amalia Miller, Joao de Mello, Rosemarie Nagel, Andreas Ortmann, Luigi Pistaferri, Daniel Tsiddon, Ed Vytlacil, Pierre-Olivier Weill, Toni Wegner, Catherine Weinberger, and Nese Yildiz for helpful comments. I am grateful to Harvard Business School for generous support and hospitality and thanks the support of the Barcelona Economics Program of CREA.

†Department of Economics and Business, Universitat Pompeu Fabra, Ramon Trias Fargas, 25-27, Barcelona 08005, Spain. Email: carmit.segal@upf.edu. Phone: (+34)93 542 2565.

1 Introduction

The inferences regarding test scores and their associations with economic outcomes and cognitive skills of individuals and groups are mostly based on tests administered without performance-based incentives to survey participants. Thus, there is no a-priori reason to assume that survey participants try their best to solve the test. As a result, the issue of effort, or motivation, may be crucial to the interpretation of the empirical findings. Specifically, on tests without performance-based incentives higher scores do not generally imply higher cognitive ability. Instead, higher scores may be caused by higher test-taking motivation, associated with personality traits. Therefore, it is possible that individuals that look less able are actually less motivated and that associations between higher scores and economic success should also be attributed to favorable personality traits.

To demonstrate these points, I identify a test that due to its simplicity, its scores may especially depend on individuals' test-taking motivation. This test was administered without performance-based incentives to participants in the National Longitudinal Survey of Youth 1979 (NLSY) and with incentives to potential recruits to the armed forces. The highly motivated, though less educated, population (potential enlists) scored higher on this test than the less motivated one (NLSY). Furthermore, I show that its non-incentivized scores, are positively related to future income of NLSY participants, controlling for conventional measures of cognitive skills. To gather definite evidence, I conducted controlled experiments in which this test was taken with and without performance-based incentives. I find heterogeneous responses to the lack of incentives, relating to individual characteristics but not to cognitive skills. Roughly a third of the participants, though as able as their fellow participants (as their scores on an incentivized test indicate), were less motivated and invested less effort when performance-based incentives were not provided. As a result they scored significantly worse on tests without incentives.

Economic theory indicates that if costly effort is needed to solve a test, then without performance-based incentives test-takers invest the lowest effort possible. However, survey participants' rarely score zero on unincentivized tests. This may be due to psychic benefits they gain from high scores. If high ability test-takers have lower costs of effort and/or find high test scores more rewarding, then they will have higher test scores than low ability ones. As a result, test scores will always provide a correct ranking according to ability. However, if the most able test-takers do not gain the highest psychic benefits from having high scores, then test scores, in general, will not provide correct ranking according to ability. In this case, low test scores will not imply that individuals or groups have low cognitive ability.[1] Moreover, if test-taking motivation relates to personality traits then these traits may also be a source of associations between test scores and economic outcomes. A likely candidate to affect both test-taking motivation and economic success is conscientiousness.[2]

[1] This intuition is modeled in section 6, below.

[2] Conscientiousness is a personality trait that has been repeatedly found to be positively correlated with labor market outcomes (see for example, Judge et al., 1999). I discuss it in detail below.

To investigate the relationship between test-taking motivation, cognitive ability, and test score, ideally, one would like to have both low- and high-stakes scores for each individual for a given test. With this data, the comparison between individual rankings according to their low- and high-stakes test scores can answer the question whether test-taking motivation relates to cognitive skills. If in addition data regarding economic outcomes is available, then one could investigate the importance to outcomes of personality traits, associated with test-taking motivation, relative to the importance of cognitive skills. However, to the best of my knowledge, there exists no such data. Instead, I utilize three different data sources: The NLSY, test scores of potential recruits to the armed forces, and experimental data. Each is used to provide evidence regarding a part of the puzzle.

As the ideal data is not available, selecting a proper test may be crucial. While all low-stakes test score may be affected by test-taking motivation, the effect may be more pronounced, and thus easier to detect, in tests which do not require specialized knowledge. The coding speed test may fulfill this requirement. The task in the coding speed test is to match words with four digit numbers (an example of the test is given in Figure 1). To find out which word matches to which number, test-takers need to look at the key, in which the associations between each four digit number and each word is given. As the knowledge necessary to answer the coding speed test is minimal, it is likely that effort is the main contributor to high scores. Still, the time allotted to the coding speed test is short, and thus it could be that its scores measure cognitive ability.[3]

The coding speed test is part of the Armed Service Vocational Aptitude Battery (ASVAB).[4] Participants in the NLSY were not provided with direct performance-based incentives to take the ABSVAB. Thus, for them it is a low-stakes test. The ASVAB also serves as the "entrance exam" to the armed forces. As such, it is a high-stakes test for prospective enlists. NLSY participants, though more educated, scored worse on the coding speed test than potential recruits. This is expected if test-taking motivation is important for the coding speed test.

I use the NLSY data to show that the coding speed test, though simple, when administered without performance-based incentives, measure traits highly valued in the market. I find that controlling for conventional measures of cognitive skills (the Armed Forces Qualification Test (AFQT) scores),[5] the coding speed scores are significantly associated with earnings of male NLSY participants, 23 years after they took the test. While this finding does not ensure that the skills measured by the coding speed scores are cognitive, I find that the relationship between the coding speed scores and earnings follows patterns documented for non-cognitive skills (Segal, 2005, Heckman et al., 2006). Specifically, I find that the coding speed scores are relatively more important to

[3]Using factor analysis Heckman (1995) and Cawley et al. (1997) have shown that the coding speed test and the numerical operations test, which includes very simple arithmetic computations, correspond to a different factor than the other ASVAB tests and that together they are highly correlated with earnings. The authors suggest that these tests measure "fluid intelligence or problem solving ability" (Heckman, 1995, p. 1105).

[4]The 10 ASVAB tests are described in Table A1 in Appendix A.

[5]The AFQT has been widely used as a measure cognitive skills and has been found to be correlated with NLSY participants' income (see for example, Herrnstein and Murray, 1994, Heckman, 1995, Neal and Johnson, 1996).

earnings of low educated workers.

The evidence from the NLSY and the comparison to potential recruits suggest that it is possible that the coding speed scores relate to test-taking motivation and to personality traits associated with it. To gather conclusive evidence, I conducted a controlled experiments, in which motivation was induced via the provision of incentives. Subjects in the experiment took the coding speed test three times. Twice for a fixed payment, where the first version was called "practice" test and the second "The" test. Monetary performance-based incentives were provided for the third version.

The model implies that the provision of incentives may change subjects' ranking if test-taking motivation differs across subjects. I find that subjects changed their ranks between the tests. This rank change is due to subjects' heterogeneous responses to the lack of incentives. Specifically, participants can be divided into two groups. While the first group (62% of subjects) consists of subjects whose own performance did not improve with provision of performance-based incentives, the performance of participants of the second group improved significantly. When no performance-based incentives were provided the test score distribution of the first group first order stochastically dominated the test score distribution of the second one. Thus, subjects of the second group appear less able. However, subjects of both groups had the same test score distributions when incentives were provided. Taken together these results suggest that those who performed worse on tests without incentives invested less effort and were unmotivated, though not less able.

Utilizing participants' answers to a psychological survey, I find that male participants who invested high effort only incentives were provided were less conscientious. In addition, women were more likely to invest high effort even without incentives. Consistent with the experimental results, I find no relationship between subjects' effort choices in the experiment and their SAT scores.

Taken together, the evidence in this paper suggests that due to the simplicity of the coding speed test, its scores are highly correlated with test-taking motivation when no performance-based incentives are provided. This is the first paper demonstrating that, at least for the coding speed test, higher test scores on tests without incentives do not imply higher cognitive ability. Instead, when incentives are not provided, individual characteristics affect effort invested in solving the test.

The relationship between motivation and test scores has been investigated before. While the working assumption of the psychometric literature seems to be that all test-takers are highly motivated, there is substantial evidence, dating back to the 1900's, that motivation affects performance on tests and may be related to personality traits (for an excellent summary see Revelle (1993) and citations therein).[6] In economics, the evidence obtained through lab and field experiments clearly indicates that performance on tests is positively related to (high enough) incentives (see for example, Gneezy and Rustichini, 2000, Angrist and Lavy, 2004, Kremer et al., 2005).

This paper also relates to the recent literature investigating the validity of the basic premises of

[6]The focus of psychologists (and lately of economists) has been on the crowding out effects that extrinsic incentives may have on intrinsic motivation (see for example, Camerer and Hogarth, 1999, Benabou and Tirole, 2003).

agency theory. Namely, that individuals invest little effort unless provided with proper incentives or monitored.[7] This literature suggests that economic theory can predict the behavior of non-negligible fraction of individuals. For example, Nagin et al. (2005) show that about 40% of employees in a calling center shirked when they inferred that they are not being monitored. Fehr and Falk (1999) show experimentally that in response to higher flat wages about 25% of participants always provided minimal effort, while the rest responded by choosing higher effort. This paper demonstrates that these insights are present in a testing situations too.[8] Furthermore, it shows that heterogeneous responses to the lack of performance-based incentives are not driven by differences in abilities.

Lastly this paper relates to the literature relating cognitive and non-cognitive skills to earnings (see for example, Bowles et al., 2001, Persico et al., 2004, Kuhn and Weinberger, 2005, Segal, 2005, Heckman et al., 2006). Rather than looking for a proxy for non-cognitive skills, I focus on the non-cognitive component of test scores available in surveys, which are the main measure of cognitive skills. I argue that the lack of performance-based incentives allows personality traits, i.e., non-cognitive skills, to affect test scores.[9] While the regression results using the NLSY data only provide suggestive evidence on the relationship between test-taking motivation and personality traits, the experimental results provide a direct one.

Next I briefly describe the NLSY data and discuss in detail the tests used in the analysis. I proceed by investigating the relationship between the coding speed scores and earnings in NLSY data and then provide the comparison to potential recruits. To highlight how test-taking motivation can be detected, I introduce the model. Lastly, I describe the experiment, its results and conclude.

2 Data

The analysis in Section 4 relies on the National Longitudinal Survey of Youth 1979 (NLSY). A nationally representative sample of over 12,000 individuals that were first surveyed in 1979, when they were between the ages of 15 and 22, and then re-surveyed annually until 1994 and biannually afterwards. For the purposes of this paper, this source is exceptional in combining detailed labor market data with a battery of tests, which is also administered to a non-survey population. Since the NLSY is a well-known survey, this section will focus on aspects particular to this paper, namely, the tests administered to NLSY participants. Due to its main role in the analysis the coding speed test is described in the next section. Details regarding the sample restriction and variable construction

[7] While not directly related to this paper, there is a growing literature in economics investigating how other-regarding preferences alters individual behavior (for an excellent summary see Fehr and Schmidt, 2003).

[8] This, is documented in the literature in psychology (Revelle, 1993). In economics, this effect can first be found in Gneezy and Rustichini (2000) where the effect of incentives was to move some participant scores away (when incentives were high) and toward (when incentives were low) zero scores on the test. Lately, Borghans et al. (2008) show that when given IQ questions some participants respond to incentives mainly by investing more time in answering questions while others do not. The authors relate this response to incentives to personality traits.

[9] Recently several studies have shown that test scores correlates with personality traits and preferences parameters (Benjamin et al., 2005, Borghans et al., 2008, Dohmen et al., 2008).

can be found in Section A1 of Appendix A. The military data is described in Section 5 and in Section A2 of Appendix A. The experimental data is described in Section 7.

2.1 The Tests Used in the Analysis

The ASVAB - The ASVAB is a battery of 10 tests, described in Table A1 in Appendix A. It serves as the screening and sorting exam to the armed forces. As the U.S. Department of Defense (DOD) had to establish a national norm for the ASVAB, it had to be administered to a represented sample of Americans. The DOD and the U.S. Department of Labor decided to utilize the NLSY sample for this purpose. The administration of the ASVAB to the NLSY participants took place between June and October of 1980. Participants in the NLSY were paid $50 for completing the test.[10] However, no direct performance-based incentives were provided.[11] Thus, for the NLSY participants the ASVAB is a low-stakes test.

AFQT - The Armed Forces Qualification Test (AFQT) scores are created by adding the scores of four of the ASVAB subtests (word knowledge, paragraph comprehension, arithmetic reasoning, and mathematics knowledge). The AFQT is the most commonly used test in studies using the NLSY data set (see for example, Herrnstein and Murray, 1994, Heckman, 1995, Neal and Johnson, 1996).

3 The Coding Speed Test

The coding speed test is central to the analysis. Thus, I start by describing the test and the reasons why it has been chosen. The instructions and an example of the questions asked in the coding speed test are given in Figure 1. The coding speed test is one of the ASVAB subtests.[12] The task in the coding speed test is to match words with four digit numbers. To figure out which word matches to which number, test-takers need to look at the key, in which the association between each word and a four digit number is given. 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 from five possible codes.[13] The NLSY participants took a paper and pencil version of the test that lasts for 7 minutes and consists of 84 questions.

Ideally, in order to test whether test takers differ in their motivation to take a test, we would like to find a test, such that all test takers have the knowledge necessary to correctly answer all questions, if they so desire. The coding speed test seems a likely candidate to fulfill this requirement. It seems

[10] "...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..." (Department of Defense (1982), p. 12).

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

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

[13] Note that even though the name coding speed may suggest complicated reasoning task, unlike IQ tests, test takers do not need to infer the relationships between the words and the numbers, as they are given in the key.

likely that everyone that knows how to read has the knowledge to correctly answer questions on the test. Therefore, due to its simplicity, test-taking motivation may play a large role in determining its scores.[14] Nevertheless, as the time allotted to the coding speed test is short, it is possible that not all test-takers are able to achieve a perfect score. Thus, the coding speed test may also measure cognitive ability related to speed. This ability may be different than the one that is being measured by the AFQT. For example Heckman (1995) suggests that the coding speed (and numerical operation) tests measure fluid intelligence or problem solving ability.

4 The NLSY Data: The Coding Speed Scores and Earnings

In this section, I present evidence that the coding speed scores are correlated with earnings of the NLSY participants. The results presented in this section are for men only, as a full treatment of the selection problem associated with female earnings is beyond the scope of this paper.[15] The results for women are very similar to ones for men. For completeness the basic means and regressions results for women are presented in Tables B1 and B2, respectively, in Appendix B.

The coding speed test seems to be a very simple test. Nevertheless, its scores are highly correlated with future economic success of NLSY participants. Table 1 presents the means of the key variables, breaking them down by a coding speed dummy for men.[16] The coding speed dummy is set to zero for all men whose coding speed scores were lower than the mean (47% overall), and is set to one otherwise. The story that will be told in detail below shows up in the simple means. More than two decades after NLSY participants took the ASVAB test, men who had low coding speed scores had lower educational attainment and are less likely to be employed in 2003. Moreover, conditional on being employed, those who had low coding speed scores earn on average 35% less than those who had high scores. While the coding speed scores seem to be correlated

[14]The ASVAB contains another test that may seem appropriate to use: the Numerical Operation test. This test consists of 50 simple algebraic questions (e.g., 2+2=?, 16/8=?, etc.) and lasts 3 minutes. However, it is possible, that some individuals may not have the knowledge necessary to correctly solve these questions. A more serious concern is that individuals with high math skills may invest higher effort in solving the numerical operations test than individuals with low math skills. Thus, the scores may include a larger component of knowledge than the content of the questions may suggest. Psychologists investigating motivation (or the lack of it) suggest that its effects are more pronounced the longer the task lasts (see Revelle, 1993). The coding speed test is more than twice as long as the numerical operation test, suggesting that the effects of test-taking motivation may be more pronounced for it. In addition, while 16% of NLSY participants correctly solved at least 90% of the numerical operation questions, the corresponding number for the coding speed test is 1%. Thus, the coding speed test may serve as a better measure since its scores range is less restricted.

[15]Several papers had cautioned against inferences made from female earnings regressions to offered wages due to severe selection problems. For example, Neal (2004) have shown that while non-working white women tend to be mothers supported by their spouse, non-working black women tend to be single mother receiving government aid. Mulligan and Rubinstein (2005) suggest that selection is an important determinant in female wages.

[16]As is discussed in Appendix A, 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 year and September 30^{th} of the following one. 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.

with economic outcomes, the simple means presented above do not take other factors into account. Below, I investigate whether the coding speed scores are associated with labor market outcomes once conventional measures of cognitive ability and educational attainment are accounted for.

4.1 The Basic Regression Results

The model estimated in this section is of the form $ln(earnings)_i = \beta + \beta_{AFQT} AFQT_i + \beta_{CS} CS_i + \beta_X X_i + \varepsilon_i$, where i indexes individuals, $earnings$ are earnings in 2003, $AFQT$ are the AFQT scores, CS are the coding speed scores, X denote individual characteristics, and ε is an error term.

Before turning to the regression results it is useful to discuss how to interpret different possible results. There are two distinct cases to consider. 1. The coding speed scores proxy for traits valued in the market and different from the one measured by the AFQT. These traits may be cognitive skills (possibly fluid intelligence) or the personality traits associated with test-taking motivation. In either case, controlling for the AFQT scores, β_{CS} should be positive.

2. Earnings are only a function of cognitive skills (presumably measured by the AFQT). However, as the AFQT was administered without performance-based incentives, for the NLSY participants it is a low-stakes test. Thus, its scores should be a function of cognitive skills and test-taking motivation. Hence, adding the coding speed scores (that presumably measure test-taking motivation) to regressions that include the AFQT ones, should increase β_{AFQT}. Moreover, β_{CS} should be negative. The intuition is simple. By itself test-taking motivation masks the relationship between the underlying cognitive skills and the AFQT that supposed to measure them. To see that consider two individuals having the same AFQT scores but different levels of test-taking motivation. The one who is more motivated works harder, but nevertheless only manages to solve as many questions as his fellow participant who works less hard. Thus, the unmotivated participant has higher cognitive skills. However, this can be inferred only if test-taking motivation is known.[17]

Table 2 presents the basic regression results. The dependent variable is log of earnings in 2003 of male civilian workers not enrolled in school. In column 1 only age and race dummies serve as controls. 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 are highly correlated with earnings, suggesting that they measure a trait which is highly valued in the market. Column 3 adds to the regression in column 1 the coding speed scores instead of the AFQT ones. The coding speed scores are highly correlated themselves with earnings. One standard deviation increase in the coding speed scores corresponds to an increase of 27.8% in earnings. Thus, the coding speed scores measure a trait that is positively priced in the labor market. In Column 4 both the AFQT and the coding speed scores are added to the regressions. The coefficients on both the coding speed and the AFQT scores are positive and highly significant (the F-test for whether the two are jointly equal to zero

[17] A similar argument can be made if the coding speed scores measure speed (in particular reading speed) if the time allotted to the AFQT is not long enough to allow all individuals to try solving all AFQT questions.

yields $p < 0.01$). Controlling for the AFQT scores, one standard deviation increase in the coding speed scores is associated with an increase 9.6% in earnings.

In comparison to columns 2 and 3, the point estimates on both the AFQT and the coding speed scores in Column 4 are reduced in magnitude, indicating that they share a common component. This common component may be test-taking motivation as both were administered without performance-based incentives. Alternatively, it may be reading ability. The AFQT includes two verbal parts one of which (reading comprehension) was introduced to "...help solve the problem of assessing literacy" (Maier and Sims, 1986, p. A-9). Thus, it is possible that the coding speed scores correlates with earnings only because the regressions reported in Table 2 regressions the math and verbal parts of the AFQT are not allowed to vary independently. To investigate this question, Table 3 repeats the regressions in columns 2 to 5 of Table 2 using the 4 tests separately. Table 3 clearly indicates that none of the main findings is different; the coding speed scores are still significantly correlated with earnings and so is the sum of the 4 tests.[18] Thus, the results in Table 3 suggest that the relationship between the coding speed scores and earnings do not stem from reading ability alone.

Of particular interest is whether the relationships between both test scores and earnings relate to educational attainment. Once years of schooling completed are controlled for, in Column 5, the coefficient on the coding speed scores are (insignificantly) reduced by 30%. However, the association between the coding speed scores and earnings is still economically large and statistically significant. One standard deviation increase in coding speed scores is associated with an increase of 6.6% in earnings. The association between the AFQT scores and earnings is significantly reduced by a larger amount - 66%. Nevertheless, the association between the AFQT scores and earnings are still significant. Interestingly, the coefficients on the AFQT and the coding scores are no longer significantly different than one another (F-test for the equality of the coefficients yields $p = 0.27$).

4.2 For Whom Do the Coding Speed Scores Matter the Most?

The results in Table 2 suggest that the coding speed scores are significantly associated with earnings. They are associated with earnings by themselves and after controlling for conventional measures of cognitive skills like the AFQT scores and educational attainment. This suggests that the coding speed scores measure skills which are positively priced in the labor market. In light of the discussion in the beginning of the section, we can conclude that the coding speed scores either proxy for cognitive skills (different than the ones measured by the AFQT) or for personality traits that relate to test-taking motivation. Below, I try to shed light on what skills the coding speed scores may

[18]The four tests are highly correlated with one another, thus caution should be taken when trying to draw any conclusions regarding the patterns of their correlations with earnings. Nevertheless, a few comments may be warranted. The comparison between columns 2 and 3 suggests that the coefficients on the verbal parts of the AFQT are reduced by larger fraction once the coding speed scores are added to the regressions. However, these reductions are insignificant both individually and jointly. Actually the correlations between the coding speed scores and earnings are reduced by the largest amount when the math and not the verbal parts of the AFQT are added to the regressions.

measure. The evidence presented below is suggestive. The interpretation relies on what is already known about the relationship between cognitive and non-cognitive skills and earnings.

If the AFQT measure cognitive ability we may expect that its scores will be more important for earnings of highly educated individuals. Similarly, if the coding speed scores measure problem solving ability or fluid intelligence as was suggested by Heckman (1995), then it also seems likely that they would be more important for earnings of individuals who are highly educated. The last 3 columns of Table 2 investigate this issue. In column 6, I allow for the coefficients on the AFQT scores to vary between those who at least graduated 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 a 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 8.[19] In contrast, the coding speed scores are related to earnings for individuals of all education levels. In Column 7, I allow the coefficients on coding speed scores to vary between workers who at least graduated from college and those who did not. The two coefficients on the coding speed scores are identical (F-test for the equality of the coefficients yielded p = 0.998), though the one for highly educated workers is imprecisely estimated.

The relative importance of the traits measured by the coding speed scores vary 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 = 0.038 for Column 7 specification). In contrast, for individuals with lower education levels, the AFQT and the coding speed scores are as important to earnings (F-test for the equality of the coefficient yields p = 0.669 for Column 7 specification). This suggests that the skills that are being measured by the coding speed test are relatively less important for earnings of highly educated workers. The evidence in the literature suggests that non-cognitive skills are more important to low educated people (see Segal, 2005, Heckman et al., 2006).

To further shed light on what the coding speed may measure, I examine 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 occupations. Here I look at two extreme examples, production workers, working with machines, and managers and professional. Table 4 describe the results from regressions where the dependent variable is the log of wages in 2004 for the job the workers reported.[20,21] 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

[19] F-test for the equality of the two coefficients on the AFQT scores in Column 8 yields p = 0.035.

[20] Since 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 not enrolled in school reporting positive wages in 2004 on job number 1, for whom dada on schooling in 2004 is available. See section A1 in Appendix A for details.

[21] For completion, the basic regression results when ln(wage2004) is the dependent variable are reported in Table B3 in Appendix B.

and professional with at least an Associate of Arts degree. As can be clearly seen in Table 4, for production workers coding speed scores are the only test scores that relate to wages. In contrast, for managers and professional only the AFQT scores relate to wages. It seems reasonable to assume that production workers are required 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 (e.g., comes on time and is not frequently absent) may be more valuable than one that has great mental skills. As far as the coding speed 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 the regressions for managers and professionals the explanatory power of both the AFQT and coding speed 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 some crucial explanatory variables are missing.

4.3 The Coding Speed Scores and Family and School Characteristics

In this section I discuss the relationships between the coding speed scores and family and school characteristics. The purpose is two folded: To investigate the relationship between the coding speed scores and family and schools characteristics. But also, to find out what can be learned from the comparison between these relationship for the AFQT and coding speed scores.

Table 5 presents the coefficients from regressions of the cohort-adjusted AFQT (columns 1-3) and coding speed scores (columns 4-6) on family and school characteristics.[22] In columns 1 and 3 only family characteristics are used as explanatory variables. Family characteristics are related to both the AFQT and the coding speed scores. Thus, higher educated parents, working in (probably) high paying jobs, fewer siblings, and reading material at home are statistically significant and economically meaningful predictors of the AFQT and the coding speed scores. The difference is in their explanatory power. While family characteristics explains almost 40% of the variation in the AFQT scores, they only explain about 20% of the variation in the coding speed scores.

In the remaining columns of Table 5 schools characteristics are added to the regressions. 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 changed, the restricted sample includes 20% more black men and 17% more Hispanic men. Thus the results reported for the restricted sample may not represent

[22]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. 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.

the unrestricted one. Therefore, columns 2 and 5 in the table repeat the regressions reported in columns 1 and 3 for the restricted sample. While the coefficients are somewhat different the qualitative relationship between both test scores and family background characteristics remains the same. Columns 3 and 6 clearly display that lower student/teacher ratio, less dropouts and teacher turnover are positively associated with an increase in both the AFQT and the coding speed scores. Again, we see that these variables explain more than twice the variations in the AFQT scores than in the coding speed ones. These results are consistent with the findings in Segal (forthcoming), where family and school characteristics seem to explain substantial part of the variation in achievement test scores, but not in non-cognitive skills.

5 Indirect Evidence from the Armed Forces

The results in the last section imply that the coding speed scores measure traits positively priced in the market and different than the ones measured by the AFQT. Moreover, the results are consistent with the interpretation that the coding speed scores measure non-cognitive skills associated with test-taking motivation. Next I provide indirect evidence that the on the test day, the coding speed scores relate to test-taking motivation. If the lack of performance-based incentives results in lower test scores, then higher test scores are expected when the same test is administered to highly motivated population, everything else equal. Moreover, if test-taking motivation is particularly important for the coding speed scores then the effect should be more pronounced for this test. As the ASVAB is the screening and sorting exam for potential enlists to the armed forces, this hypothesis can be tested, by comparing the scores of the NLSY participants and potential recruits, who should have the incentives to do well on the ASVAB. However, as potential recruits are less educated and more racially diverse population than the NLSY participants, this is not a perfect test. Nevertheless, this comparison may serve as an indication whether the effect exists.

When establishing the national norm for the ASVAB, Maier and Sims (1983) first discovered the problems in comparing the ASVAB scores between potential recruits to the armed forces and NLSY participants of comparable ages (i.e., born before 1/1/1963). Specifically, Maier and Sims (1983) show that while potential recruits score higher on the speeded tests (i.e., coding speed and numerical operations tests) than the NLSY participants, they did worse on any other test.[23] The latter part was expected since the NLSY participants were more educated than potential recruits (Maier and Hiatt, 1986).[24] The former part was not. Maier and Hiatt (1986) suggest that the gaps on the speeded tests are the result 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

[23] The possible solutions suggested by military researchers are discussed in detail in Section A2 of Appendix A.

[24] It is not clear, though, if the NLSY participants scored as high as can be expected given their education level.

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 the NLSY participants did not know these strategies. However, they may not cared enough to follow them.

Unfortunately, none of the above mentioned sources provide the raw test score distributions of potential recruits. However, using the information provided in Maier and Hiatt (1986, Appendix A, pp. A1-A10), I was able to reconstruct the coding speed score distribution for the 1984 male applicants for enlistment (IOT&E 1984).[25] Figure 2 presents the cumulative coding speed scores for 3 groups of males: NLSY civilian sample born before 1/1/1963, NLSY military sample, and the IOT&E 1984 sample. Figure 2 clearly displays that the NLSY civilian population has the lowest test scores, in particular for the lower 80% of the test score distribution.[26] If indeed the coding speed 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 are.

Thus, the comparison between potential recruits to the armed forces and the NLSY participants provides indirect evidence that the coding speed scores may be highly related to motivation to take the ASVAB. However, there may be other possible explanations to account for the differences between the two populations. Therefore, in order to gather direct evidence that indeed motivation plays an important role in determining the coding speed scores I turn to the controlled experiment.

6 The Model

In this section I model how individual differences in test-taking motivation may affect their test scores. I start with the case in which individuals differ only in their skills. I then extend the model to include individual differences in test-taking motivation. The main purpose is to understand under what conditions individuals' ranking according to their test scores corresponds to their ranking according to their skills. This will allow to derive testable predictions that in turn will allow for detection of test-taking motivation in the experiment, if it exists.

6.1 The Basic Model

Agents differ from one another only by their endowment of skills, denoted by x, that a given test is supposed to measure. The random variable x has a density $f(x)$.

Test scores are being produced using two inputs: skill and effort, denoted by e. The production function of test scores is given by $TS(x,e)$, where test scores are increasing in skills and effort, i.e. $TS_e > 0$ and $TS_x > 0$. I assume further that $TS_{ex} \geq 0$, i.e., a given increase in effort results in weakly higher test scores for agents with higher skills. Producing test scores is costly. The costs associated with effort, $C(x,e)$, are increasing and convex in effort, i.e., $C_e > 0$ and $C_{ee} > 0$. It is

[25] See Section A2 in Appendix A for the construction of this distribution.

[26] 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.

natural to assume that the costs are lower for individuals with higher skills, i.e., $C_x < 0$, and that a given increase in effort is weakly less costly for agents with higher skills, i.e., $C_{xe} \leq 0$.

If no performance-based incentives are supplied and agents gain no psychic benefits from higher test scores, then there are no benefits associated with higher test scores. As effort is costly, agents invest the minimal effort level. In a testing situation the feasible minimal effort is solving no question. However, most survey participants get scores much higher than zero even without performance-based incentives. Thus, I assume that agents obtain psychic benefits from having higher test scores. By definition, when agents take a high-stakes test, i.e., a test in which performance-based incentives (of any kind) are provided, their benefits from higher test scores are not psychic alone. I focus on the provision of piece rate monetary incentives as this is the relevant case for the experiment.[27] In this case, agents' benefits also include their monetary gains from having higher test scores, given by $M(TS;\phi) = A + \phi TS$, where $A \geq 0$ is a constant, and $\phi \geq 0$ is the piece rate amount paid for each correct question (when no monetary incentives are provided $\phi = 0$).

Thus, agents' benefits are given by $U(TS, M; \phi)$, where $U_{TS} > 0$, $U_M > 0$, i.e., agents like to have more money and higher test scores. I assume further that $U_{TS,TS} \leq 0$ and $U_{M,M} \leq 0$, and that agents' benefits are weakly concave in test scores, i.e., $\frac{d^2U}{dTS^2} = \left(U_{TS,TS} + 2\phi U_{TS,M} + \phi^2 U_{M,M}\right) \leq 0$ (this condition is fulfilled if, for example, agents' benefits are separable in money and test scores). As usual, an agent with a skill level x chooses an effort level, e, to maximize benefits minus costs.

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

Proposition 1 indicates that if agents differ only in their skills, test scores provide a correct ranking of agents according to these skills. The result operates through two channels. First, agents with higher skills find it less costly to invest a given level of effort. Second, since test scores are produced using both effort and skill, agents with higher skills, have higher test scores for a given level of effort. As a result, they obtain higher psychic (and if $\phi > 0$ also higher monetary) benefits.

6.2 Types with Different Test-Taking Motivation

To capture the possibility that individuals may differ in their psychic gains from the same test scores, I add types to the basic model. Thus, agents of different types differ their test-taking motivation, i.e., in their psychic gains from the same test scores. The extended setting is as follows.

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' benefits from money, i.e., $U_{M,\theta} = 0$. Assuming, as before, that agents' benefits are a function of test scores and money, we can write their benefits as $U(TS, M; \theta)$. As before, agents are endowed with skills,

[27] The results can be easily extended to situations in which test scores affect agents' future.

denoted by, x. The random variable x has a density which may depend on the type, denoted by $f(x;\theta)$. I assume that given a skill level, x, and an effort level, e, agents of different types will have the same test scores and the same cost function. Hence, the the production and cost functions do not depend on θ and the assumptions regarding these functions are unchanged.[28] An agent of type θ with a skill level x chooses an effort level, e, to maximize benefits minus costs.

Proposition 2 *Conditional on θ, test scores provide a correct ranking of agents according to their skills. If the marginal utility of money is increasing in ϕ, holding θ fixed, an increase in ϕ results in higher effort and higher test scores. Moreover, holding skill level fixed, agents with higher values of θ invest more effort, and have higher test scores.* The Proof is given in Appendix C.

Proposition 2 suggests that if individuals have different test-taking motivation, then, unless the test score distributions of different types do not overlap, test scores do not provide a correct ranking of the population according to individuals' skills. The intuition is as follows. Agents with lower values of θ choose to invest less effort, as they have lower marginal benefits from higher test scores, and the same marginal costs. As test scores are produced using both skill and effort, and types with lower values of θ systematically invest less effort, they have lower test scores. Thus, the comparison of test scores across types is uninformative with respect to their relative skills. A possible way to recover the rank according to skill in the population as a whole is to induce test-takers to invest maximum effort levels. This may be achieved by providing incentives to test-takers.

Proposition 3 *Denote the skill of type θ_i by $x_i(\theta_i)$. $x_i(\theta_i)$ is a random variable with a density function $f(x_i;\theta_i)$, and support $\underline{x}_i \leq x_i(\theta_i) \leq \overline{x}_i$, where $i = 1, 2$ and $\theta_1 > \theta_2$. If $TS(x_1;\phi,\theta_1)$ first order stochastically dominates $TS(x_2;\phi,\theta_2)$, this does not imply that $x_1(\theta_1)$ first order stochastically dominates $x_2(\theta_2)$. However, if all individuals have the same value of θ, denoted by $\tilde{\theta}$, then if $TS(x_1;\phi,\tilde{\theta})$ first order stochastically dominates $TS(x_2;\phi,\tilde{\theta})$, then $x_1(\tilde{\theta})$ first order stochastically dominates $x_2(\tilde{\theta})$.* The Proof is given in Appendix C.

Proposition 3 implies that even if we find, two groups such that the unincentivized test score distribution of one group first order stochastically dominates the unincentivized test score distribution of the other, this may not be the case when incentives are provided. If individuals differ in their test-taking motivation, it is possible that the group with low values of θ (i.e., low test-taking motivation), may have the same (or even higher) skill level than the group with high values of θ.

The three propositions suggest how to investigate whether individuals differ in their test-taking motivation and not only in their skills. The comparison between propositions 1 and 2 suggests that in this case the relative ranking of individuals according to their test scores may change with the provision of incentives. Propositions 3 suggests that first order stochastic domination of test score distribution of one group over another may change with the provision of incentives.

[28] Thus, 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 is given by $C(x, e)$, where $C_e > 0$, $C_{ee} > 0$, $C_x < 0$,and $C_{xe} \leq 0$. Agent of type θ with a skill level x needs to choose an effort level, e, to maximize benefits minus costs.

Note that Proposition 1 shows that even if all individuals value test scores in the same manner, under some conditions (for example, if $U(TS, M) = \tilde{U}(TS) + M$), the provision of incentives will result in an increase in effort, and as a result an increase in test scores. Therefore, investigating whether test scores increase with the provision of incentives is uninformative regarding the existence of individual differences in test-taking motivation.

7 Experimental Evidence

The model implies that to test whether individuals vary in their test-taking motivation one needs to investigate whether individual relative ranking according to their test scores changes under varying incentives schemes. Improvement between tests and narrowing of gaps between groups are feasible even when test-taking motivation do not vary across individuals, as long as they differ in their ability. To find changes in relative ranking one needs to examine test scores of the same individuals for the same test under different incentives schemes. As this data is not available in any of the conventional data sets, I conducted an experiment to investigate whether test-taking motivation determines (at least in part) the coding speed scores. Next I describe the experiment and its results.

7.1 Experimental Design

The experiment consisted of 2 treatments, described below. As the model implies that in order to distinguish between varying test-taking motivation and varying ability one needs to look at rank changes, the design chosen is within subject design. In each of the treatments participants solved three versions of the coding speed test. Each test lasted 10 minutes and consisted of 140 questions.[29]

The experiment was conducted at Harvard using the CLER subject pool and standard recruiting procedures. Overall 127 individuals participated in the two treatments: 99 in six sessions conducted in Spring 2006 for the main treatment (50 men and 49 women) and 28 (14 men and 14 women) in one session conducted in Fall 2006 for 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 had, and that one will be randomly chosen for payment at the end of the experiment. However, participants were only informed of the tasks they need to perform in each part and the compensation scheme immediately before performing the task. The instructions are given in Section D1 in Appendix D. The specific compensation schemes and tasks were as follows. **Main Treatment: Part 1 – Fixed Payment:** Participants were asked to solve two versions of

[29]The tests were constructed in the following manner. For each test, 200 words were randomly chosen from a list of 240 words, and were then randomly ordered to construct 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. Given this construction process, there is no reason to believe tests vary in their degree of difficulty.

the coding speed test, the first was called a practice test. Their payment, if part 1 is randomly chosen for payment, was \$10. Below, I refer to these two tests as the practice test and the \$10 test.

The practice test was administered for two reasons. First, if learning occurs it may be restricted to the duration of this test. Second, if learning is not an issue, then the practice test, as the \$10 test, is administered without performance-based incentives, though (some) participants may have been less motivated to take it. Thus, the comparison between these two tests may help assessing if individual differences in valuation of money are driving the results.

Part 2 – Piece Rate Compensation: Participants were asked to solve a third 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 (known) performance on the \$10 test and a payment based on their future performance on the incentives test. Their payment, if part 2 is randomly selected for payment, was the following. If they chose to be paid according to their past performance they received \$10×(the fraction of \$10 test questions solved correctly). If they chose to be paid according to their future performance they received \$30×(the fraction of incentives test questions solved correctly).

The main purpose of the experiment is to find whether test-taking motivation plays a large role in determining the coding speed scores. To achieve this goal, there has to be a treatment in which participants are motivated to take the test. Thus, if participants are choosing the piece rate, this can serve, at least to some degree, as an indication that the incentives scheme is desirable. 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 (for Learning) Treatment: All three parts in this treatment were identical. In each, participants were asked to solve the coding speed test. They were told that if the current part is randomly selected for payment, they will receive \$10.

Survey: At the end of the experiment, after subjects solved the three tests, they were asked to answer a survey and a psychological questionnaire, designed to detect the "Big Five" constructs.[30,31]

The Testing Program - Performance Measures and Guessing: Figure D1 in Appendix D depicts a typical screen of the testing program. The key and the answers are on the left hand side, while the answer sheet (an electronic "bubble sheet") is on the right. To answer a question subjects had to press one of the radio buttons associated with the question. To see the next (previous) key and the answers associated with it subjects had to press the "Continue" ("Go Back") button. Similarly, subjects could move between the answers on the answer sheet by pressing the "Next" and "Previous" buttons. The testing program recorded all the answers given when any of these buttons was pressed and recorded all answers given every 30 seconds. Using the information gathered by the program, it is possible to identify the 30-second intervals in which participants were guessing,

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

[31] Given the evidence on framing effects (see for example, Tversky and Kahneman, 1981) and stereotype threat effects (see for example, Steele. and Aronson, 1998), the survey was conducted at the end of the experiment.

i.e., answered questions which were part of keys they did not see. Moreover, for each participant we know how many questions they correctly answered in up to twenty 30-second periods.

7.2 Basic Experimental Results

The results reported below include all participants, as all 99 subjects chose the piece rate scheme in the second part. Table 6 reports the means and standard deviations of performance for the three tests. In the first 3 columns, performance is measured by the sum of correct answers on each test. In the last 3 columns, the measure of performance used is the number of correct answers per 30-second period. As it is impossible to know how many questions participants answered correctly in the periods after the first guess (since some of the questions were already guessed correctly), I restrict attention to the periods before the first guess. Participants' performance has 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 unequal 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 seconds (a one-sided t-test allowing for unequal variances yields $p < 0.001$). Between the \$10 and incentives tests participants significantly improved even further, and correctly solved on average 8.2 more questions (a one-sided t-test allowing for unequal variances yields $p = 0.003$). Participants also correctly solved 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 one-sided t-test allowing for unequal variances yields $p < 0.001$).

In addition, when examining the variance of the total number of correct answers in each test a pattern is emerging. 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 and $p = 0.005$ for equality of variances between the \$10 and the practice tests).

The improvement in performance may be in response to the incentives scheme or may indicate learning. To separate the two explanations we would like to know what would have been participants' test score had they took the coding speed test repeatedly without a change in the (implicit and explicit) incentives. The control treatment, in which participants took the test three times for a fixed payment, answers this question directly.[32] The results of the control treatment are very dif-

[32] As it is possible that learning occurs only if incentives are supplied, a treatment in which subjects take the test repeatedly under a piece rate pay scheme will not help in separating the effects of learning from those of incentives.Moreover, changing the order of the tasks (i.e., first administering a test with incentives and then one without) will not help either. The expected result is that subjects will not experienced an increase in scores from an incentivized to an unincentivized test. Interesting as it may be, the reason has nothing to do with learning, but with a crowding out effect the extrinsic incentives will have in this case.

ferent than the ones in the main treatment. Specifically, mean performances (standard deviations) were 90.3 (21.1), 93.6 (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 unequal variances yields $p = 0.26$) instead of a significant increase of 8.2 in the main treatment. Between the first two tests there was an insignificant improvement of 3.3 correct answers on average (a one-sided t-test allowing for unequal variances yields $p = 0.3$) instead of a significant increase of 14 correct answers on average in the main treatment.[33] In addition, while the test score distributions in the first test are not significantly different between the two treatments (Mann-Whitney test yields $p = 0.90$), they differ for the last two tests (Mann-Whitney test yield $p = 0.04$ for the second test and $p < 0.001$ for the third test). The results of the control treatment show that even if learning occurs between the tests, it only occurs if incentives are provided.[34]

7.3 Change in Relative Ranking

The improvement in average performance documented above does not rule out the possibility that the coding speed scores provide a correct ranking of individuals according to their skills when no performance-based incentives are supplied. Next, I investigate this question. Specifically, I ask whether participants who have the same test scores in one test have the same test scores on another and whether participants react differently to the lack of performance-based incentives. While investigating the rank changes directly seems to be the most straightforward way, it necessitates the most ad hoc assumptions, as even small changes in test scores may lead to changes in ranking. For completeness, these results are provided in Section D3 in Appendix D.

7.3.1 Do Different Tests Allow for Comparison between Participants' Ability?

If the coding speed scores provide correct ranking of individuals according to their ability then two individuals with the same test scores on a given test have the same ability. Therefore, they should have the same test scores on any other test, regardless of the incentives. I start by investigating

[33] As a robustness check, I ran the following simulations. I randomly drawn, with replacement, a group of 14 men and 14 women from the main treatment participants, and calculated the mean improvement in their test scores between consecutive tests. I repeated this exercise a 1,000,000 times. The probability that participants in the main treatment would experience an average increase between the practice and \$10 tests smaller than 4 correct answers is 0.0002. The probability that they would experience an average decrease between the \$10 and incentives test of 5 correct answers or more is less than 0.0001.

[34] Consistent with the result of the control treatment, I find little evidence for learning within the tests. In individual fixed effects regressions of the number of correct answers in the 30-second periods before first guess on period number and period number squared, I find for both the \$10 and incentives tests that the number of correct answers is decreasing over time. This decrease is a common finding in the psychological literature and is usually attributed to fatigue or boredom (see for example, Revelle, 1993). For the practice test the relationship between the number of correct answers and time is concave; after about 7 minutes, the number of correct answers is decreasing with time. Learning within the practice test may account for less than half of the increase in test score between the practice and \$10 tests. These results are reported in Table D1 in Appendix D.

whether participants who had the same scores on one test had the same scores on another. Note that this measure can only serve as a lower bound for the amount of rank changing. Specifically, pairs in which two individuals did not have the same scores on one test and changed their relative ranking without having the same scores on another test would not be captured by this measure.

To construct this measure of rank change, the meaning of "having the same test scores" had to be determined. Without adopting a (possibly ad hoc) criterion, it is impossible to use the total test scores to answer this question. Instead, I examine the performance in the 30-second periods before participants started guessing, and use statistical definitions to determine whether two mean performances are the same or not. I use a t-test allowing for unequal variances to test whether the mean performance of every pair of participants is different at the 5% significance level in any two tests. I then count the number of pairs for whom mean performance is significantly different in one test but not in the other. Of the 4656[35] 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.[36]

Even in a very restrictive measure of rank changes, I find that for more than half of the participants the coding speed scores do not provide correct indication regarding their relative abilities. Moreover, the changes in ranks between the practice and the \$10 tests suggest that rank changing cannot only stem from individual differences in the valuation of money. In particular, in terms of monetary incentives there is no change between the practice and \$10 tests. In both, payment does not depend on performance. However, participants did change their behavior. Learning 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 seems 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.

7.4 Do Individuals React Differently to the Lack of Incentives?

Having shown that the individuals' ranking according to their test scores changes, I next provide further evidence that the least motivated individuals, and not the least able ones, are the ones that do not try their best unless performance-based incentives are provided. Economic theory

[35] 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.

[36] 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.

suggests that participants will not invest effort in solving the practice and $10 tests, while they will invest effort when solving the incentives test. The left hand panel of Figure 3 provides an example. The figure depicts the number of correctly answered questions in each 30-second period, before participants numbers 84 and 89 started guessing for each of the three tests. The first 20 periods depict performance in the practice test, periods 21 to 40 performance on the $10 test, and the last 20 periods performance on the incentives one. The (significantly) improved (average) performance on the incentives test suggests that while participant 84 can solve the test, they did not care to show it. Without seeing participant 84 performance on the incentives test, we may have concluded that they cannot solve the coding speed test. The right hand panel of Figure 3 depicts a different behavior by participant 89. Participant 89 does pretty well on the practice test. Nevertheless, once they are told that the test counts (The ($10) test) they improve (significantly) their (average) performance. However, the provision of performance-based incentives does not cause participant 89 to further improve (significantly) their performance. Examining Figure 3 as a whole, we notice that while participant 84 significantly improved their performance between the $10 and incentives tests, and participant 89 did not, participant 84 is still performing worse than participant 89.

The model provide a straightforward way to test whether the participants who behave similarly to participant 84 are less motivated or less able than participants who behave like participant 89. To do that, we need to examine their test score distributions on the $10 and incentives tests and see whether we draw different conclusions regarding their relative ability from the two tests. Note that investigating the average treatment effect will not help us answer this question. As the model shows, it is possible that test scores will increase in response to incentives even without individual differences in test-taking motivation. To classify participants into groups, I examine the improvement in individuals' own performance between the different tests. The measure of performance I use is the mean number of correct answers in the 30-second periods before participants' first guess.[37] Between the $10 and the incentives tests 37 participants out of 99 significantly improved their own performance,[38] while the other 62 participants did not.[39] To simplify the exposition I will refer to the group whose members significantly improved their own performance as "Economists" (as their behavior agrees with economic theory predictions). I will refer to the other group as "Boy Scouts"

[37]Three individuals started guessing early on the $10 and incentives tests. Thus, it is impossible 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 (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 any question, or even a period in which they correctly answered no question. Thus, they all have been classified as experiencing a significant improvement between the $10 and the incentives tests. The results reported below remain qualitatively and quantitatively the same if they are excluded from the analysis.

[38]The criterion used was significance level of 5% or less using a one sided t-tests allowing for unequal variances. As a robustness check I used a significance level of 10%, and the results reported below remain qualitatively the same.

[39]Only two participants experienced a significant decline in their performance, their behavior may be consistent with incentives crowding out intrinsic motivation as modeled in Benabou and Tirole (2003). These 2 participants are assigned to the group that did not significantly improved. The results below remain qualitatively and quantitatively the same if I exclude them from the analysis or assign them to the other group.

(as they seem to be trying their best even when no performance-based incentives were supplied). While individuals in one group significantly improved their performance (the "Economists"), as is clearly demonstrated in Figure 3, the relationship between the total test score distributions of the two groups on different tests cannot be defined theoretically.

Figure 4 presents the cumulative distribution of total test scores of the two groups in the different tests. Panel A presents the total test scores in the incentives test. Panel A suggests that once performance-based monetary incentives are supplied the two distributions of total test scores are the same. To test for stochastic dominance I follow McFadden (1989). Neither the hypothesis that the tests score distribution of the "Economists" first order stochastically dominates the distribution of the "Boy Scouts" can be rejected ($p = 0.420$) nor the opposite one ($p = 0.757$). A similar picture arises when examining the maximum scores each participant achieved in the experiment (Panel B), which are the best estimate of participants' ability. Again, there is no difference in the test score distributions between the two groups. Neither the hypothesis that the tests score distribution of the "Economists" first order stochastically dominates the test score distribution of the "Boy Scouts" can be rejected ($p = 0.266$) nor the opposite one ($p = 0.839$). Thus, these two panels suggest that both groups have the same underlying ability.

A different picture arises when looking at the total test scores of participants in the \$10 and practice tests (Panels C and D, respectively). while we would expect that the "Economists" to do worse, the magnitudes are surprising. In accordance with the rank changes shown before, the hypothesis that the test score distribution of the "Economists" first order stochastically dominates the test score distribution of the "Boy Scouts" 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 be rejected ($p = 0.967$ and $p = 1$ for the practice and the \$10 tests, respectively). Moreover, the differences in the mean performance on the \$10 test are striking. While the "Economists" correctly solved on average 93.4 questions, the "Boy Scouts" correctly solved on average 110.6 questions (a t-test yields $p < 0.001$). This difference is as big as the standard deviation across participants in the incentives test. In contrast, the difference between the groups in the incentives test is 2 correct answers (111.1 for the "Economists" and 113.2 for the "Boy Scouts", a t-test yields $p = 0.57$).

Figure 4 and the subsequent tests suggest that when no performance-based monetary incentives were provided, there is a group of participants that have chosen to invest little effort (the "Economists"). Just looking at their test scores in the \$10 test (or the practice test) one would label them as 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 (the "Boy Scouts").

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 their own average performance between the \$10 and incentives

tests, 42 have significantly improved their own average performance between the practice and \$10 tests.[40] This suggest that while some participants tried their best already in the practice test (20 participants overall, their mean performance in the practice test is 96.25), others needed 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.[41]

The basic experimental results indicate that the variance of the test scores is the largest in the \$10 test. Now we have an explanation. While in the incentives and practice tests the heterogeneity in motivation amongst subjects does not play a role, as in the former all are motivated, and in the latter most invest little effort. It does play a role in the \$10 test in which about 60% invest high effort and the rest invest little effort. As a result the variance in test scores increases.

7.4.1 Individual Characteristics and Effort Choice in Tests without Incentives

In this section I investigate whether individual characteristics are correlated with participants' effort choices. I start by investigating the relationship between gender and effort choices. I find that women are more likely to invest high effort even without performance-based incentives. Of the 49 female participants, only 14 (28.6%) were classified as "Economists". In contrast, out of the 50 male participants 23 (46%) were classified as "Economists". A Chi-squared test for the equality of the distributions yields $p = 0.073$. While this may seem at odds with the evidence gathered in field experiments suggesting that females are more likely to improve their performance with the provision of incentives (see for example Angrist and Lavy, 2004), the evidence from psychology may provide an explanation. Duckworth and Seligman (2006) find that as women are more self discipline than men, they outperform men on tasks that require long term investment (like grades in school). While in field experiments students need to invest high effort for long periods of time, in the experiment, effort is concentrated in a very short period of time (10 minutes). Thus, these differences in the period during which individuals need to invest effort may explain the differences in results regarding gender. Given these differences across gender, below I examine the relationship between effort choices and individual characteristics by gender.

At the end of the experiment, as part of the survey, participants were asked to report their SAT scores, all but 2 men and a woman did so. As participants in the two groups do as well on the incentives test, we should expect to see no differences in their SAT scores. Indeed, SAT scores do not relate to individuals' effort choices. For the participants who reported SAT scores, the average SAT scores of male "Economists" is 1,433 while the average SAT scores of males "Boy Scouts" is

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

[41] 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 only did their best when monetary incentives were supplied.

1,454. The average SAT scores of female "Economists" is 1,427 while the average SAT scores of female "Boy Scouts" is 1,404. These differences are not significant, a one sided t-test allowing for unequal variances yield p = 0.32 and p = 0.72 for men and women, respectively.[42]

After solving the three tests participants were asked to answer the "Big 5" questionnaire. The "Big Five" theory is part of a research in psychology, dating back to the 1930's, trying to empirically find the most important ways in which individuals differ from one another. The common classification of personality traits is to five dimensions, which are referred to as the big five.[43] These five constructs are defined as follows (following Roberts et al., 2004). Extroversion reflects the tendency to be socially active and assertive; Agreeableness the tendency to be trusting, modest, altruistic, and warm; Conscientiousness the tendency to be rule following, task- and goal-directed, planful, and self controlled; Neuroticism contracts the experience of anxiety, worry, anger, and depression with even-temperedness; Openness to experience reflects the tendency to be open to new ideas, complex, original, and creative. The big five literature relates these traits to various aspects of individuals' life, including economic success. Of these traits, conscientiousness is the one that most consistently relates to job performance, job seeking behavior, and retention at work (Judge et al., 1999). The "Big 5" questionnaire includes 50 statements (10 for each construct).[44] Participants were asked to indicate on a five-point scale how accurately each statement describes their usual behavior. To create the five constructs the answers to the questionnaire were added within each construct.[45]

Overall, 91 participants (44 men and 47 women) had answered all 50 questions, of those 2 men and a women did not report SAT scores. The results reported below are for this restricted sample. For male participants, effort choices can be related to conscientiousness. Specifically, the average conscientiousness level of male "Economists" is 8.3 while the average conscientiousness level of male "Boy Scouts" is 13.4 (a one sided t-test allowing for unequal variances yields p = 0.024).[46] No other personality construct is related to male participants' type. I find no personality trait that can predict female effort choices.

Table 7 investigates further the relationship between the answers to the "Big 5" questionnaire

[42]For the restricted sample of 88 subjects who also answered the "Big 5" questionnaire in full, the average SAT scores of male "Economists" is 1,429 while the average SAT scores of males "Boy Scouts" is 1,457. The average SAT scores of female "Economists" is 1,427 while the average SAT scores of female "Boy Scouts" is 1,397. These differences are not significant, a one sided t-test allowing for unequal variances yield p = 0.29 and p = 0.77 for men and women, respectively.

[43]For an excellent summary on the big five theory see McCrae and John (1992).

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

[45]When a question was phrased in a negative manner (e.g., "Worry about things") the answers were subtracted.

[46]As a robustness check I used probit regressions to impute the missing answers on the "big 5" questionnaire. To predict the missing value, the answers of other participants, of the same gender, to the questions within the same personality construct were used. Using the imputed values, the average conscientiousness level of male "Economists" is 9.3 while the average conscientiousness level of male "Boy Scouts" is 12.4. This difference is statistically significant; a one sided t-test allowing for unequal variances yields p = 0.072.

and participants' choice of effort in dept. Panel A (B) of Table 7 describes the mean answers to the questions in which male (female) "Boy Scouts" significantly differ from male (female) "Economists". Panel A, shows why differences are found for male participants. Specifically, male "Boy Scouts" have consistently favorable answers to the questions in the consciousness construct. In Addition, differences in answers to some questions on other constructs suggest that male "Boy Scouts" may be extraverts and neurotic. These differences may indicate that male "Boy Scouts" are more competitive. In Psychology, Graziano et al. (1985) find extravert individuals to be more competitive. In economics, Charness and Grosskopf (2001) report that unhappy individuals have competitive preferences. Diener et al. (1990) report that the unhappy people are more likely to be neurotic.

In contrast, no coherent pattern emerges when examining female participants' answers (Panel B of Table 7). If anything, female answers seem to be affected by the experiment (female "Economists" insist that they do not shirk on their duties and have excellent English knowledge). Nevertheless, this is not a source of concern for the following reasons. All participants went through the same treatments (i.e., took the practice, \$10, and incentives tests). Moreover, "Boy Scouts" and "Economists" have the same test score distributions on the incentives test (true for men and women separately). The differences are on the \$10 test (and the practice test). Therefore, if something in the experiment caused female "Economists" to answer the questionnaire differently than female "Boy Scouts", it has to relate to their choices in the experiment. However, while female participants' answers to the questionnaire support the notion that participants of two groups behaved differently in the experiment, their effort choices cannot be linked to their personality traits.

As part of the survey participants were asked to report the university they attend at and their major. For male participants I find no correlations between effort choices and either their major or the university they attend.[47] While for female participants I find no correlations between effort choices and the university they attend,[48] I do find correlations between their effort choices and their major. Specifically, of the 16 women majoring in Arts or Humanities only one (6.25%) is an "Economist", while of the 26 women majoring in Social science or business 11 (42.3%) are "Economists" and of the 7 women majoring in Physical or Natural sciences 2 (28.5%) are "Economists". Fisher exact test yields $p = 0.03$. While female participants majoring Arts and Humanities are less likely to be "Economists" than female participants majoring in Business and Social sciences (fisher exact test yields $p = 0.015$), female participants majoring Physical or Natural sciences are as likely to be "Economists" as are female participants majoring Arts and Humanities

[47]Of the 39 men enrolled at Harvard or MIT 17 (40%) are "Economists", of the 4 men enrolled in BU 2 (50%) are "Economists", and of the 7 men enrolled in smaller universities in the Boston area 4 (57.1%) are "Economists". Fisher exact test for the equality of the distributions yields $p = 0.879$. Of the 15 men majoring in Arts or Humanities 6 (40%) are "Economists", of the 23 men majoring in Social science or business 11 (47.8%) are "Economists", and of the 12 men majoring in Physical or Natural sciences 6 (50%) are "Economists". Fisher exact test yields $p = 0.873$.

[48]Of the 32 women enrolled at Harvard 8 (25%) are "Economists", of the 11 enrolled in BU 4 (36.4%) are "Economists", and of the 6 enrolled in universities in the Boston area 2 (33.3%) are "Economists". Fisher exact test yields $p = 0.642$.

or in Business and Social sciences (fisher exact tests yield p = 0.209 and p = 0.676, respectively).

8 Discussion and Conclusions

The analysis in this paper focuses on a very simple test, the coding speed test, whose scores may strongly depend on individuals' test-taking motivation due to its simplicity. Experimental data, the NLSY survey data, and data from the armed forces are utilized to investigate the relationship between the coding speed scores, test-taking motivation, cognitive skills, and economic success.

In the NLSY sample I find that, controlling for the AFQT scores, an increase in the coding speed scores is significantly associated with an increase in earnings of male workers. Moreover, the coding speed scores are relatively more important to the earnings of low educated workers, while the AFQT scores are relatively more important to earnings of highly educated ones. The data available from the armed forces shows that potential recruits scored higher than the NLSY participants on the coding speed test. The experimental results show that subjects responded differently to the lack of incentives. About 40% of subjects improved their own performance with the provision of incentives, while the rest did not. Both groups, though, had the same test score distributions when incentives were provided. Moreover, those male participants who are more conscientious and female participants were more likely to invest high effort in the test without performance-based incentives.

While there are several explanations that could account for each of these findings alone, one simple explanation could account for all. Namely, due to the simplicity of the coding speed test, its scores, when no performance-based incentives are provided, are highly correlated with test-taking motivation. Moreover, test-taking motivation itself correlates with personality traits like conscientiousness that are valued in the market, but not with cognitive ability. Thus, potential recruits are doing better than NLSY participants on the coding speed test since they are more motivated to take it. At the same time, for the NLSY participants, the coding speed test is a low-stakes test, thus, higher scores indicate favorable (in the labor market) personality traits. As a result we find correlations between the coding speed scores and earnings.

The experimental results show that some individuals do not try their best when no performance-based incentives are provided, while others do. The individuals that do not try their best are not the least able ones. If this behavior does not depend on a particular test, then all low-stakes test scores will measure a combination of cognitive and non-cognitive skills.[49] This suggests that inferences from test score distributions to ability distributions may be questionable. Unless evidence is provided that all test-takers tried their best or other controls are used, caution should be exercised when interpreting results where test scores are either the dependent or the independent variable.

Not all the implications of the results presented in this paper are negative. In particular, as

[49]While evidence in Borghans et al. (2008) and Gneezy and Rustichini (2000) are not sufficient to determine the relationship between ability and effort invested in low-stakes IQ and SAT-like tests, they do not refute this possibility.

long as the purpose of using test scores is to have a measure of (unobserved) individual characteristics important for economic success, low-stakes test scores may even be a better measure than previously assumed. Not only low-stakes test scores combine cognitive and non-cognitive measures, the environment in which they are obtained (i.e., no explicit performance-based incentives and no explicit monitoring) probably resembles the typical work place to a large degree. Workers that invest high effort in this environment and have the necessary skills are probably more valuable.

The results in this paper may have bearing beyond the academic discussion, particularly, with the No Child Left Behind (NCLB) Act of 2001. The NCLB test is a high-stakes test for schools, which may lose funding and close if their students perform poorly. For students, though, it is a low-stakes test, as their scores do not directly affect them. While, in the short run, schools may lose funding and close even if their students possess the required knowledge, if they are unmotivated, in the long run, the NCLB act may have positive effects. Given the NCLB incentives scheme, schools can "pass" the exam only if their students know the required material and are motivated to show it. Thus, "teaching to the test" can help schools only if complemented by motivating their students.[50] Anecdotal evidence in the press suggests that schools are well aware of this. In particular, principals try to motivate students by creating "school spirits" and provide prizes to students who do well (see for example, "Successes at a Big-City System; Focus, Funding Help Turn Around Nation's 8^{th}-Largest School District", Washington Post, June 12, 2007, and "A school's comeback formula: Expel cynicism, stress reform", The Boston Globe, November 26, 2006). Thus, in the long run, the NCLB act may result in students having more knowledge and higher non-cognitive skills.

References

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 Paper No. 1146.

Benabou, R. and Tirole, J. 2003. "Intrinsic and Extrinsic Motivation". *Review of Economic Studies*, 70(3):489–520.

Benjamin, D. J., Brown, S. A., and Shapiro, J. M. 2005. "Who is 'Behavioral'? Cognitive Ability and Anomalous Preferences". Harvard University Working Paper.

Borghans, L., Meijers, H., and ter Weel, B. 2008. "The Role of Noncognitive Skills in Explaining Cognitive Test Scores". *Economic Inquiry*, 46(1):2–12.

Bowles, S., Gintis, H., and Osborne, M. 2001. "The Determinants of Earnings: A Behavioral Approach". *Journal of Economic Literature*, 39(4):1137–1176.

[50] Note that as far as parents would like school to transmit knowledge and to instil values associated with being motivated, an incentives scheme like NCLB can solve the distortion problems that could arise in multitasking environment as descried in Holmstrom and Milgrom (1991).

Camerer, C. F. and Hogarth, R. M. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework". *Journal of Risk and Uncertainty*, 19:7–42.

Cascio, E. U. and Lewis, E. G. 2006. "Schooling and the Armed Forces Qualifying Test". *Journal of Human Resources*, 41(2):294–318.

Cawley, J., Conneely, K., Heckman, J., and Vytlacil, E., "Cognitive Ability, Wages, and Meritocracy". In S. E. Devlin, D. R. Fienberg, and K. Roeder (Eds.) "Intelligence Genes, and Success: Scientists Respond to the Bell Curve", pp. 179–192. Copernicus:Springer-Verlag 1997.

Charness, G. and Grosskopf, B. 2001. "Relative Payoffs and Happiness: an Experimental Study". *Journal of Economic Behavior and Organization*, 45(3):301–328.

Department of Defense, U. S. 1982. "Profile of American Youth: 1980 Nationwide Administration of the Armed Services Vocational Aptitude Battery". Office of the Assistant Secretary of Defense.

Diener, E., Suh, E. M., Lucas, R. E., and Smith, H. L. 1990. "Subjective Well-Being: Three Decades of Progress". *Psychological Bulletin*, 125(2):276–302.

Dohmen, T. J., Falk, A., Huffman, D., and Sunde, U. 2008. "Are Risk Aversion and Impatience Related to Cognitive Ability?" CEPR Discussion Paper No. 6852.

Duckworth, A. L. and Seligman, M. E. 2006. "Self-Discipline Gives Girls the Edge: Gender in Self-Discipline, Grades, and Achievement Test Scores". *Journal of Educational Psychology*, 98(1):198–208.

Fehr, E. and Falk, A. 1999. "Wage Rigidity in a Competitive Incomplete Contract Market". *Journal of Political Economy*, 107(1):106–134.

Fehr, E. and Schmidt, K. M., "Theories of Fairness and Reciprocity: Evidence and Economic Applications". In M. Dewatripont, L. Hansen, and S. Turnovsky (Eds.) "Advances in Economics and Econometrics, Eighth World Congress of the Econometric Society", pp. 208–257. Cambridge University Press, Cambridge 2003.

Gneezy, U. and Rustichini, A. 2000. "Pay Enough Or Don'T Pay At All". *The Quarterly Journal of Economics*, 115(3):791–810.

Graziano, W. G., Feldesman, A. B., and Rahe, D. F. 1985. "Extraversion, Social Cognition, and the Salience of Aversiveness in Social Encounters". *Journal of Personality and Social Psychology*, 49(4):971–980.

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):39–98.

Heckman, J. 1995. "Lessons from the Bell Curve". *Journal of Political Economy*, 103(5):1091–1120.

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):411–482.

Herrnstein, R. J. and Murray, C., *The Bell Curve: Intelligence and Class Structure in American Life*. The Free Press, New York 1994.

Holmstrom, B. and Milgrom, P. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design". *Journal of Law, Economics and Organization*, 7:24–52.

Judge, T. A., Higgins, C. A., Thoresen, 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(3):621–652.

Kremer, M., Miguel, E., and Thorton, R. 2005. "Incentives to Learn". NBER Working Paper No. 10971.

Kuhn, P. and Weinberger, C. 2005. "Leadership Skills and Wages". *Journal of Labor Economics*, 23(3):395–434.

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.

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

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

McCrae, R. R. and John, O. P. 1992. "An Introduction to the Five-Factor Model and Its Applications". *Journal of Personality*, 60(2):175–215.

McFadden, D. L., "Testing for Stochastic Dominance". In T. Fomby and T. Seo (Eds.) "Studies in the Economics of Uncertainty", pp. 113–134. Springer, New York 1989.

Mulligan, C. B. and Rubinstein, Y. 2005. "Selection, Investment, and Women's Relative Wages Since 1975". NBER Working Paper No. 11159.

Nagin, D., Rebitzer, J., Sanders, S., and Taylor, L. 2005. "Monitoring, Motivation and Management: The Determinants of Opportunistic Behavior in a Field Experiment". *American Economic Review*, 92(4):850–873.

Neal, D. 2004. "The Measured Black-White Wage Gap among Women Is Too Small". *Journal of Political Economy*, 112(S1):S1–S28.

Neal, D. A. and Johnson, W. R. 1996. "The Role of Premarket Factors in Black-White Wage Differences". *Journal of Political Economy*, 104(5):869–895.

Neal, D. A. and Johnson, W. R., "Basic Skills and the Black-White Earnings Gap". In C. Jencks and M. Phillips (Eds.) "The Black-White Test Score Gap", pp. 480–498. The Brookings Institute, Washington DC 1998.

Persico, N., Postlewaite, A., and Silverman, D. 2004. "The Effect of Adolescent Experience on Labor Market Outcomes: The Case of Height". *Journal of Policital Economy*, 112(5):1019–1053.

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):157–169.

Revelle, W., "Individual Differences in personality and motivation: 'Non-cognitive' determinants of cognitive performance". In A. Baddeley and L. Weiskrantz (Eds.) "Attention: Selection, Awareness and Control: A tribute to Donald Broadbent", pp. 346–373. Oxford University Press, Oxford 1993.

Roberts, B. W., Robins, R. W., Trzesniewski, K. H., and Caspi, A., "Personality Trait Development in Adulthood". In M. J. S. Jeylan T. Mortimer (Ed.) "Handbook of the Life Course", pp. 579–598. Springer, New York 2004.

Segal, C. 2005. "Misbehavior, Education, and Labor Market Outcomes". Stanford University Working Paper.

Segal, C. forthcoming. "Classroom Behavior". *Journal of Human Resources*, 43(4). 2008.

Steele., C. M. and Aronson, J., "Stereotype Threat and the Test Performance of Academically Successful African American". In C. Jencks and M. Phillips (Eds.) "The Black-White Test Score Gap", pp. 401–427. The Brookings Institute, Washington DC 1998.

Tversky, A. and Kahneman, D. 1981. "The Framing of Decisions and the Psychology of Choice." *Science*, 211(4481):453–458.

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

Table 1: Mean and Standard Deviation of Key Outcome Variables for Men by Cohort-Adjusted Coding Speed Test Scores[1,2]
Low Coding Speed Test Scores - Men with Coding Speed Test Sores Below the Mean[3]
High Coding Speed Test Scores - Men with Coding Speed Test Sores Above the Mean[3]

Variable	Low Coding Speed Scores Mean	Low Coding Speed Scores Standard Deviation	High Coding Speed Scores Mean	High Coding Speed Scores Standard Deviation	Difference	Observations
% 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 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. 47% of men had coding speed scores lower than the mean.
4. 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 2: Earnings Men
Dependent Variable: Log of Earnings 2003

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Black	-0.555	-0.212	-0.378	-0.202	-0.286	-0.276	-0.276	-0.287
	(0.066)***	(0.069)***	(0.066)***	(0.069)***	(0.069)***	(0.070)***	(0.070)***	(0.070)***
Hispanic	-0.331	-0.120	-0.245	-0.123	-0.150	-0.151	-0.151	-0.150
	(0.076)***	(0.072)*	(0.071)***	(0.071)*	(0.069)**	(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	0.262	0.289
						(0.067)***	(0.067)***	(0.066)***
AFQT - Less than College Degree						0.095	0.095	0.132
						(0.044)**	(0.044)**	(0.042)***
CS - College Graduates or More							0.069	
							(0.031)**	
CS - Less than College Degree							0.069	
							(0.051)	
Years of Schooling Completed 2003					0.104	0.084	0.084	0.088
					(0.014)***	(0.016)***	(0.016)***	(0.015)***
Age in 2003	0.026	0.015	0.020	0.015	0.025	0.023	0.023	0.024
	(0.031)	(0.029)	(0.030)	(0.029)	(0.029)	(0.029)	(0.029)	(0.029)
Constant	9.681	10.042	9.892	10.062	8.251	8.588	8.588	8.494
	(1.257)***	(1.191)***	(1.221)***	(1.188)***	(1.223)***	(1.186)***	(1.184)***	(1.187)***
Observations	1187	1187	1187	1187	1187	1187	1187	1187
R-squared	0.05	0.18	0.12	0.18	0.23	0.24	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 1^{st} 1961 and September 30^{th} 1964, who competed the ASVAB test, were not given "Spanish Instructions Cards", and did not belong to the over-sample. The sample was restricted further to include only civilian workers who reported positive earnings and 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 A for details).

Table 3: Earnings Men - Using the 4 AFQT Sub-Tests Separately
Dependent Variable: Log of Earnings 2003

	(1)	(2)	(3)	(4)
Black	-0.245	-0.378	-0.232	-0.299
	(0.069)***	(0.066)***	(0.069)***	(0.068)***
Hispanic	-0.132	-0.245	-0.133	-0.155
	(0.071)*	(0.071)***	(0.070)*	(0.068)**
Arithmetic Reasoning	0.075		0.075	0.064
	(0.043)*		(0.042)*	(0.040)
Math Knowledge	0.180		0.157	0.055
	(0.051)***		(0.051)***	(0.055)
Word Knowledge	0.024		0.016	-0.028
	(0.045)		(0.045)	(0.043)
Paragraph Comprehension	0.085		0.058	0.046
	(0.047)*		(0.049)	(0.048)
Coding Speed		0.245	0.087	0.062
		(0.026)***	(0.028)***	(0.027)**
Years of Schooling 2003				0.104
				(0.014)***
Age in 2003	0.014	0.020	0.013	0.024
	(0.029)	(0.030)	(0.029)	(0.029)
Constant	10.097	9.892	10.119	8.311
	(1.191)***	(1.221)***	(1.189)***	(1.227)***
Observations	1187	1187	1187	1187
R-squared	0.18	0.12	0.19	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 1^{st} 1961 and September 30^{th} 1964, who competed the ASVAB test, were not given "Spanish Instructions Cards", and did not belong to the over-sample. The sample was restricted further to include only civilian workers who reported positive earnings and 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 A for details).

Table 4: 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 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 1^{st} 1961 and September 30^{th} 1964, who competed the ASVAB test, 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 A for details).
3. See Appendix A for details regarding the occupation data.

Table 5: Relationship between AFQT and Coding Speed Scores and Family Background and School Characteristics

	Dependent Variable: AFQT Scores			Dependent Variable: Coding Speed Scores		
	(1)	(2)	(3)	(4)	(5)	(6)
Black	-0.675	-0.565	-0.497	-0.466	-0.411	-0.356
	(0.051)***	(0.073)***	(0.074)***	(0.058)***	(0.083)***	(0.085)***
Hispanic	-0.289	-0.161	-0.068	-0.046	0.116	0.195
	(0.061)***	(0.088)*	(0.087)	(0.068)	(0.090)	(0.093)**
Mother High School Grad.	0.308	0.276	0.254	0.186	0.197	0.179
	(0.057)***	(0.079)***	(0.078)***	(0.065)***	(0.090)**	(0.091)**
Mother College Grad.	0.415	0.323	0.300	0.203	0.225	0.208
	(0.093)***	(0.129)**	(0.121)**	(0.123)*	(0.172)	(0.166)
Father High School Grad.	0.208	0.207	0.204	0.204	0.172	0.170
	(0.058)***	(0.079)***	(0.078)***	(0.062)***	(0.085)**	(0.085)**
Father College Grad	0.558	0.605	0.575	0.360	0.382	0.358
	(0.082)***	(0.107)***	(0.103)***	(0.099)***	(0.134)***	(0.130)***
Mother Professional	0.207	0.175	0.155	-0.001	-0.087	-0.102
	(0.083)**	(0.120)	(0.118)	(0.120)	(0.177)	(0.179)
Father Professional	0.133	0.045	0.034	0.092	-0.050	-0.057
	(0.078)*	(0.109)	(0.105)	(0.098)	(0.146)	(0.145)
Did Not Live with Both	-0.028	-0.099	-0.078	-0.041	-0.074	-0.057
Biological Parents at Age 14	(0.054)	(0.076)	(0.075)	(0.059)	(0.085)	(0.084)
Number of Siblings	-0.044	-0.039	-0.041	-0.038	-0.028	-0.030
	(0.009)***	(0.013)***	(0.013)***	(0.010)***	(0.014)**	(0.014)**
No Reading Materials	-0.346	-0.428	-0.398	-0.279	-0.394	-0.368
at Age 14	(0.081)***	(0.114)***	(0.112)***	(0.102)***	(0.134)***	(0.135)***
Numerous Reading Materials	0.314	0.336	0.313	0.236	0.213	0.194
at Age 14	(0.049)***	(0.068)***	(0.067)***	(0.054)***	(0.075)***	(0.075)***
Student/Teacher Ratio			-0.019			-0.016
			(0.007)**			(0.008)*
Disadvantage Student Ratio			-0.002			-0.001
			(0.002)			(0.002)
Dropout Rate			-0.004			-0.003
			(0.001)***			(0.001)**
Teacher Turnover Rate			-0.010			-0.009
			(0.004)**			(0.005)*
Age 1980	0.030	0.037	0.026	0.008	0.029	0.021
	(0.023)	(0.033)	(0.033)	(0.027)	(0.037)	(0.037)
Constant	-0.817	-0.957	-0.225	-0.304	-0.686	-0.090
	(0.410)**	(0.585)	(0.601)	(0.467)	(0.655)	(0.676)
Observations	1961	1027	1027	1961	1027	1027
R-squared	0.38	0.33	0.35	0.18	0.15	0.16

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 competed the ASVAB test, 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 2,3,5, and 6 individuals with missing school data were excluded).
2. AFQT and coding speed scores are school-year cohort adjusted (see Appendix A for details).
3. All specifications also include a dummy equal to one if the information regarding parents' educational attainments is 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.

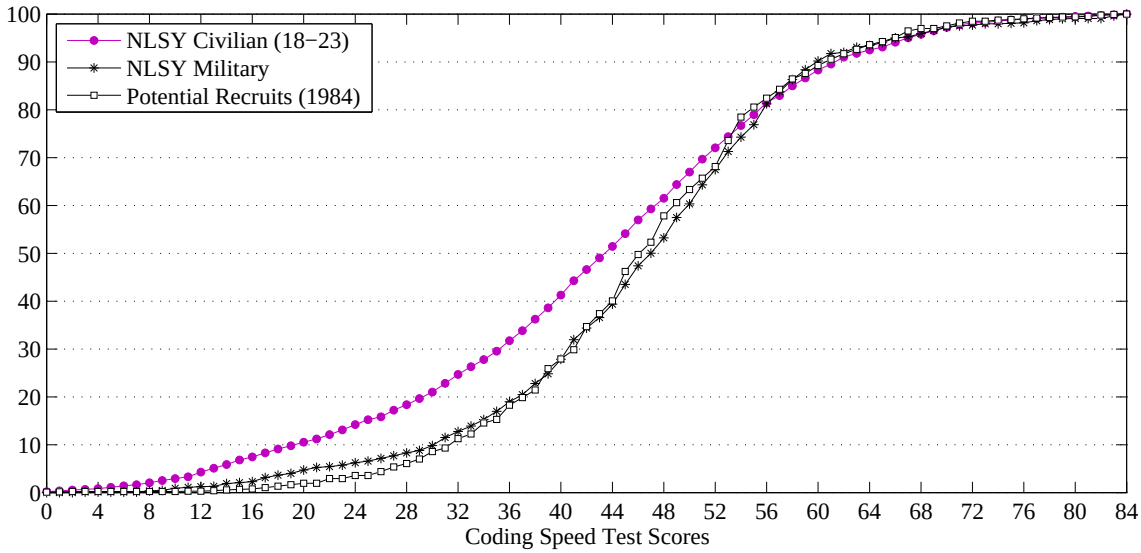


Figure 2: CDFs of Raw Coding Speed Test Scores for NLSY participants and Potential Recruits to the Armed Forces - Men

Table 6: 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

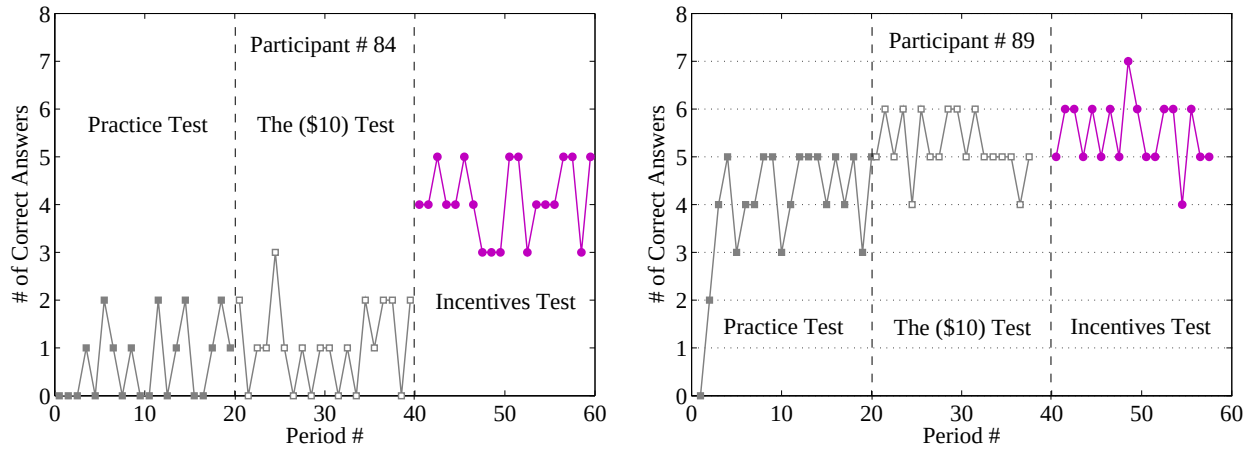
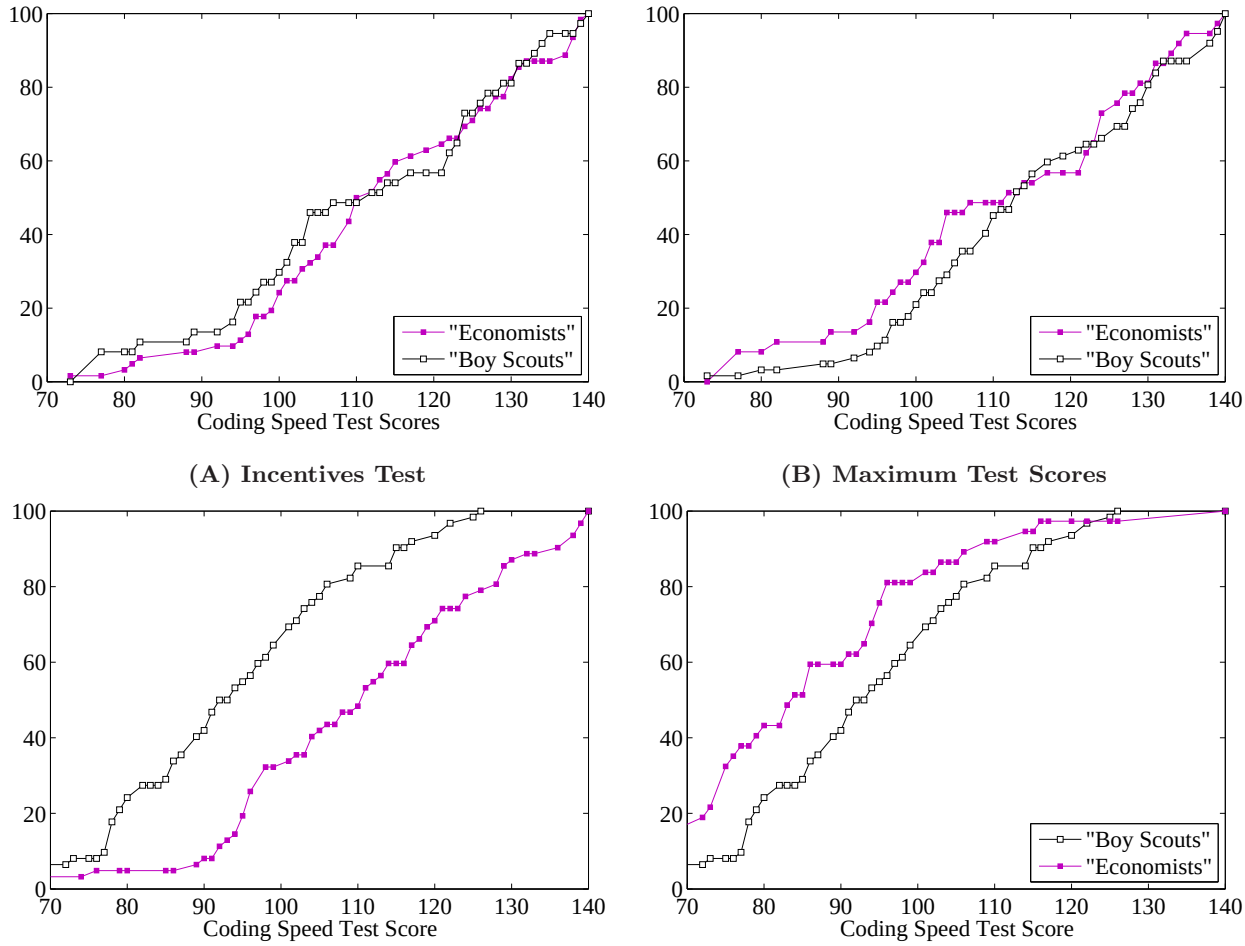


Figure 3: Number of Correct Answers in 30-second Periods before First Guess by Test



(A) Incentives Test **(B) Maximum Test Scores**

(C) $ 10 Test **(D) Practice Test**

Figure 4: CDFs of Coding Speed Test Scores by Compensation Scheme and Participants Type. "Economists" are participants who improved their own average performance significantly between the $10 and incentives tests. "Boy Scouts" are participants who did not.

Table 7: Differences in Mean Answers to the "Big Five" Questionnaire by Participants' Effort Choices

Panel A: Men

Question from "Big five" Questionnaire	"Boy Scouts"	"Economists"	Difference
Consciousness			
Am always prepared (+)	3.633	3.2	0.436 (p = 0.053)
Pay attention to details (+)	4.182	3.6	0.582 (p = 0.024)
Get chores done right away (+)	3.182	2.3	0.882 (p = 0.010)
Follow a schedule (+)	3.682	3.1	0.582 (p = 0.045)
Like order (+)	3.909	3.35	0.559 (p = 0.067)
Shirk my duties (-)	1.864	2.35	-0.486 (p = 0.057)
Often forget to put things back in their proper place (-)	2.409	3	-0.591 (p = 0.088)
Neuroticism			
Worry about things (-)	3.727	3.5	0.477 (p = 0.061)
Seldom feel blue (+)	2.727	3.05	-0.322 (p = 0.155)
Change my mood a lot (-)	3.136	2.65	0.486 (p = 0.060)
Extraversion			
Keep in the background (-)	2.864	3.2	-0.336 (p = 0.174)
Don't like to draw attention to myself (-)	2.955	3.7	-0.745 (p = 0.013)
Agreeableness			
Take time out for others (+)	3.545	4	-0.454 (p = 0.039)

Panel B: Women

Question from "Big five" Questionnaire	"Boy Scouts"	"Economists"	Difference
Consciousness			
Pay attention to details (+)	4.344	3.857	0.487 (p = 0.069)
Shirk my duties (-)	2.188	1.714	0.473 (p = 0.062)
Neuroticism			
Am relaxed most of the time (+)	3.5	2.857	0.7642 (p = 0.021)
Worry about things (-)	3.719	3.286	0.433 (p = 0.072)
Openness to Experience			
Have a rich vocabulary (+)	3.688	4.285	-0.598 (p = 0.006)
Have excellent ideas (+)	4.219	3.857	0.362 (p = 0.036)
Use difficult words (+)	3.406	4	-0.593 (p = 0.011)
Am full of ideas (+)	4.25	3.857	0.393 (p = 0.093)

Notes:

1. "Economists" are participants who improved their own average performance significantly between the $10 and incentives tests. "Boy Scouts" are participants who did not.
2. (+) and (-) next to each question indicates whether the question indicates a positive or negative trait.
3. The sample is restricted to include only participants who answered the Big 5 Questionnaire in full and reported their SAT scores.
4. p-values are for the one sided t-test allowing for unequal variances.
5. Cells in grey: Significance of differences across types depends on the sample used. In particular, significance changes when the sample is only restricted to include all participants who answered a specific question.

Appendix A: Data Appendix

Table A1: 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 (speeded)
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 math principles
Mechanical Comprehension	19	25	Measures knowledge of mechanical and physical principles, and the ability to visualize how illustrated objects work
Electronics Information	9	20	Tests knowledge of electricity and electronics

A.1 NLSY Sample Restrictions and Variable Construction

The sample used in Section 4 is restricted to include only individuals that have valid test scores who were surveyed in 2004.[51] To try and avoid endogeniety problems, in particular that either test scores or test-taking motivation may be affected by labor market experience, the sample was restricted further to include only the three youngest school-year cohorts discussed below, i.e., participants born between September 1^{st} 1961 and August 31^{st} 1964.[52]

Participants in the NLSY took the ASVAB exam in the summer of 1980 when they were 16-23 years old. This ages differential implies differences in educational attainment, which may affect test scores (see for example Hansen et al., 2004, Cascio and Lewis, 2006). In addition, it is possible that older individuals may be more mature, which may affect their test scores. Indeed, the ASVAB scores increase with age, thus, in order to compare test scores of individuals of different age groups an adjustment is needed. I adjust the test scores used in the analysis in Section 4 by school-year cohorts, where a school year-cohort includes all individuals born between October 1^{st} of one calendar year and September 30^{th} of the subsequent one. Specifically, the residuals from regressions of the test scores variables on school-year cohort indicators for the restricted sample,

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

[52] 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 Section 4 remains qualitatively and quantitatively the same if the sample is restricted to include only the two youngest cohorts or all individuals born before January 1^{st} 1961.

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 scores than men, the adjustment was done separately by gender. 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 one. An additional benefit of the school-year cohort normalization is that it excludes participants who were born after September 30^{th} 1964 from the analysis, which are believed to be a non-random sample.[53]

The variable years of schooling completed used in the earnings regressions 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. For individuals that reported that they have not received a high school diploma the actual years of schooling reported were used. Individuals who reported receiving a high school diploma were assigned 12 years of schooling. For participants who reported completing at least a year of post secondary degree but did not receive a degree 13 years of schooling were assigned.[54] 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 earnings regressions for individuals of different occupations, I used the wage and the occupation reported for job number 1 in 2004. The sample was restricted to include all civilian workers not enrolled in school reporting positive wages on job number 1 in 2004 for whom dada on schooling and test scores is available. The occupation is determined using the 2000 Census occupational categories. Production workers are workers who reported a job in the categories Production and Operating Workers or Setter, Operators, and Tenders. Mangers are workers who reported a job in Executive, Administrative, and Managerial Occupation or in Management Related Occupations. Professionals are workers who reported a job in Mathematical and Computer Scientists, Engineers, Architects, and Surveyors, Physical Scientists, Social Scientists and Related Workers categories, or reported being Lawyers or Judges, Magistrates, and Other Judicial Workers.

In all the regression results reported in Section 4 I use sampling weights. In regressions where the AFQT or the coding speed scores are the dependent variable I use the provided ASVAB sampling weights. In regressions where earnings in 2003 or wages in 2004 are the dependent variable I use the 2004 cross-sectional weights.

[53]The NLSY sample includes "too few" participants born after September 30^{th} 1964 in comparison to the general population (NLSY79 User guide pp.19-20).

[54]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 degree. Thus, to maintain that those with more years of competed schooling actually have higher educational attainment this coding was chosen.

A.1.1 Problem with ASVAB Norming

The purpose of administrating the ASVAB to the NLSY participants was to obtain data on the vocational aptitudes of American youth during the 1980s and to establish national norm for the ASVAB. Previously, military recruits had been compared statistically to adult males who were tested during World War II. To obtained this norm, the ASVAB had to be administered to a representative sample of American youth. As the U.S. Department of Labor had already started conducting the NLSY survey, it was decided to administer the ASVAB to the NLSY participants. The ASVAB adminstration process is known as the Profile of American Youth (PAY80).

As discussed in Section 5, while trying to establish the norm for the ASVAB Maier and Sims (1983) had problems in comparisoning the ASVAB scores between potential recruits to the armed forces and NLSY participants of comparable ages (i.e., born before 1/1/1963).[55] By 1985 the problem was considered solved by the military (Maier and Sims, 1986). The differences in the scores between the NLSY participants and potential recruits were attributed to differences in the shape of answer sheets (the NLSY participants filled a "bubble sheet" while potential recruits filled "slim rectangles") and its layout (the answer sheet used by the military corresponded exactly with the layout of the questions). Ree and Wegner (1990) have shown in a large-scale experiment that potential recruits do worse on the "NLSY answer sheet" than on the military one. Comparing two groups of potential enlists, the authors found that the gaps in scores between the "NLSY answer sheet" and the military one increase with GT scores (the sum of arithmetic reasoning, word knowledge, and paragraph comprehension standardized scores).[56] However, Maier and Sims (1983) have shown that gaps in speeded tests' scores between the two populations actually decrease with GT.[57] Not surprisingly then, with the introduction of new ASVAB forms, the problems with the norming of the speeded tests got even worse (Maier and Hiatt, 1986), and resulted in the recommendation, that was accepted in 1989, to take the numerical operation test out of the AFQT.

A.1.2 Armed Forces - Variable Construction for Figure 2

The test score distribution for the 1984 male applicants for enlistment (IOT&E 1984), reported in Figure 2, was constructed using the data provided in Maier and Hiatt (1986). The authors provide, in Table A-4 pp. A9-A10, a conversion between the coding speed scores of the 1980 Youth Population (i.e., all NLSY participants who were born before 1/1/1963) and the IOT&E 1984 scores. This conversion was done by setting the raw scores of individuals from the two populations

[55] The reason the speeded tests attracted so much attention in the military was that the numerical operation test was until 1989 part of the AFQT that serves as the "entrance exam" to armed forces.

[56] If people with high GT scores work faster on average, then their speeded test scores would suffer the most if it takes them longer to record their answers on the "NLSY answer sheet".

[57] Moreover, potential recruits participating in the experiment were asked to take the speeded tests related to the research before taking the whole ASVAB (including these two tests) "for real". Therefore, they may not have the right incentives to answer questions in the research part. In addition, all individuals that displayed large change in speeded test scores between the research part and the ASVAB part were deleted from the data.

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, paragraph comprehension, and mechanical comprehension standardized scores. I used this conversion, the relative weights for each of the HST intervals for the IOT&E 1984 and the 1980 Youth Population (Maier and Hiatt, 1986, Table A-1, p. A2), and the data available in the NLSY data set to construct the distribution of raw coding speed score for the IOT&E 1984. There were two decisions to be made while reconstructing the IOT&E 1984 coding speed scores. The first Maier and Hiatt (1986) do not provide an equivalent to test scores of zero. However, 12 NLSY participants had this score. I have set the equivalent test score to zero since it will make the IOT&E 1984 population look worse. The second, Maier and Hiatt (1986) report a range for coding speed test score of two (2-10), I have again taken the lowest value (2).

Appendix B: Supplemental Regression Results

Table B1: Mean and Standard Deviation of Key Outcome Variables for Women by Cohort-Adjusted Coding Speed Test Scores[1]

Low Coding Speed Test Scores - Women with Coding Speed Test Sores Below the Mean[2]
High Coding Speed Test Scores - Women with Coding Speed Test Sores Above the Mean[2]

Variable	Low Coding Speed Scores Mean	Low Coding Speed Scores Standard Deviation	High Coding Speed Scores Mean	High Coding Speed Scores Standard Deviation	Difference	Observations
% 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 (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, and were calculated for women who where born between October 1^{st} 1961 and September 30^{th} 1964, who have completed the ASVAB and were not given "Spanish Instructions Cards" and were interviewed in 2004.
2. 45% of women had coding speed scores lower than the mean.
3. Working women in 2003 are civilians with valid ASVAB scores, who were nor enrolled in school in 2003, and reported positive earnings.

Table B2: 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 A for details).

Table B3: 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 A for details).

Appendix C - Theoretical Appendix

Proof. Proposition 1:

When performance based incentives 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 condition is: $D\equiv TS_{ee}(U_{TS}+\phi U_M)+TS_e^2\left(U_{TS,TS}+2\phi U_{TS,M}+\phi^2U_{M,M}\right)-C_{ee}$. 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

$$\frac{dTS}{dx}=TS_x+TS_e\frac{de^*}{dx} \tag{2}$$

In order to figure out how the optimal level of effort, e^*, depends on skill differentiate e^* 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 (2) and (3) 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, test scores will always increase with skill, regardless of the relations between optimal 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 e^* 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 utility is 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 intensity of the incentives ,ϕ, will result in an increase in effort, and a corresponding increase in test scores. ■

Proof. Proposition 2:

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

The first part of the proof is identical to the proof of Proposition 1. 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.

To find how the optimal level of effort, e^*, depends the type, θ, differentiate e^* with respect to θ to get $\frac{de^*}{d\theta} = -\frac{1}{D}TS_eU_{TS,\theta}$ 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 3:

I start by proving the second part. For brevity I use the following notations: $TS_1(\phi) = TS(x_1;\phi,\tilde{\theta})$, $TS_2(\phi) = TS(x_2;\phi,\tilde{\theta})$, $\underline{TS}_1(\phi) = TS(\underline{x}_1;\phi)$, and $\underline{TS}_2(\phi) = TS(\underline{x}_2;\phi)$. Denote by $\tilde{f}_i(TS_i;\phi)$ the test score distribution of group i under incentives scheme ϕ. Since $TS_1(\phi)$ first order stochastically dominates $TS_2(\phi)$ then

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

for all z. As all individuals have the same test-taking motivation, test scores provide a correct ranking according to skills and thus $\underline{TS}_1(\phi) \geq \underline{TS}_2(\phi)$ and $\overline{TS}_1(\phi) \geq \overline{TS}_2(\phi)$.

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

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

where $f_i(x_i;\phi) = \tilde{f}_i(TS_i\,(x_i;\phi))$ 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 $TS(z_2;\phi) = \widetilde{TS} = TS(z_1;\phi)$, Proposition 1 implies that $z_2 = z_1$. Moreover, it implies that $f_i(x_i;\phi) = f_i(x_i)$. Hence, $\int_{x=\underline{x}_i}^{TS_i^{-1}(x)} f_i(x_i)\frac{dTS}{dx_i}dx_i = \int_{x=\underline{x}_i}^{TS^{-1}(x)} f_i(x)\frac{dTS}{dx}dx$, where $i = 1,2$. Similarly, since $\underline{TS}_1 \geq \underline{TS}_2$ and $\overline{TS}_1 \geq \overline{TS}_2$ then $\underline{x}_1 \geq \underline{x}_2$ and $\overline{x}_1 \geq \overline{x}_2$. Thus we can rewrite (6) as

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

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

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

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

The first part can be proven by noticing that the proof above fails already at the first assertion. Proposition 2 states that $\frac{dTS}{dx}\big|_{\theta} > 0$, i.e., test scores provide a correct ranking according to skills only for individuals of the same type. Thus, the condition $\underline{TS}_1 \geq \underline{TS}_2$ does not guarantee that $\underline{x}_1 \geq \underline{x}_2$, 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 2 implies that $\underline{x}_1 < \underline{x}_2$, and similarly if $\overline{TS}_1 = \overline{TS}_2$ then $\overline{x}_1 < \overline{x}_2$. Hence, if either $\underline{x}_1 < \underline{x}_2$, or $\overline{x}_1 < \overline{x}_2$ there will be some values of x for which $\int_{x=\underline{x}_2}^{x} f(x,\theta_1)dx > \int_{x=\underline{x}_1}^{x} f(x,\theta_2)dx$. Thus, it is not true that $x(\theta_1)$ first order stochastically dominates $x(\theta_2)$.

Therefore, without making more assumptions on the skill distributions of the two types, which are what we want to recover, even in the case where $TS\left(x;\phi,\theta_1\right)$ first order stochastically dominates $TS\left(x;\phi,\theta_2\right)$ we cannot guarantee that $x\left(\theta_1\right)$ first order stochastically dominates $x\left(\theta_2\right)$.[58] ■

Appendix D - Experimental Appendix

D.1 Instructions

D.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

[58] Note also that in additions to further assumptions regarding the support of the skill distributions of the two types, we would need to assume that $\frac{d^2TS}{dxd\theta} \geq 0$ in order to get that stochastic dominance in test scores imply stochastic dominance in skills levels.

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?

D.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(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 2(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?

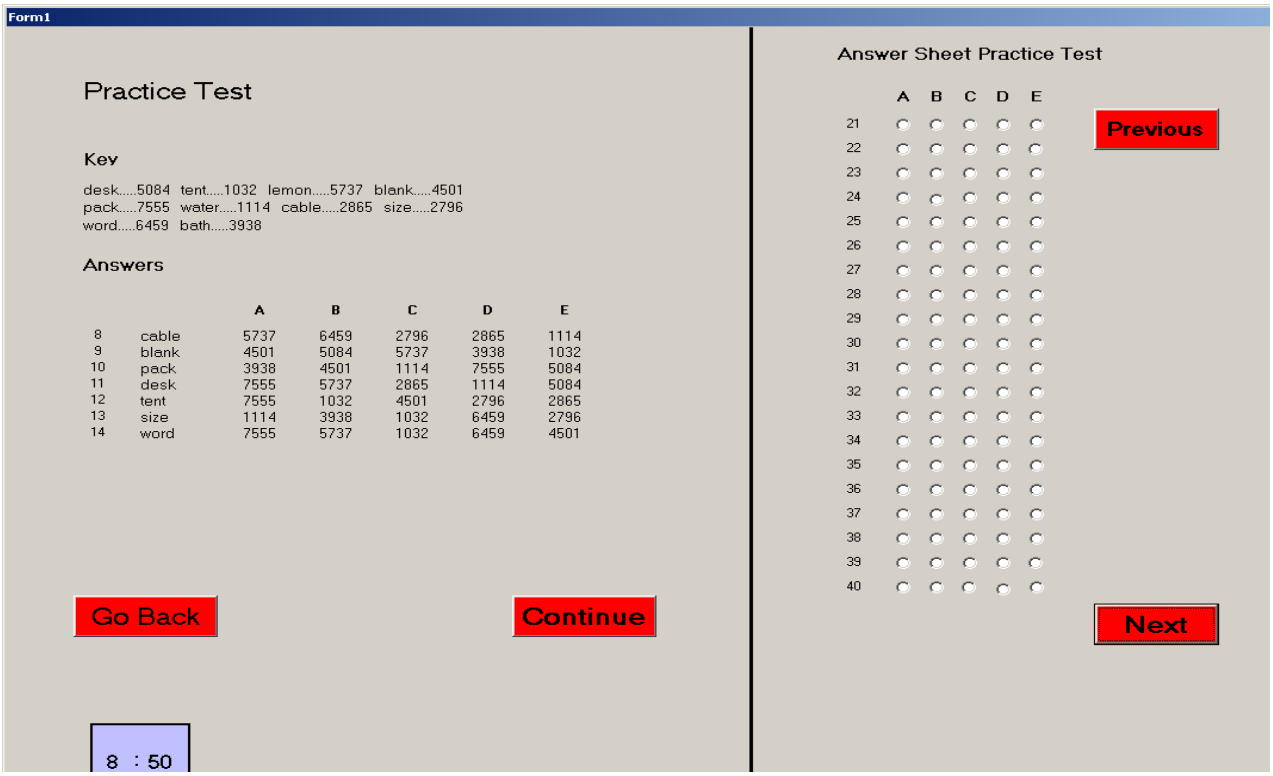


Figure D1: Snapshot of Testing Program's Screen

Table D1: 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)
Period2		-0.002		0.0003		-0.0001
		(0.001)*		(0.001)		(0.001)
Observations	1863	1863	1784	1784	1767	1767
R-Squared	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.

D.2 Does the Relative Ranking Change?

To further provide evidence on the change in ranks, I examine directly the change in ranks according to total test scores. To do so, I assign within each test a higher rank to participants with lower test scores. Thus, participants with the highest test scores are assigned rank 1; participants with the second to highest scores are assigned rank 2, and so on. This is a very conservative measure of rank change. In particular, if participant's scores improved between any two tests and now she is one of the participants with the highest scores, only her own rank will be changing. As it is not

clear whether small differences in test scores indicate to real differences in ability, I examine the percentage of participants who have changed their ranks above a certain threshold. In addition, I inspect the distributions of the absolute rank change to get a sense of the magnitudes.

Table D2 reports the percentage of participants who experienced an absolute rank change bigger than 4 (i.e., a change that move them outside a decile of ranks centered around their own rank) and the mean and the maximum rank change. The Table indicates that at least 53% of participants changed their absolute relative rank by more than 4 ranks. In addition, the average participant experienced an absolute rank change bigger than 6 ranks between any two tests.

Next, I investigate the different sources of noise that may create rank changes. There are three obvious such sources that usually play a role in testing: having a good/bad day, getting "lucky" with some test-questions, and guessing. Given the experimental design and the coding speed test, it seems that only guessing is likely to create noise that may lead to rank changes. Specifically, the experiment lasted for less than hour. Therefore, participants who were having a good or a bad day would have had it throughout the experiment. In tests that measure knowledge, a possible source of noise is the specific questions asked. In particular, test-takers may get higher scores on a particular test since some of its questions relate to a firsthand knowledge they have.[59] However, as all the coding speed questions require identical knowledge (recognize words and numbers), this source cannot operate here. Thus, the only source left is guessing. To create a measure of rank change taking away the random component introduced by guessing, I use the average number of questions correctly solved in the 30-second periods before the first guess.[60] To make the resulting scores comparable to the original ones, I assume that participants, had they not been guessing, would have had this average performance throughout the test, i.e., for twenty 30-second periods. This assumption means that test-taking motivation has no effect on subjects' boredom and fatigue as the test progresses. However, psychologists investigating motivation claim that lack of motivation causes boredom and fatigue that increase in the task's length (Revelle, 1993). To further ensure compatibility of the resulting scores, I truncate the test scores at 140 and allow them to take only integer values. Ranks are assigned given this constructed set of total test scores as before.

Table D2 reports the percentage of participants who experienced an absolute rank change bigger than 4 and its mean and maximum. Even once the guesses are taken out at least 48.5% of participants changed their absolute relative ranking by more than 4 ranks. The average participant experienced an absolute rank change bigger than 5 between any two tests. The distributions of the absolute rank change reported in the two panels are not statistically different (Mann-Whitney tests yield $p = 0.33$ for the absolute rank change between the \$10 and incentives tests and $p = 0.97$ for the absolute rank change between the practice and \$10 tests). Thus, the change in ranks between

[59] For example, if a test taker spent the summer in England, she may know that London is its capital, even though she may not be able answer any other question regarding capitals of other countries.

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

the different tests does stem from random error associated with guessing.

Next I use simulations to investigate whether the rank change reported above stem from another source of noise. Specifically, I simulate, for a given test, potential total scores and relative ranking for each participant, and investigate whether noise can explain changes in the simulated ranks between tests. Here some cautious is warranted. First, the division of the tests to 30-second periods itself creates noise. Even if all participants solved the test at a constant pace, this pace does not necessarily coincide with the 30-second periods. Therefore, to reduce this measurement error, I examine 1-minute and 2-minute periods. Second, we want to distinguish possible noise from lack of test-taking motivation. The latter may explain rank change within the first two tests. However, lack of test-taking motivation should not explain rank change within the incentives test.[61] Therefore, I restrict attention to rank change implied by participants' actions in the incentives test, to test whether the rank change 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 scores in the incentives test) this can serve as an upper bound for the effects of noise.

To simulate the rank change within the incentives test, for each participant I re-sample twice from the periods before her first guess. 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 two sets of relative ranking, where the highest scores are ranked 1, the second highest test scores are ranked 2, etc, and compute the absolute rank change between these 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 times. Figure D2 depicts the CDF's of absolute rank change obtained from these simulations. It is clear from the figures that the amount of rank change between the tests is much higher than the amount of rank changes within the incentives test. To quantify this, I calculate the fraction of absolute rank change that is bigger than x, where x = 1,2..., for each simulation, and calculate the percentage of simulations in which this fraction was bigger within the incentives test than between any two tests. For the 1-minute periods simulations, the fraction of absolute rank change bigger than one is bigger within the incentives test than between the tests in at most 6% of simulations. The percentage is reduced to less than 1% when looking at fraction of absolute rank changes bigger than three. For 2-minute periods the simulations, this fraction is less than 1% already when looking at absolute rank changes bigger or equal to one. While the simulations suggest that some rank change occurs 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 change between the tests cannot be attributed to noise.

[61] Although, one needs to keep in mind that if participants started guessing after periods in which their pace fell, not all time periods are interchangeable. This may create additional rank change even within the incentives test.

Table D2: Change in Relative Ranking Between Tests Using Total Test Scores

Panel A: Using Total Test Scores[2]

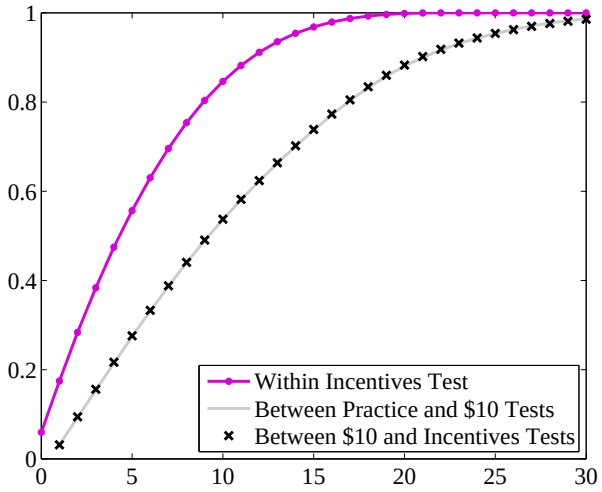
	% Participant that Changed Ranking by more than | 4 |	Mean Absolute Rank Change	Maximum Absolute Rank Change	Observations
Between Incentives Test and				
$10 Test	54.5	6.04	25	99
Practice Test	53.5	6.38	32	99
Between $10 Test and				
Practice Test	58.6	6.9	31	99

Panel B: Using Test Scores based on Participants' Average Performance Before they Started Guessing[3]

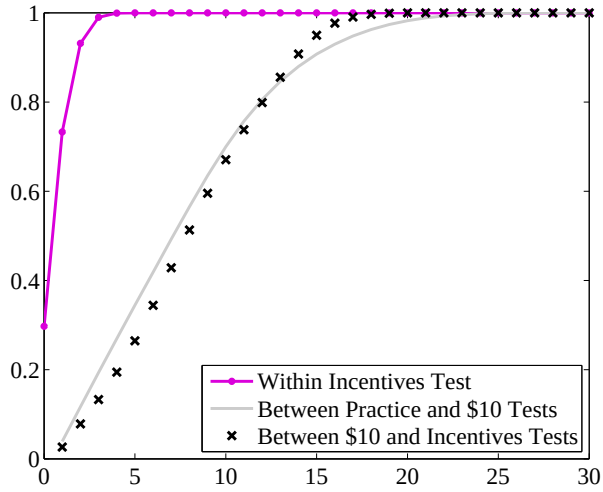
	% Participant that Changed Ranking by more than | 4 |	Mean Absolute Rank Change	Maximum Absolute Rank Change	Observations
Between Incentives Test and				
$10 Test	48.5	5.45	21	97
Practice Test	62.2	6.62	28	98
Between $10 Test and				
Practice Test	62.2	6.66	28	97

Notes:

1. 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.
2. 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.
3. The test scores were constructed using the 30-second periods before participants' first guess, see text for details. The maximum ranks are 46 for the practice test, 53 for the $10 test, and 49 for the incentives test.



CDFs of Absolute Rank Change: 1-Minute Periods



CDFs of Absolute Rank Change: 2-Minute Periods

Figure D2: CDFs of Absolute Rank Change Simulation Results (see text for details).