This article was downloaded by: [155.246.103.35] On: 06 April 2017, At: 14:56
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Working When No One Is Watching: Motivation, Test Scores, and Economic Success

Carmit Segal,

To cite this article:
Carmit Segal, (2012) Working When No One Is Watching: Motivation, Test Scores, and Economic Success. Management Science 58(8):1438-1457. http://dx.doi.org/10.1287/mnsc.1110.1509

Full terms and conditions of use: http://pubsonline.informs.org/page/terms-and-conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2012, INFORMS

Please scroll down for article—it is on subsequent pages

INFORMS is the largest professional society in the world for professionals in the fields of operations research, management science, and analytics.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

RIGHTSLINK

MANAGEMENT SCIENCE
Vol. 58, No. 8, August 2012, pp. 1438–1457
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1110.1509
© 2012 INFORMS

Working When No One Is Watching: Motivation, Test Scores, and Economic Success

Carmit Segal
Department of Economics, University of Zurich, 8006 Zurich, Switzerland, carmit.segal@econ.uzh.ch

This paper provides evidence that scores on simple, low-stakes tests are associated with future economic success because the scores also reflect test takers' personality traits associated with their level of intrinsic motivation. To establish this, I use the coding speed test that was administered without incentives to participants in the National Longitudinal Survey of Youth (NLSY). I show that, controlling for cognitive ability, the coding speed scores are correlated with future earnings of male NLSY participants. I provide evidence that the coding speed scores relate to intrinsic motivation. I show that the scores of the highly motivated, though less educated, group (potential recruits to the U.S. military), are higher than the NLSY participants' scores. I use controlled experiments to show directly that intrinsic motivation is an important component of the unincentivized coding speed scores and that it relates to test takers' personality traits.

Key words: organizational studies; motivation–incentives; behavior; labor; economics; utility preference; applications
History: Received July 15, 2010; accepted October 15, 2011, by Teck Ho, behavioral economics. Published online in *Articles in Advance* April 18, 2012.

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

1. Introduction

Classical economic theory predicts that if an action, like taking a standardized test, requires costly effort, then without performance-based incentives individuals will invest the lowest effort possible in performing the action. Therefore, one might expect that scores on low-stakes tests (tests administered without performance-based incentives) would be uninformative about an individual's cognitive ability, and thus that low-stakes tests scores would be uncorrelated with test takers' future economic outcomes. However, nearly all researchers find strong positive correlations between high scores on unincentivized tests and future labor market success. This paper suggests that unincentivized scores capture not only cognitive ability, but also intrinsic motivation. Moreover, it suggests that intrinsic motivation is associated with favorable personality traits, such as conscientiousness, which lead test takers to put effort into low-stakes tests and to future labor market success.[1]

To investigate why unincentivized test scores relate to economic success, ideally one would like to have both low- and high-stakes scores for each individual for a given test and data on outcomes. This paper suggests that if such data were available, both low- and high-stakes scores would be associated with favorable economic outcomes: low-stakes scores primarily because they relate to personality traits associated with intrinsic motivation, and high-stakes scores because they relate to cognitive skills. The comparison between individual rankings according to their low- and high-stakes scores could provide direct evidence on the relationships between intrinsic motivation and unincentivized test scores to support the explanation above. However, to the best of my knowledge, no such data exist. Instead, I investigate the relationship between intrinsic motivation and economic success using two data sets and a laboratory experiment. Specifically, I use the National Longitudinal Survey of Youth (NLSY) to establish correlations between test scores and outcomes, data from the U.S. military to limit potential interpretation of these correlations and to argue that the intrinsic motivation explanation is possible, and experimental data to provide direct evidence on the role that intrinsic motivation plays in the unincentivized test scores.

Because the ideal data are not available, selecting a proper test is crucial. All low-stakes test scores may be affected by the individual's level of intrinsic motivation, but the effect may be larger for tests that do not require specialized knowledge. To ensure the positive impact of monetary incentives, one should also

[1] I make a distinction between "intrinsic motivation," which is a trait of the individual, and "effort," which is a choice variable that can be affected by intrinsic motivation and cognitive ability. Thus, when performance-based incentives are provided, individuals invest high effort or are highly motivated to take the test, though they may have low intrinsic motivation to do so. When I refer to the personality trait, I always use the term intrinsic motivation.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

use simple tasks (Ariely et al. 2009). The coding speed test, which was part of the Armed Service Vocational Aptitude Battery (ASVAB), fulfills these requirements. The coding speed test asks participants to match words with four-digit numbers (see an example in Figure 1 in §4). To find out which word matches to which number, test takers need to look at the key, in which the association between each four-digit number and each word is given. The prior knowledge necessary to answer the coding speed test is minimal, so it is likely that level of effort is the main contributor to high scores. Still, the time allotted to the test is short, so its scores to some degree also measure cognitive ability related to mental speed, possibly fluid intelligence as suggested by Heckman (1995) and Cawley et al. (1997).[2]

The ASVAB was administered, for research purposes, without performance-based incentives to the NLSY participants. The NLSY data show that male participants' unincentivized coding speed scores are significantly and positively associated with earnings 23 years after taking the test. This implies that the unincentivized coding speed scores measure a trait (or traits) highly valued in the labor market, but it does not reveal what the trait(s) may be.

Based on military research, I argue that the unincentivized coding speed scores cannot only measure cognitive ability. The ASVAB is the screening and job assignment exam used by the U.S. military. As such, it is a high-stakes test for potential enlistees. Therefore, their coding speed scores should be an excellent measure of any cognitive ability measured by the test. If the cognitive component of the coding speed scores is the one causing the associations between the scores and economic success, we would expect to find a positive relationship between the coding speed scores and success in the military. Utilizing *hundreds of military studies*, Hunter (1986) reported that whereas the two speeded ASVAB test scores predict military performance, after controlling for the nonspeeded ASVAB scores, the speeded test scores are no longer associated with military performance.[3] This implies that the cognitive ability measured by the coding speed test is important for military performance, but that the nonspeeded ASVAB tests measure it better.[4]

[2] Using factor analysis, these papers show that the two ASVAB speeded tests, the coding speed and the numerical operation (NO), belong to a different factor than the other ASVAB tests, and that this factor is highly correlated with earnings.

[3] The only exception to this pattern is clerical jobs.

[4] Moreover, it may explain why the military dropped the speeded tests from the ASVAB in 2002, even though computerized testing allowed it to more accurately assess potential recruits' mental speed (Wolfe et al. 2006), and why it hardly used the coding speed scores even before 2002.

The findings reported by Hunter (1986) suggest that one or more of the eight nonspeeded ASVAB tests can serve as a measure for the cognitive ability component of the coding speed scores. Therefore, if after controlling for the nonspeeded ASVAB test scores the coding speed scores are still correlated with earnings in the NLSY sample, the reason cannot be that they measure cognitive ability. In regressions in which I control for the eight nonspeeded ASVAB test scores, I find, in the NLSY sample, associations between the coding speed scores and earnings that are large and highly significant. This suggest that in these regressions the associations between the coding speed scores and earnings do not stem from cognitive ability, but should be attributed to personality traits associated with intrinsic motivation.

The rest of this paper provides evidence that indeed intrinsic motivation is an important determinant of the unincentivized coding speed scores. I use military data to provide indirect field evidence. I show that the NLSY participants, though more educated, scored worse on the coding speed test than potential recruits. That result is to be expected if indeed motivation is important for the coding speed test scores.

I then present the results of controlled experiments in which motivation was induced through monetary incentives. The experimental results provide direct evidence on the roles that levels of effort and intrinsic motivation play in the unincentivized coding speed scores. The model described in §2 implies that if intrinsic motivation varies across individuals, then their ranking according to their unincentivized scores may differ from their ranking according to their incentivized scores. To investigate whether this happens, participants in the experiment took the coding speed test three times: twice for a fixed payment, where the first version was called a "practice" test and the second "the" test, and then a third time with performance-based monetary incentives.

Participants in the experiment responded differently to the lack of performance-based incentives. Sixty-two percent of participants did not improve their performance with provision of such incentives, but the performance of 38% of participants improved significantly. When no incentives were provided, the test score distribution of the first group first-order stochastically dominated the test score distribution of the second one. However, participants from both groups have an equal distribution of cognitive ability, as measured by their SAT scores and their incentivized coding speed scores. Together with the evidence regarding participants' guessing behavior, this suggests that this second group of participants needed incentives to try their best to solve the test, whereas the first group did so from the moment the test was called "the" test. I estimate that the variation in intrinsic

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

motivation can account for approximately 43% of the variation in the unincentivized coding speed scores. Using participants' answers to a psychological survey, I find that for males, the intrinsically motivated group (i.e., the group that worked hard irrespective of incentives provided) had higher conscientiousness than the intrinsically unmotivated group (i.e., the group that needed incentives to work hard). In addition, women were more likely to be intrinsically motivated (i.e., to belong to the group that worked hard irrespective of incentives provided).

The experimental results reinforce the following explanation of the evidence from the NSLY and the military. Potential recruits are doing better than NLSY participants on the coding speed test because they are more motivated to take it. The military is not using the coding speed test because in their incentivized form its scores no longer capture the personality traits associated with intrinsic motivation, and it has a better measure for the cognitive ability component of the test. At the same time, for the NLSY participants, the coding speed test is a low-stakes test, so higher scores also indicate favorable (in the labor market) personality traits. As a result, the coding speed scores are correlated with earnings even after controlling for cognitive ability measured by the nonspeeded ASVAB test scores and by level of education.

The relationship between motivation and test scores has been investigated before. Starting in the 1900s, psychologists have shown that level of motivation affects test scores and may be related to personality traits (for an excellent summary, see Revelle 1993 and citations therein; for current research, see Duckworth et al. 2009). In economics, the evidence obtained through lab and field experiments clearly indicates that test scores are positively related to (high enough) incentives (see, for example, Gneezy and Rustichini 2000, Angrist and Lavy 2009, Kremer et al. 2009). Recently, several studies in economics have shown that test scores correlate with personality traits and preferences parameters (Benjamin et al. 2005; Borghans et al. 2008, 2009; Dohmen et al. 2010; Heckman et al. 2009). However, as is the case in the psychology literature, these papers do not examine whether the most motivated test takers are the most cognitively able ones. Thus, these papers cannot provide any empirical evidence regarding sources other than cognitive skills that may create relationship between unincentivized test scores and outcomes. This is the first paper showing that, at least for the coding speed test, low scores on unincentivized tests do not necessarily imply low cognitive ability. Instead, they also imply unfavorable personality traits. As such, this is the first paper that provides empirical evidence suggesting that the relationship between unincentivized test scores and economic success is not solely due to cognitive skills.

This paper also relates to the literature investigating the validity of the basic premises of agency theory. Those premises suggest that individuals invest little effort unless provided with proper incentives or monitored. The literature suggests that economic theory can predict the behavior of a nonnegligible fraction of individuals. For example, Nagin et al. (2005) showed that approximately 40% of employees in a calling center shirked when they inferred that they were not being monitored. Fehr and Falk (1999) showed experimentally that in response to higher flat wages, approximately 25% of participants always provided minimal effort, whereas the rest responded by choosing higher effort. This paper demonstrates that these insights are present in testing situations too.[5] Furthermore, it shows that heterogeneous responses to the lack of performance-based incentives are not driven by differences in abilities.

Last, this paper contributes to the literature relating cognitive and noncognitive skills to earnings (see, for example, Bowles et al. 2001, Heckman and Rubinstein 2001, Persico et al. 2004, Kuhn and Weinberger 2005, Segal 2012, Heckman et al. 2006). Rather than looking for a proxy for noncognitive skills, I focus on the noncognitive 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., noncognitive skills, to affect test scores. Whereas the NLSY data provide only suggestive evidence on the relationship between intrinsic motivation and personality traits, the experimental results provide direct evidence of a relationship between the two.

2. The Model

In this section, I model how individual differences in intrinsic motivation may affect test scores. I model the case in which individuals differ in both their cognitive ability and their intrinsic motivation, thus making the case of no differences in intrinsic motivation a special case. The goal is to understand how intrinsic motivation affects test scores. This will allow the derivation of testable predictions that can detect intrinsic motivation in the experiment, if it exists.

Agents differ from one another in their cognitive skills, denoted by x, that a given test is supposed to

[5] 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 from zero (when incentives were high) and toward zero (when incentives were low). Recently, Borghans et al. (2008) showed that when given IQ questions, some participants respond to incentives mainly by investing more time in answering questions, whereas others do not. The authors related this response to incentives to personality traits.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

measure. The random variable x has a density $f(x)$. In the case of the coding speed test, this cognitive ability has to do with speed and may even be fluid intelligence.

Test scores are being produced using two inputs: cognitive skills and effort, denoted by e.[6] The production function of test scores is given by $TS(x, e)$, where test scores are increasing in cognitive skills and effort, i.e., $TS_e > 0$ and $TS_x > 0$. I assume further that $TS_{ex} > 0$, i.e., a given increase in effort results in higher test scores for agents with higher cognitive skills. Producing test scores is costly. The costs associated with effort, $C(e)$, are increasing and convex in effort, i.e., $C_e > 0$ and $C_{ee} > 0$. In this model, investing effort means concentrating or working fast. Thus, I assume that the costs of effort do not depend on the cognitive skills of the individuals.[7]

Because effort is costly, if there are no benefits associated with higher test scores, agents will invest the minimal effort, i.e., will solve no question on the exam. However, most survey participants get scores much higher than zero even without performance-based incentives. This suggests that agents obtain psychological benefits from having higher test scores. To capture the possibility that individuals may differ in their psychological gains from having the same test scores, I add types to the model. Thus, agents are of different types, denoted by θ, where agents with lower values of θ gain fewer psychological benefits from test scores. Thus, if we denote the agents' benefits by U, then $U_{TS,\theta} > 0$. I assume that the production and cost functions do not depend on θ.

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 only psychological. Below, I focus on the provision of piece-rate monetary incentives, because this is the relevant situation in the experiment.[8] In this case, agents' benefits also include their monetary gains from having higher test scores, which are 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. The type θ does not affect agents' benefits from money, i.e., $U_{M,\theta} = 0$. Thus, I model agents' benefits by $U(TS, M; \phi, \theta)$, where $U_{TS} > 0$, $U_M > 0$, i.e., agents like to have more money and higher (psychological benefits from) 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., $d^2U/dTS^2 = (U_{TS,TS} + 2\phi U_{TS,M} + \phi^2 U_{M,M}) \leq 0$ (this condition is fulfilled if, for example, agents' benefits are separable in money and test scores). As usual, an agent with cognitive ability x chooses an effort level, e, to maximize benefits minus costs.

Proposition 1. *If agents obtain psychological benefits from higher test scores and/or monetary performance-based incentives are provided, then, conditional on θ, the resulting test scores provide a correct ranking according to agents' cognitive skills. If the marginal benefits of money are increasing in ϕ, holding θ fixed, an increase in ϕ results in higher effort and higher test scores. Holding cognitive ability fixed, agents with higher values of θ invest more effort and have higher test scores.*

The proof is in Appendix A.

Proposition 1 suggests that conditional on θ, or if all agents have the same intrinsic motivation, test scores provide a correct ranking of agents according to their cognitive skills. The intuition is simple. Because test scores are produced using effort and cognitive skills, agents with higher cognitive skills have higher test scores for a given level of effort. As a result, they obtain higher psychological (and if $\phi > 0$, also higher monetary) benefits. However, if agents have heterogeneous levels of intrinsic motivation, then test scores do not provide a correct ranking according to their cognitive skills.[9] The intuition is as follows. Agents with lower values of θ choose to invest less effort, because they have lower marginal benefits from higher test scores and the same marginal costs. Because test scores are produced using both cognitive ability 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 provides only limited information regarding their relative cognitive skills. A possible way to recover the rank according to cognitive ability 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 2. *Denote the cognitive skills 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 $\underline{x}_i \leq x_i(\theta_i) \leq \bar{x}_i$, 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*

[6] For simplicity I assume that e is not bounded. All the main results go through if $e \in [\underline{e}, \bar{e}]$.

[7] If the costs of effort depend on cognitive skills, then it is natural to assume that $C_x < 0$, and $C_{xe} \leq 0$, which will not change the results below.

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

[9] In two cases, test scores will provide a correct ranking of individuals: (1) if the test score distributions of different types do not overlap and (2) if $e \in [\underline{e}, \bar{e}]$ and *all* agents choose the same (corner) effort level (i.e., either $e = \underline{e}$ or $e = \bar{e}$). An agent will choose the highest or lowest effort level if the first-order conditions are not fulfilled because the marginal costs are too low or too high, respectively.

RIGHTSLINK

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 in Appendix A.

Proposition 2 implies that even if we find two groups such that the unincentivized test score distribution of one group first-order stochastically dominates the other, this may not be the case when incentives are provided. If individuals differ in their intrinsic motivation, it is possible that the group with low values of θ (i.e., low intrinsic motivation) may have the same (or even higher) cognitive skills as the group with high values of θ.

The propositions suggest how to investigate whether individuals differ in their intrinsic motivation and not only in their cognitive skills. Proposition 1 suggests that in this case the relative ranking of individuals according to their test scores may change with the provision of incentives. Proposition 2 suggests that first-order stochastic domination of the test score distribution of one group over another may change with the provision of incentives. Moreover, Proposition 1 shows that under some conditions (for example, if $U(TS, M) = \tilde{U}(TS) + M$), the provision of incentives could result in an increase in effort, and thus an increase in test scores. This could be the case even if all individuals value test scores in the same manner (i.e., have the same θ), as long as effort is bounded and agents do not invest the highest effort possible. Therefore, the fact that test scores increase with incentives does not prove that individuals differ in their intrinsic motivation. To prove this, one needs to show a change in rankings, a change in stochastic dominance, or that a fraction of individuals invested the highest level of effort in the unincentivized test.

3. Data

The analysis in §5 relies on the NLSY, a nationally representative sample of over 12,000 individuals who were first surveyed in 1979, when they were between the ages of 14 and 22, and then resurveyed annually until 1994 and biennially after that. For this paper, this source is exceptional in combining detailed labor market data with a battery of tests, which is also administered to a nonsurvey population. Because the NLSY is a well-known survey, this section focuses on aspects particular to this paper, namely, the tests administered to participants in the NLSY. Because of its main role in the analysis, the coding speed test is described in §4. Details regarding sample restriction and variable construction can be found in Appendix B. The military data are described in §6 and in Appendix B. The experimental data are described in §7.

3.1. The ASVAB

The ASVAB is a battery of 10 tests, described in Table B.1 in Appendix B. It contains two speeded tests (coding speed and numerical operations) and eight nonspeeded ones.[10] It is the screening and sorting exam for the U.S. military. The U.S. Department of Defense (DOD) had to establish a national norm for the ASVAB, so it had to be administered to a representative sample of Americans. The DOD and the U.S. Department of Labor decided to use the NLSY sample for this purpose. The ASVAB was given to the NLSY participants between June and October of 1980. Participants in the NLSY were paid \$50 for completing the test, but no direct performance-based incentives were provided.[11] Thus, for the NLSY participants, it is a low-stakes test.

4. The Coding Speed Test

Figure 1 gives the instructions and an example of the questions asked in the coding speed test. The task in this 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. The NLSY participants took a paper and pencil version of the test that lasts for seven minutes and consists of 84 questions.

Ideally, to investigate whether test takers differ in their intrinsic 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 likely that everyone who knows how to read has the knowledge to correctly answer questions on the test. Therefore, because of its simplicity, intrinsic motivation may play

[10] The scores of four of the nonspeeded tests (word knowledge, paragraph comprehension, arithmetic reasoning, and mathematics knowledge) are added to create the Armed Forces Qualification Test (AFQT). The AFQT is the most commonly used test in studies using the NLSY data (see, for example, Herrnstein and Murray 1994, Heckman 1995, Neal and Johnson 1996, Johnson and Neal 1998).

[11] "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. . . ." (U.S. Department of Defense 1982, p. 12). Some indirect incentives may have been provided by promising participants that at a future date they would get their own test scores, which might help them to make plans for their future.

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

RIGHTSLINK

Figure 1 The Coding Speed Test—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

a large role in determining its unincentivized scores.[12] Nevertheless, because the time allotted to the test is short, it is possible that not all (highly motivated) test takers are able to achieve a perfect score. Thus, the coding speed scores may also measure cognitive ability related to mental speed that may be related to fluid intelligence—novel problem solving ability—as suggested by Heckman (1995) and Cawley et al. (1997).

5. Evidence from the NLSY: Coding Speed Scores and Earnings

In this section, I present evidence that the coding speed scores of the NLSY participants are correlated with their earnings. I restrict attention to men, as a full treatment of the selection problem associated with female earnings is beyond the scope of this paper.[13] As far as the coding speed scores are concerned, the results for females are very similar to the results for males.

The model estimated in this section is of the form $\ln(earnings)_i = \beta + \beta_{CS} CS_i + \beta_{TS} TS_i + \beta_X X_i + \varepsilon_i$, where i indexes individuals, *earnings* are earnings in 2003, CS is the coding speed scores, TS is the scores on the nonspeeded ASVAB tests,[14] X denotes individual characteristics, and ε is an error term.

Table 1 presents the earnings 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 the coding speed scores. The coding speed scores are highly correlated with earnings. One standard deviation increase in the coding speed scores corresponds to an increase of 27.8% in earnings. This implies that the coding speed scores measure traits valued in the labor market. However, it does not tell us what they may be.

To figure out whether the coding speed scores are correlated with earnings only because they measure cognitive ability (of any kind), a measure of the relevant cognitive ability is needed. To find the appropriate measure, I take advantage of the fact that the ASVAB is the screening and job assignment exam used by the U.S. military. Thus, potential recruits are likely to be highly motivated to take it, so their coding speed scores should provide mainly an estimate of the cognitive ability measured by the test. Using meta-analysis over *hundreds of military studies*, Hunter (1986) found that the two speeded ASVAB tests predict military performance.[15] However, Hunter (1986) also reported

[12] The ASVAB contains another test that may seem appropriate to use: the numerical operation test, which consists of 50 simple algebraic questions (e.g., $2+2=?$, $16/8=?$, etc.) and lasts three minutes. Although it seems that everyone can answer these questions, a possible concern is that individuals with high math skills may invest higher effort in solving the NO test than individuals with low math skills. Thus, its scores may include a larger knowledge component than the content of the questions might suggest. Psychologists suggest that the effects of lack of motivation are more pronounced the longer the task is (see Revelle 1993). The coding speed test is more than twice as long as the NO test, so the effect of lack of motivation would therefore be easier to detect on the coding speed test. In addition, whereas 16% of NLSY participants correctly solved at least 90% of the NO questions, the corresponding number for the coding speed test is 1%. Thus, the coding speed test may serve as a better measure because its range of scores is larger.

[13] Several papers had cautioned against inferences made from female earnings regressions to offered wages because of severe selection problems. Specifically relevant is the paper by Mulligan and Rubinstein (2008), who suggested that selection is an important determinant in female wages and that the relationships between cognitive skills and wages for women are nontrivial.

[14] All test score variables have been adjusted for school-year cohort; see Appendix B for details.

[15] Wolfe et al. (2006) provided evidence that supports the analysis of Hunter (1986) for the computerized version of the ASVAB and for more recent performance data.

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

Table 1 Earnings of Men (Dependent Variable: ln of *Earnings 2003*)

	(1)	(2)	(3)	(4)
Black	−0.555 [0.066]***	−0.378 [0.066]***	−0.218 [0.073]***	−0.256 [0.072]***
Hispanic	−0.331 [0.076]***	−0.245 [0.071]***	−0.12 [0.074]	−0.125 [0.072]*
Coding speed		0.245 [0.026]***	0.091 [0.028]***	0.064 [0.027]**
Arithmetic reasoning			0.068 [0.045]	0.048 [0.043]
Mathematics knowledge			0.153 [0.051]***	0.05 [0.054]
Word knowledge			−0.004 [0.055]	−0.054 [0.053]
Paragraph comprehension			0.051 [0.049]	0.033 [0.048]
General science			−0.007 [0.042]	−0.053 [0.043]
Auto and shop information			0.005 [0.037]	0.06 [0.036]*
Mechanical comprehension			−0.003 [0.047]	0.008 [0.046]
Electronic information			0.05 [0.046]	0.064 [0.045]
Years of schooling completed (*2003*)				0.112 [0.015]***
Age in 2003	0.026 [0.031]	0.02 [0.030]	0.014 [0.030]	0.025 [0.029]
Constant	9.681 [1.257]***	9.892 [1.221]***	10.102 [1.199]***	8.148 [1.241]***
Observations	1,187	1,187	1,187	1,187
R^2	0.05	0.12	0.19	0.24

Notes. The sample includes men who where born between October 1, 1961, and September 30, 1964, who completed the ASVAB test, were not given "Spanish Instructions Cards," and did not belong to the oversample. 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. All test score variables are school-year-cohort adjusted (see Appendix B for details). The coefficients on the nonspeeded tests are jointly different from zero ($p < 0.01$). Robust standard errors are in brackets.

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

that after controlling for the eight nonspeeded ASVAB scores, the two speeded tests are no longer correlated with military performance. The only exception to this pattern is clerical jobs, and even for these jobs, the contribution of the speeded tests is small (Hunter 1986, pp. 341–342, 358). This evidence suggests that the cognitive component of the speeded tests predicts military performance. However, the nonspeeded ASVAB tests measure this very same cognitive component better. Thus, the findings of Hunter (1986) imply that the nonspeeded ASVAB test scores can serve as controls for the cognitive ability component of the coding speed scores.[16]

Therefore, in column (3), I add the eight nonspeeded ASVAB test scores to the regression of column (2). The inclusion of the nonspeeded ASVAB scores reduces the correlations between the coding speed scores and earnings of the NLSY participants. Nevertheless, the associations are still economically large and statistically significant. Thus, one standard deviation increase in the coding speed scores is associated with an increase of 9.5% in earnings. Column (4) adds years of schooling completed to the regression. The association between the coding speed scores and earnings is insignificantly reduced, and it is still economically large and statistically significant. One standard deviation increase in the coding speed scores is associated with an increase of 6.6% in earnings.[17]

The findings of Hunter (1986) suggest that the nonspeeded ASVAB scores can serve as controls for the cognitive component of the coding speed scores. Therefore, the regression results of columns (3) and (4) imply that the relationship between the unincentivized coding speed scores and earnings does not stem from any cognitive ability.[18] I suggest that because the NLSY participants' coding speed scores are unincentivized ones, they proxy for personality

[16] These findings may explain why the military used the coding speed scores very sparingly: The scores were not used to determine who is eligible to enlist or to become an officer, and of the 30 different composite tests used to screen soldiers into jobs, only seven used the coding speed scores; of those, five were used to screen for clerical positions (Segall 2006) (the other speeded test (NO) was used only five times, three of which were together with the coding speed test). Moreover, it may explain why since 2002 the speeded tests have no longer been part of the ASVAB even though computerized testing allowed the military to more accurately assess potential recruits' speed (Wolfe et al. 2006).

[17] Based on Hunter (1986) I report results using the eight nonspeeded ASVAB scores as controls for cognitive ability. The results are very similar when the AFQT is used instead. Specifically, I find that after controlling for the AFQT scores and education, one standard deviation increase in the coding speed scores is associated with an increase of 6.6% in earnings. Moreover, when I allow the test scores variables to vary between those who are college graduates and those who are not, I find that the association between the AFQT scores and earnings is larger for highly educated workers (F-test for the equality of the two coefficients yields $p = 0.026$). However, this is not the case for the coding scores; 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 regression also indicates that whereas for college graduates the AFQT scores are significantly more important to earnings than the coding speed scores, for less-educated workers the two scores are equally important (F-tests for the equality of the coefficients yield $p = 0.038$ and $p = 0.669$, respectively). Previous research (see Segal 2012, Heckman et al. 2006) suggests that in comparison to cognitive skills, noncognitive skills are relatively less important to the earnings of highly educated people. These results are reported in Table C.1 in Appendix C.

[18] These regressions also rule out the possibility that the coding speed scores correlates with earnings only because they measure reading ability, because the reading comprehension test is used to assess literacy (Maier and Sims 1986).

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

traits that relate to intrinsic motivation (once cognitive ability is controlled for in the form of the nonspeeded tests).[19] In the next sections, I provide indirect and direct evidence that indeed the coding speed scores relate to intrinsic motivation.

6. Indirect Evidence from the Armed Forces

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 populations, everything else being equal. Moreover, if being motivated is particularly important for the coding speed scores, then the effect should be more pronounced for this test. I test this hypothesis by comparing the scores of the NLSY participants and potential recruits, who are highly motivated to do well on the ASVAB. Potential recruits are a less educated and more racially diverse population than the NLSY participants (Maier and Sims 1983), so 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 problems in comparing the ASVAB scores between potential recruits to the armed forces and NLSY participants of comparable ages (i.e., born before January 1, 1963). Specifically, Maier and Sims (1983) showed that whereas potential recruits scored higher on the speeded tests (i.e., the coding speed and NO tests) than the NLSY participants, they did worse on every other test.[20] The latter part was expected because the NLSY participants were more educated than potential recruits (Maier and Hiatt 1986). The former part was not. Maier and Hiatt (1986) suggested 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 is not known. What is known is that the 1980 Youth Population generally did not know or follow these strategies . . ." (Maier and Hiatt 1986, pp. 5–6). It seems unlikely that the NLSY participants did not know these strategies. However, they may not have cared enough about the outcome of the test to follow them.

[19] If the coding speed scores only proxied for how motivated the NLSY participants were to take the ASVAB, but were unrelated to their general level of intrinsic motivation, i.e., to their (stable) personality traits, we would not have expected the coding speed scores to be related to economic outcomes 23 years after they took the test.

[20] As a result, in 1989, the mathematics knowledge test replaced the numerical operation test in the AFQT.

None of the above-mentioned sources provides the raw test score distribution of potential recruits. However, using the information provided by Maier and Hiatt (1986), I was able to reconstruct (as described in §B.2 in Appendix B) the coding speed score distribution for the 1984 male applicants for enlistment (Initial Operational Test and Evaluation (IOT&E) 1984). Figure 2 presents the cumulative distribution function (CDF) of the coding speed scores for two groups of males: NLSY civilians born before January 1, 1963, and the IOT&E 1984 sample. Figure 2 clearly shows that the NLSY civilian population has the lowest test scores, in particular for the lower 80% of the test score distribution.[21] If indeed the coding speed scores measure effort, this is exactly what we would have expected to see if the potential recruits are highly motivated to take the ASVAB, whereas (not all) the NLSY participants are highly motivated.

The comparison between potential recruits and the NLSY participants provides indirect evidence that the coding speed scores may be highly related to motivation to take the ASVAB. However, other explanations are possible. Thus, to provide direct evidence that intrinsic motivation is important in determining the unincentivized coding speed scores, I turn to the controlled experiment.

7. Experimental Evidence

The model implies that to test whether individuals vary in their intrinsic motivation, one needs to investigate whether the relative ranking of individuals according to their test scores changes under varying incentive schemes. Improvement between tests is feasible even when intrinsic motivation does not vary across individuals, as long as they differ in their cognitive ability. To find changes in relative ranking, one needs to examine test scores of the same individuals for the same test under different incentive schemes. Because these data are not available, I conducted an experiment to investigate the part intrinsic motivation plays in determining the unincentivized coding speed scores.

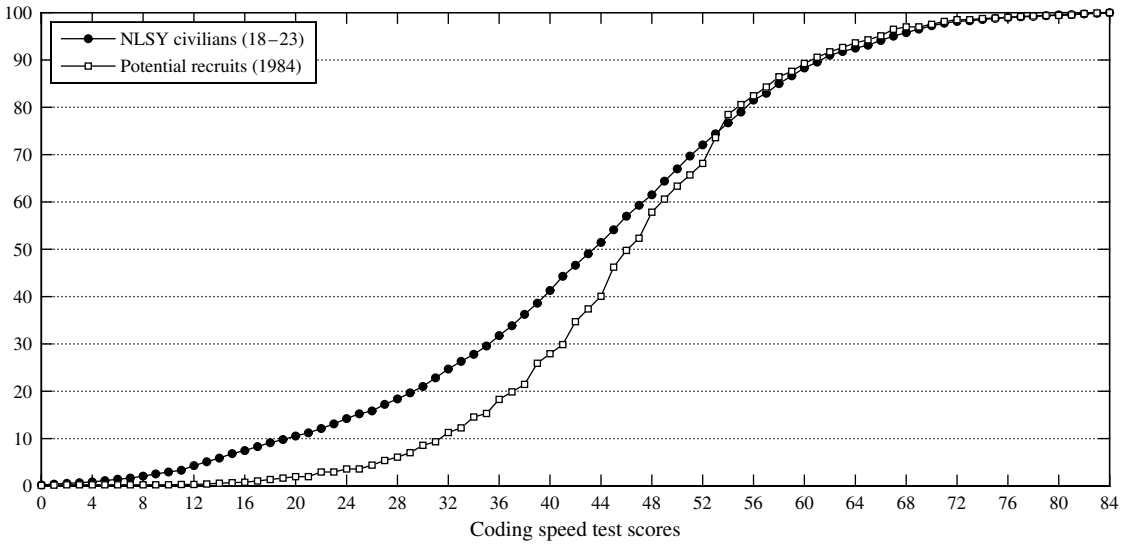
7.1. Experimental Design

The experiment consisted of two treatments, described below. The model implies that to distinguish between varying intrinsic motivation and varying cognitive ability, one needs to inspect rank changes. Thus, the design chosen is a within-subject design. In each of the treatments, participants solved three versions of the coding speed test. Each test lasted 10 minutes

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

RIGHTSLINK

Figure 2 CDFs of the Coding Speed Test Scores for NLSY Participants and Potential Recruits—Men



and consisted of 140 questions.[22] The experiment was conducted at Harvard using the Computer Lab for Experimental Research 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. Participants received a \$10 showing-up fee and an additional \$5 for completing the experiment. Participants were told in advance how many parts the experiment had, and that one would be randomly chosen for payment at the end of the experiment. Participants were informed about the tasks they needed to perform in each part and the compensation scheme immediately before performing the task. The instructions are given in the online appendix (available at https://sites.google.com/site/carmitsegal/). The specific compensation schemes and tasks in each treatment as well as the testing program are described below.

7.1.1. Main Treatment. *Part 1—Fixed Payment.* Participants were asked to solve two versions of the coding speed test. The first was called a "practice test." Their payment, if Part 1 was randomly chosen for payment, was \$10. Below, I refer to these 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, like the \$10 test, is administered without performance-based incentives, though (some) participants may have been less intrinsically motivated to take it. Thus, its scores can serve as a first indication of the importance of intrinsic motivation in determining the coding speed scores.

Part 2—Piece-Rate Compensation. Participants were asked to solve a third version of the coding speed test, 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 was 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 the role intrinsic motivation plays in determining the unincentivized 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 incentive 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.

[22] The tests were constructed in the following manner. For each test, 200 words were randomly chosen from a list of 240 words and were randomly ordered to construct 20 keys. For each word in the keys, a random number between 1,000 and 9,999 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.

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

Table 2 Means and Standard Deviations 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	1,864	1,785	1,768

Note. See §7.1.4 for details on guessing detection.

7.1.2. 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 was randomly selected for payment, they would receive \$10.

7.1.3. Survey. In both treatments, at the end of the experiment, after subjects solved the three tests, they were asked to answer a survey and a psychological questionnaire.[23]

7.1.4. The Testing Program—Performance Measures and Guessing Detection. The testing program allowed participants to move back and forth between the different keys and the answers associated with them by pressing buttons. Similarly, participants were allowed to move freely on the answer sheet in which answers were grouped in groups of 20. The 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, i.e., answering 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 20 30-second periods.

7.2. Basic Experimental Results

All 99 participants chose the piece-rate scheme in Part 2. Thus, all participants are included in the analysis. Table 2 reports the means and standard deviations of performance for the three tests. In the first three columns, performance is measured by the sum of correct answers. In the last three 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 (because some of the questions were already guessed correctly), I examine only the periods before the first guess. Participants' performance 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 $p < 0.001$).[24] Inspecting the number of correct answers in the periods before guessing, a similar picture arises. Between the first two tests, participants improved significantly by 0.82 correct answers ($p < 0.001$). Between the \$10 and incentives tests, participants significantly improved even further, and correctly solved, on average, 8.2 more questions ($p = 0.003$). Accordingly, on average, between the two tests, participants improved by 0.32 correct answers per 30 seconds ($p < 0.001$).

When examining the variance of the number of correct answers, a pattern emerges. The variance in test scores is the largest for the \$10 test. It increases by 77% compared to the incentives test and by 54% compared to the practice test (a two-sided F-test yields $p = 0.005$ and $p = 0.033$ for equality of variances between the first two tests and the last two tests, respectively).

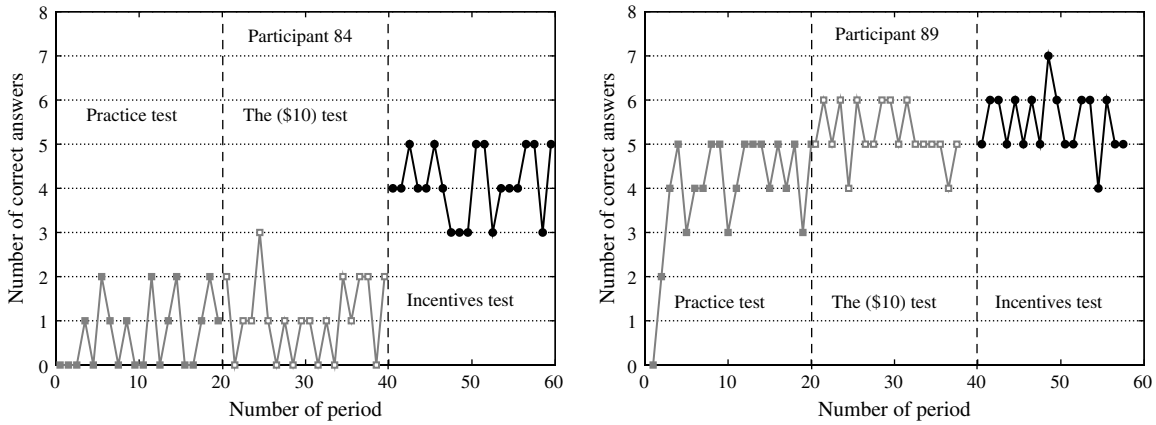
The improvement in performance may be in response to the incentive scheme or may indicate learning. To separate the two explanations, we would like to know what would have been participants' test score had they taken the coding speed test repeatedly without a change in the (implicit and explicit) incentives. The control treatment, in which participants took the test three times for a fixed payment, answers this question directly. The results of the control treatment are very different from those of 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 tests, respectively. Thus, participants have actually experienced an insignificant decrease of 5.1 correct answers on average between the second and the third tests ($p = 0.26$), instead of a significant increase in the main treatment. Between the first two tests, there was an insignificant improvement of 3.3 correct answers on average ($p = 0.3$), instead of a significant increase. In addition, whereas the test score distributions in the first test are not significantly different between the two treatments (the

[23] 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.

[24] Unless otherwise noted, the reported statistics refer to one-sided t-tests allowing for unequal variances.

RIGHTSLINK

Figure 3 Number of Correct Answers in 30-Second Periods Before First Guess by Test



Mann–Whitney test yields $p = 0.90$), they differ for the last two tests (the Mann–Whitney test yields $p = 0.04$ for the second test and $p < 0.001$ for the third test).[25] Thus, the results of the control treatment show that even if learning occurs, it occurs only if incentives are provided.

7.3. Do Individuals React Differently to the Lack of Incentives?

Classical economic theory predicts that participants will invest effort in solving the incentives test, but not in the practice and $10 tests. The left-hand panel of Figure 3 provides an example for such behavior by participant 84. Figure 3 depicts the number of correctly answered questions in each of the 30-second periods before participants 84 and 89 first guesses, for each of the three tests. Periods 1 to 20 depict performance on the practice test, periods 21 to 40 performance on the $10 test, and periods 41 to 60 performance on the incentives test. The (significantly) improved (average) performance of participant 84 on the incentives test suggests that although they were capable of solving the test, they did not care to show it. Without seeing participant 84's performance on the incentives test, we might have wrongly concluded that they were not capable of solving the coding speed test. The right-hand panel of Figure 3 depicts a different behavior by participant 89. Participant 89 did pretty well on the practice test. Nevertheless, once they were told that the test counted (the ($10) test), they improved (significantly) their (average) performance. However, the provision of performance-based incentives does not cause participant 89 to further improve (significantly) their performance. Nevertheless, on the incentives test participant 89 is still performing better than participant 84, who significantly improved their performance between the $10 and incentives tests.

The model provides a straightforward way to test whether the participants who behave like participant 84 are less intrinsically motivated or less cognitively 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 cognitive ability from the two tests.[26]

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 period before the participant's first guess.[27] Between the $10 and the incentives tests,

[25] The results when using the test scores in the periods before the first guess are qualitatively similar. Specifically, the mean performances (standard deviations) were 4.58 (1.47), 5.08 (1.46), and 5.03 (1.64) for the first, second, and third test, respectively. To examine the changes in performances between the tests and across treatments, I ran a difference-in-differences regression that included individual fixed effects and in which standard errors were clustered on the individual. I found that the average change in test scores between the last two tests in the treatment group relative to the control group (i.e., the double difference estimator) is 0.418 with a standard error of 0.123, which is highly significant (the coefficient (standard error) on the first difference between the last two tests for the control group is −0.064 (0.116)). Examining the differences between the first two tests (relative to the second test), I found that the coefficient (standard error) on the first difference for the control group is −0.554 (0.095). This shows that there is learning between the first two tests in the control treatment. However, the difference is significantly larger (by 40%) for the treatment group, as the double difference estimator (standard error) is −0.223 (0.103). Thus, the fact that in the treatment group the second test was called "the" test already had a significant effect on test scores.

[26] The increase in test scores between the $10 and the incentives tests cannot be attributed to differences in intrinsic motivation before showing (at least) that a fraction of participants invested the highest effort possible in the $10 test.

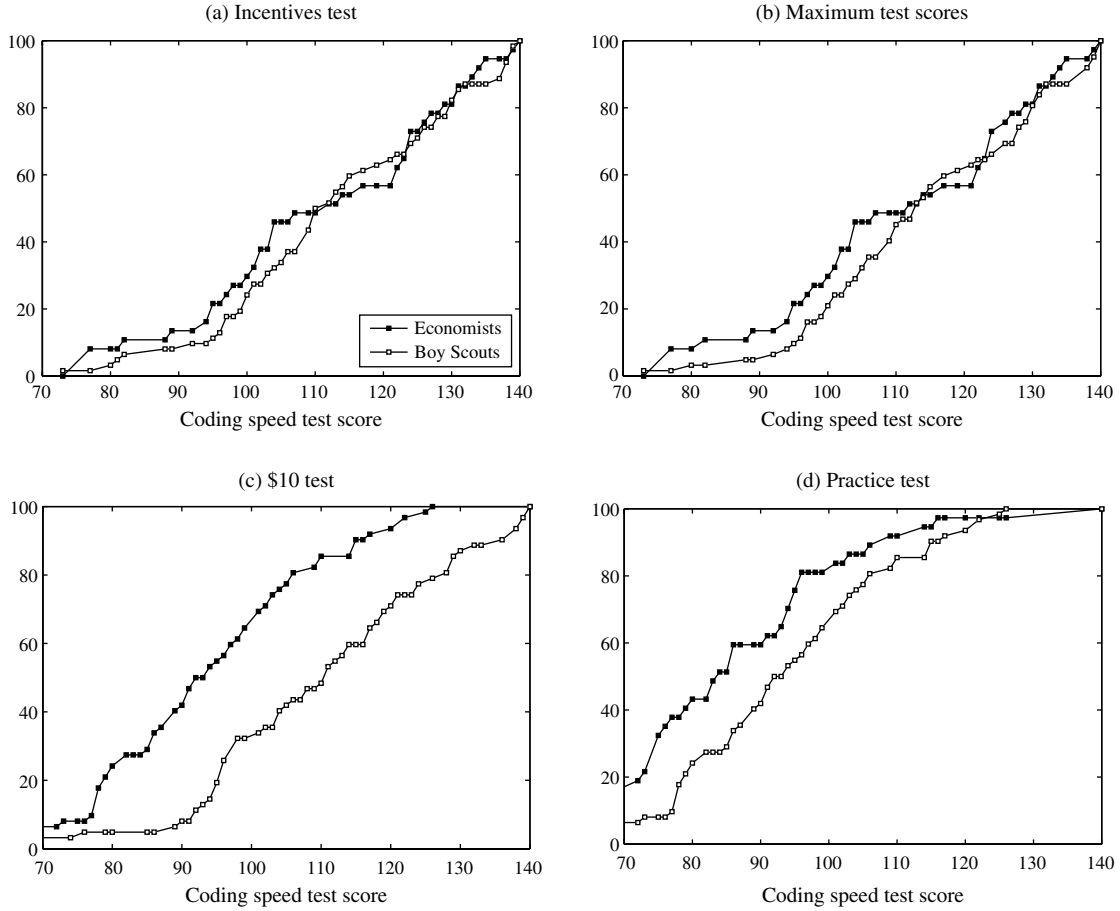
[27] Three individuals started guessing early on the $10 and incentives tests. Thus, it is impossible to test whether they significantly improved between the tests. However, all three had multiple periods in the $10 test in which they did not try to solve any question (one guessed the whole test in the first two minutes and then ended

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

Figure 4 CDFs of Coding Speed Test Scores by Compensation Scheme and Participants' Type



Notes. Economists are participants who improved their own average performance significantly between the $10 and incentives tests. Boy Scouts are participants who did not.

37 participants significantly improved their own performance,[28] whereas the other 62 participants did not. Of those, only two participants experienced a significant decline in their performance.[29] To simplify the exposition, I refer to the group whose members significantly improved their own performance as "Economists" (because their behavior agrees with economic theory predictions). I refer to the other group as "Boy Scouts" (because they seem to be trying their best even when no performance-based incentives were provided). Although the Economists significantly improved their performance, 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 test score distributions become the same. To test for stochastic dominance, I follow McFadden (1989). Neither the hypothesisthat the test score distribution of the Economists first-order stochastically dominates the distribution of the Boy Scouts ($p = 0.420$) nor its converse ($p = 0.757$) can be rejected. A similar picture arises when examining the maximum scores each participant achieved in the experiment (panel (b)), which are the best estimate of participants' cognitive ability. Again, there is no difference in the test score distributions between the two groups. Neither the hypothesis

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 answered no questions correctly. 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.

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

[29] The results below remain qualitatively and quantitatively the same if I exclude these two participants from the analysis or assign them to the other group.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

that the test score distribution of the Economists first-order stochastically dominates the test score distribution of the Boy Scouts ($p = 0.266$) nor its converse ($p = 0.839$) can be rejected. Thus, these two panels suggest that the underlying cognitive ability distribution of the Economists is not statistically different from that of the Boy Scouts.

A different picture arises when examining the total test scores of participants in the \$10 and practice tests (panels (c) and (d), respectively). 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 neither test can the converse 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. Whereas on average the Economists had 93.4 correct answers, the Boy Scouts had 110.6 ($p < 0.001$). This difference is as big as the standard deviation across participants in the incentives test. In contrast, the difference between the groups on the incentives test is two correct answers (111.1 for the Economists and 113.2 for the Boy Scouts, $p = 0.57$).

7.4. Effort vs. Other Explanations

Figure 4 and the subsequent tests suggest that when no monetary incentives were provided, a group of participants invested little effort (the Economists). Examining the test scores of this group on the \$10 test (or the practice test), one would have labeled the group as a low-cognitive-ability group. However, once performance-based monetary incentives were provided, it turned out that participants in this group had the same cognitive ability distribution as their fellow participants who chose to work hard all along (the Boy Scouts). In this section, I discuss alternative explanations and provide additional evidence in support of the hypothesis that indeed lack of effort can explain the group differences in the unincentivized coding speed scores.

At the end of the experiment, as part of the survey, participants were asked to report their SAT scores, and all but two men and one woman did. Participants' SAT scores do not vary across groups. The average SAT scores are 1,430.8 and 1,426.3 for the Economists and the Boy Scouts, respectively.[30] These differences are not significant ($p = 0.44$).[31] Research in psychology suggests that the SAT scores are highly correlated with conventional measures of fluid intelligence and with mental speed (Rohde and Thompson 2007). This may suggest that Economists and Boy Scouts have the same distributions of fluid intelligence, which is hypothesized to be the cognitive component part in the coding speed test. Together with the fact that both groups have the same incentivized coding speed scores, this implies that the differences in the unincentivized scores are the result of lack of effort on the part of the Economists.

More direct evidence can be found when examining the guessing behavior of the participants. Table 3 depicts the means of several performance variables for the participants who guessed and break them down by test and group. Table 3 suggests that when motivated, participants are solving more questions before the first guess, they solve each question faster (with at most minor losses to their accuracy), and, if they have the time, they keep solving the problems after guessing. The Boy Scouts are already doing it in "the" (\$10) test, whereas the Economists need the monetary incentives. Table 3 indicates that in the incentives test, the Boy Scouts and the Economists have different strategies regarding the timing of the first guess. The Economists started guessing earlier in the test (in terms of attempted questions and time), when they may have been less tired (as may be indicated by the highest fraction of correct answers before guessing). After guessing, they got back to the test and solved correctly a significantly larger fraction of the questions than the Boy Scouts. The different strategies yield the same number of attempted questions and fraction of correct answers and hence the same total test scores. This is not the case for the \$10 test. In this test, the Economists have lower test scores than the Boy Scouts, even though they attempted the same number of questions, because they answered correctly a

[30] Ideally, one would like the participants in the experiment to represent the general population. As is often the case in academic research, most participants are students. Although the participants in the experiment have above-average SAT scores, the variation in their scores is far from being negligible: the standard deviation of SAT scores in the experiment is 140 points (and among all SAT takers it is 212). Nevertheless, one may be concerned about the possibility to generalize the results of the experiment to the general population. In particular, one may be concerned that individuals with lower cognitive ability may not be able to respond to incentives because they have problems with answering the questions or have limited ability to concentrate. The coding speed test seems to be less problematic regarding these issues. The simplicity of the test suggests that the fraction of the population that cannot answer its questions is very small. The short length of the test suggests that relatively few individuals may not be able to concentrate for 10 minutes if they so chose. However, for this fraction of the population the results of the experiment cannot be generalized, and it will look as if it was composed only of Boy Scouts. Because the evidence suggests that conscientiousness is uncorrelated with cognitive ability (see, for example, Judge et al. 2007), we should be able to generalize the experimental results to all other populations.

[31] For the sample of 88 subjects who answered the psychological questionnaire in full, the average SAT scores of Economists is 1,428.2, whereas for Boy Scouts it is 1,421.6. These differences are not significant ($p = 0.42$). The two groups also have the same SAT scores when I test separately by gender.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

Table 3 Performance, Effort, and Guessing Behavior of Participants

	Practice test			The $10 test			Incentives test		
	Boy Scouts	Economists	*p*-value of diff.	Boy Scouts	Economists	*p*-value of diff.	Boy Scouts	Economists	*p*-value of diff.
Attempted	124.77	133.41	0.972	136.75	134.54	0.204	137	135.69	0.271
Fraction correct [%]	78.16	67.37	0.001	81.05	69.64	0.004	81.14	81.26	0.519
Attempted before First guess	85.5	80.17	0.186	101.87	82.24[a]	0.004	101.56	92.32[b]	0.068
Fraction correct before First guess [%]	96.91	95.16	0.054	97.8	96.95[a]	0.161	96.47	98.14[b]	0.992
Fraction correct after First guess [%]	36.14	25.99	0.013	32.05	31.01	0.364	38.77	44.58	0.922
First guess period	18.53	18.53	0.498	18.91	17.24[a]	0.049	18.9	17.54[b]	0.054
Total test score	96.00	89.82	0.102	110.78	93.84	0.002	111.15	110.28	0.410
Observations	30	17		36	26		41	29	

Note. The *p*-values are obtained from a one-sided *t*-test allowing for different variances where the hypothesis is that the Boy Scouts have higher values than the Economists.

[a]These numbers exclude a participant who guessed in the first period. If this person is added, average questions attempted before the first guess by the Economists in the $10 test is 79.08 (the corresponding *p*-value is 0.001), and the average period in which the Economists first guessed in the $10 test is 16.62 (the corresponding *p*-value is 0.026).

[b]These numbers exclude a participant who guessed in the first period. If this person is added, the average questions attempted before the first guess by the Economists in the incentives test is 89.14 (the corresponding *p*-value is 0.022), and the average period in which the Economists first guessed in the incentives test is 16.97 (the corresponding *p*-value is 0.029).

significantly lower fraction of questions. This is not because they solved a smaller fraction of questions correctly before guessing, but because they attempted (as in the incentives test) significantly fewer questions before guessing. However, in the $10 test this is not a strategy; here the Economists worked at a slower pace and never compensated for their guessing behavior. Even though they had the time, as they started guessing significantly earlier than the Boy Scouts, they did not get a higher fraction of questions right after guessing. Thus, the results presented in Table 3 suggest that the Economists just decided to stop working at some point during the $10 test and needed incentives to work harder and to maintain effort throughout the test.[32]

These results suggests that while the Boy Scouts were trying their best already in the $10, the Economists needed monetary incentives. Alternative explanations are that the Economists are slow learners or that the Boy Scouts are more risk-averse or have a lower valuation of money. In the control experiment, participants displayed no learning between the second and the third tests. This implies that a learning explanation cannot be that the Economists are inherently slow learners, but rather that they need incentives to put the effort and learn. The last two explanations suggest that the Boy Scouts are just smarter individuals (i.e., they can have high test scores without putting in effort) who did not react to the incentives. However, we have seen that the Boy Scouts and the Economists have the same cognitive ability as measured by their SAT scores. Therefore, it is hard to argue that the Boy Scouts are smarter. Moreover, Table 3 indicates that they increased their effort between the practice and $10 tests.[33] Last, even if we assume that the cognitive ability component of the coding speed test is different from the one in the SAT, it still seems unlikely that the improvement of the Economists will be such that the final test scores are the same even though the groups have different levels of cognitive skills.

We noted that the variance of the test scores is the largest in the $10 test. The evidence above suggests an explanation. In the incentives and practice tests, the heterogeneity in effort amongst subjects does not play a role, because in the former all invest high effort, and in the latter most invest little effort. It does play a role in the $10 test, in which approximately 60% of participants invest high effort and the rest invest little effort. As a result, the variance in test scores increases.

7.5. Individual Characteristics and Effort Choice

This paper suggests that the correlations between the coding speed scores and earnings in the NLSY stem from favorable personality traits of individuals who were intrinsically motivated to take the ASVAB.

[32] Table 3 also suggests that the differences in behavior have started in the practice test. Thus, although both groups started guessing at the same period, the Economists have a lower fraction of correct questions before and after the first guess. So, they paid less attention before guessing and tried less hard afterward.

[33] They did so by increasing the number of attempted questions before guessing ($p = 0.001$) and answering correctly more questions per 30-second period before guessing ($p < 0.001$).

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

In this section I investigate directly whether individual characteristics are correlated with participants' effort choices (i.e., with intrinsic motivation).

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, of the 50 male participants, 23 (46%) were classified as Economists. A chi-squared test for the equality of the distributions yields $p = 0.073$. Given these differences across gender, below I examine the relationship between effort choices and individual characteristics by gender.

After solving the three tests, participants were asked to answer the "Big Five" questionnaire.[34] The Big Five theory classifies personality traits into five dimensions (extroversion, agreeableness, conscientiousness, neuroticism, and openness to experience) and relates them to various aspects of individuals' life, including economic success (see Judge et al. 1999 for definitions and an excellent summary). Of these traits, conscientiousness, which is related to individuals' self-control, need for achievement, orderliness, and persistence, is the one that most consistently relates to job performance, job-seeking behavior, and retention at work (Judge et al. 1999). Moreover, low conscientiousness in adolescence is negatively related to success on the job (Roberts et al. 2003) and predicts counterproductive work behavior (Roberts et al. 2007) of young adults.

Overall, 91 participants (44 men and 47 women) answered all 50 questions; of those, two men and one woman did not report SAT scores. For this restricted sample, I find that male participants' effort choices can be related to conscientiousness. Specifically, the average conscientiousness level of male Economists is 8.6, whereas that for male Boy Scouts is 13.4 ($p = 0.024$). No other personality construct is related to male participants' type, and none relate to female participants' type.

7.6. Separating Effort and Cognitive Ability in the Coding Speed Scores

In the experiment I find that even when incentives are provided, participant test scores still differ. Thus, the coding speed scores contain a cognitive component in spite of the test's simplicity. This component captures participants' mental speed and thus may be related to fluid intelligence—novel problem-solving ability. In this section I use the experimental results to decompose the variance in scores to their components of cognitive ability and intrinsic motivation. This exercise will allow me to make precise statements about the relative importance of cognitive ability and intrinsic motivation to the unincentivized coding speed test scores.

To separate the different components of the test scores, I start by observing that if participants have indeed invested the highest level of effort in the incentives test, then the variation in their test scores on this test can be attributed only to cognitive ability or noise.[35] Thus, given a measure of noise, the variation in the incentives test scores can be used to identify the variation in cognitive ability, and therefore the variation in effort in the \$10 test can be identified too (with further functional form assumptions discussed below).

To perform this exercise, it is necessary to make an assumption about the production function of test scores. I assume that this function is multiplicative in cognitive ability and effort. This assumption could generate positive correlations between effort and cognitive ability found in the data (see Footnote 37) and is consistent with the psychometric literature on speeded tests (Rasch 1960).[36] Thus, I assume that the test scores of individual i on test j, where $j = 10$, and inc for the \$10 and the incentives test, respectively, is given by $TS_i^j = x_i e_i^j \epsilon_i^j$, where x denotes cognitive ability, e denotes effort invested in the test, and ϵ denotes noise. On the incentives test, individuals tried their best, hence, $e_i^{inc} = \bar{e}$ and $\widetilde{TS}_i^{inc} = \tilde{x}_i + \tilde{\bar{e}} + \tilde{\epsilon}_i^{inc}$, where the tilde denotes the natural logarithm of the variable. Thus, $\mathrm{Var}(\widetilde{TS}^{inc}) = \sigma_{\tilde{x}^2} + \sigma_{\tilde{\epsilon}}^2$, where σ_k^2 denotes the variance of the variable k. On the \$10 test, $\widetilde{TS}_i^{10} = \tilde{x}_i + \tilde{e}_i^{10} + \tilde{\epsilon}_i^{10}$. Thus,

$$\mathrm{Var}(\widetilde{TS}^{10}) = \sigma_{\tilde{x}}^2 + \sigma_{\tilde{e}^{10}}^2 + 2\,\mathrm{Cov}(\tilde{x}_i, \tilde{e}_i^{10}) + \sigma_{\tilde{\epsilon}}^2.$$

[34] The Big Five questionnaire included 50 statements (10 for each construct). The survey was taken from http://ipip.ori.org/newQform50b5.htm. It was administered without incentives, mainly because it is unclear how to provide incentives for such a test. Participants were asked to indicate on a five-point scale how accurately each statement described their usual behavior. To create the five constructs, the answers were added (or if negatively framed, were subtracted) within each construct.

[35] If effort in the incentives test is not maximal, its scores will depend on other parameters such as intrinsic motivation, risk aversion, money valuation, etc. Therefore, it will be impossible to identify cognitive ability separately from the other components. However, the results below present an upper bound on the importance of cognitive skills as long as these parameters are either not correlated or positively correlated with cognitive ability (the results above suggest that intrinsic motivation and cognitive ability are uncorrelated, and Dohmen et al. 2010 found positive correlations between risk aversion and cognitive ability).

[36] Following Rasch's (1960) stochastic model of speeded tests, the psychometric literature assumes that the expected number of correct answers in a given time unit is a multiplicative function of the individual's cognitive ability and a parameter describing the test or the experimental condition. In the case of this paper, the latter could be effort.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

Note that because the Boy Scouts invested the highest effort already in the \$10 test, for them $\widetilde{TS}_i^{10,BS} = \tilde{x}_i + \tilde{\tilde{e}} + \tilde{\epsilon}_i^{10}$. Therefore, $\Delta\widetilde{TS}_i^{BS} = \widetilde{TS}_i^{inc,BS} - \widetilde{TS}_i^{10,BS} = \tilde{\epsilon}_i^{inc} - \tilde{\epsilon}_i^{10}$. Hence, $\mathrm{Var}(\Delta\widetilde{TS}^{BS}) = 2\sigma_{\tilde{\epsilon}}^2$. Finally, $\mathrm{Cov}(\widetilde{TS}^{inc}, \widetilde{TS}^{10}) = \sigma_{\tilde{x}}^2 + \mathrm{Cov}(\tilde{x}_i, \tilde{e}_i^{10})$. Effort in the \$10 test is determined by cognitive ability and intrinsic motivation. Because $\mathrm{Cov}(\tilde{x}_i, \tilde{e}_i^{10})$ captures the variation in effort that is attributed to cognitive ability, the $\sigma_{\tilde{e}^{10}}^2$ estimate is solely attributable to intrinsic motivation. Because we can always normalize test scores, to ease the interpretation I normalized the natural logarithm of the \$10 test scores to have mean zero and standard deviation one and used the same transformation on the natural logarithm of the incentives test scores (which under this normalization have a mean of 0.32 and a variance of 0.263). Thus, using the experimental data and the equations above I get that $\sigma_{\tilde{\epsilon}}^2 = 0.026$, $\sigma_{\tilde{x}}^2 = 0.236$, $\mathrm{Cov}(\tilde{x}_i, \tilde{e}_i^{10}) = 0.152$, and therefore $\sigma_{\tilde{e}^{10}}^2 = 0.434$.[37] Thus, I find that a large fraction of the variation in unincentivized test scores (about 43%) can be attributed to variation in intrinsic motivation.

8. Conclusions

This paper uses experimental data, the NLSY survey data, and data from the military to investigate the relationship between the coding speed scores, intrinsic motivation, cognitive skills, and economic success. I use the NLSY data to establish that the unincentivized coding speed scores measure traits valuable in the labor market. I apply the military data to argue that the scores do not only measure cognitive ability. Specifically, the military is no longer using the test, even though it has incentivized scores that should provide an excellent measure of the cognitive component of the test. This suggests either that the military is now using another test that better measures the cognitive ability that was previously being measured by the coding speed test, or that this cognitive ability is unimportant for military performance. The evidence in Hunter (1986) rules out the latter and supports the former, and thus implies that the nonspeeded ASVAB scores can serve as controls for this cognitive ability. Controlling for all eight ASVAB nonspeeded scores, I find that an increase in the coding speed scores is significantly associated with an increase in earnings of male workers. I suggest that this association should be attributed to personality traits associated with intrinsic motivation. The experimental results establish that intrinsic motivation plays a significant role in determining the unincentivized coding speed scores. Moreover, it shows that intrinsic motivation does not relate to participants' cognitive ability as measured by their SAT scores and the incentivized coding speed scores. Instead, I find that male participants who are intrinsically motivated are significantly more conscientious.

The analysis in this paper has limitations. An external validity problem may arise because the experimental participants had above-average SAT scores. Another concern is that standard recruiting procedures in economic labs emphasize the opportunity to earn money. This could have resulted in the participants being unrepresentative in terms of their need for extrinsic rewards and therefore lack of intrinsic motivation. However, I find that compared to a larger, more diverse Internet sample of participants of similar age used by Srivastava et al. (2003), the experimental participants are more conscientious. Thus, if intrinsic motivation is being captured by participants' conscientiousness, this seems less of a problem.

This paper may have practical implications for preemployment screening. Specifically, to judge a candidate, managers would like to have answers to three questions: Is the candidate qualified? If so, is he likely to accept a job offer? And if he is, will he do the job well? To figure out what candidates know, management can rely on formal qualifications or administer a variety of knowledge tests. Answering the second question is a notoriously challenging task for employers in congested job markets (Coles et al. 2011). Currently, firms can get a reliable indication regarding applicants' intentions only in some centralized markets (see Coles et al. 2010 on the market for new economists). The evidence in this paper suggests a possible solution that does not necessitate any special market institutions. Specifically, firms could set a simple task/test that requires (substantial) effort to fulfill but no knowledge, and thus identify the applicants who scored the highest as the ones most interested in the job. This paper also suggests new possibilities answering the third question without relying on job applicants to truthfully report their personality traits. Specifically, if the firm could mimic a low-stakes environment during preemployment screening, it may be able to identify the least conscientious applicants by their low scores on very simple tests. Although creating the appearance of low-stakes environment is not easily done, it may be possible.[38] However, more

[37] Using these parameters, I find that for the Economists (who changed their effort levels between the tests), $\mathrm{Cov}(\tilde{x}, \tilde{e}^{10}) > 0$, which means that (conditional on intrinsic motivation) effort increases with cognitive ability. Based on Equation (A3) in Appendix A, a multiplicative production function can generate such a relationship.

[38] Some ideas with potential are as follows. Applicants could be told that they should quit solving the (simple) test as soon as they think that they have demonstrated their skill. Discussing an earlier version of this paper, Ayres and Nalebuff (2007) suggest telling job applicants that only half of the questions in the test would count. In both cases, the applicants who need explicit incentives will invest less effort solving the test and therefore will have lower scores.

RIGHTSLINK

research is needed before these possibilities could become tools in the hand of management.

Acknowledgments

The author thanks Ed Lazear, Muriel Niederle, and Al Roth for their encouragement, useful suggestions, and numerous conversations. She thanks Harvard Business School for generous support and hospitality, and the associate editor and two referees for their helpful comments and suggestions. She thanks George Baker, Greg Barron, Vinicius Carrasco, Pedro Dal Bo, Liran Einav, Florian Englmaier, Itay Fainmesser, Richard Freeman, Ed Glaeser, Avner Greif, Ben Greiner, Felix Kubler, Steve Leider, Aprajit Mahajan, Tatiana Melguizo, Guy Michaels, Amalia Miller, Joao de Mello, Rosemarie Nagel, Andreas Ortmann, Luigi Pistaferri, Daniel Tsiddon, Lise Vesterlund, Ed Vytlacil, Pierre-Olivier Weill, Toni Wegner, Catherine Weinberger, and Nese Yildiz for helpful comments on earlier drafts. The author also thanks multiple discussants and seminar participants for their insights. All errors are the author's responsibility.

Appendix A. Proof of Propositions

Proof of Proposition 1. When performance-based incentives are provided and/or agents obtain psychological benefits from higher test scores, then the optimal level of effort, e^*, solves

$$\frac{\partial TS(x, e^*)}{\partial e}(U_{TS}(\theta) + U_M\phi) - C_e(e^*) = 0. \tag{A1}$$

The second-order condition is $D \equiv TS_{ee}(U_{TS} + \phi U_M) + TS_e^2(U_{TS,TS} + 2\phi U_{TS,M} + \phi^2 U_{M,M}) - C_{ee}$. Given the assumptions, a sufficient condition ensuring that an internal solution is a maximum (i.e., $D < 0$) is $TS_{ee} \le 0$.

Note that because U_{TS} is a function of θ, then e^* depends on θ. At the optimal level of effort, $e^*(\theta)$, the relations between test scores and cognitive ability, conditional on θ, are given by

$$\frac{dTS}{dx} = TS_x + TS_e\frac{de^*(\theta)}{dx}. \tag{A2}$$

To find how the optimal level of effort, $e^*(\theta)$, depends on cognitive skills, differentiate $e^*(\theta)$ with respect to x to get

$$\frac{de^*(\theta)}{dx} = -\frac{1}{D}\big[TS_{ex}(U_{TS} + \phi U_M) + TS_e TS_x(U_{TS,TS} + 2\phi U_{TS,M} + \phi^2 U_{M,M})\big]. \tag{A3}$$

Using Equations (A2) and (A3), we get $dTS/dx = (1/D)[(TS_x TS_{ee} - TS_e TS_{ex})(U_{TS} + \phi U_M) - TS_x C_{ee}] > 0$.[39] Therefore, conditional on θ, test scores increase with cognitive skills, regardless of the relations between $e^*(\theta)$ and x.

To see that, conditional on θ, an increase in the incentives, i.e., an increase in ϕ, will result in an increase in the optimal level of effort, differentiate $e^*(\theta)$ with respect to ϕ, to get that $de^*(\theta)/d\phi = -(TS_e/D)[U_M + TS(U_{TS,M} + \phi U_{M,M})]$.

[39] If the costs function depends on cognitive ability, then the numerator in Equation (A3) has an additional term which is $-C_{ex}$. In addition, the numerator in dTS/dx has an additional term equal to $+TS_e C_{ex}$. Even with this, $dTS/dx > 0$.

Under the assumption that the marginal benefits are increasing in ϕ (i.e., $U_M + TS(U_{TS,M} + \phi U_{M,M}) > 0$), it is clear that $de^*(\theta)/d\phi$. As a result $dTS/d\phi = TS_e(de^*(\theta)/d\phi)$ 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.

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

Proof of Proposition 2. 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. Because $TS_1(\phi)$ first-order stochastically dominates $TS_2(\phi)$, then

$$\int_{TS=\underline{TS}_1(\phi)}^{z} \tilde{f}_1(TS_1; \phi)\, dTS \le \int_{TS=\underline{TS}_2(\phi)}^{z} \tilde{f}_2(TS_2; \phi)\, dTS \tag{A4}$$

for all z. Because all individuals have the same intrinsic motivation, test scores provide a correct ranking according to cognitive skills. Thus, $\underline{TS}_1(\phi) \ge \underline{TS}_2(\phi)$ and $\overline{TS}_1(\phi) \ge \overline{TS}_2(\phi)$.

By Proposition 1, because all agents have the same θ, then test scores are monotonically increasing function of cognitive skills, i.e., $dTS/dx > 0$, $\forall x$. 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

$$\int_{x=\underline{x}_1(\phi)}^{TS_1^{-1}(z)} f_1(x_1; \phi)\frac{dTS}{dx_1}\, dx_1 \le \int_{x=\underline{x}_2}^{TS_2^{-1}(z)} f_2(x_2; \phi)\frac{dTS}{dx_2}\, dx_2, \tag{A5}$$

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

From Proposition 1 we know that because all agents have the same θ, the test scores provide a correct ranking according to cognitive skills, i.e., there is one-to-one mapping between test scores and cognitive skills for all agents, 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, $\int_{x=\underline{x}_i}^{TS_i^{-1}(x)} f_i(x_i)(dTS/dx_i)\, dx_i = \int_{x=\underline{x}_i}^{TS^{-1}(x)} f_i(x)(dTS/dx)\, dx$, where $i = 1, 2$. Similarly, because $\underline{TS}_1 \ge \underline{TS}_2$ and $\overline{TS}_1 \ge \overline{TS}_2$, then $\underline{x}_1 \ge \underline{x}_2$ and $\bar{x}_1 \ge \bar{x}_2$. Thus, we can rewrite (5) as

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

Let R be the lowest value of dTS/dx in the range, i.e., $R \le dTS/dx$ for all x. Then,

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

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

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

The first part can be proven by noticing that the proof above fails already at the first assertion. Proposition 1 states that $(dTS/dx)|_\theta > 0$, i.e., test scores provide a correct ranking according to cognitive skills only for agents 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 1 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 cognitive skills distributions of the two types, which are what we want to recover, even in the case where $TS(x_1;\phi,\theta_1)$ first-order stochastically dominates $TS(x_2;\phi,\theta_2)$ we cannot guarantee that $x_1(\theta_1)$ first-order stochastically dominates $x_2(\theta_2)$.[40] □

Appendix B. Data

B.1. NLSY Sample Restrictions and Variable Construction

The sample used in §5 includes only men who were surveyed in 2004 with valid test scores.[41] Because it is possible that test scores or intrinsic 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., men born between October 1, 1961, and September 30, 1964.

Participants in the NLSY took the ASVAB exam in the summer of 1980 when they were 15–23 years old. This age differential may affect test scores because of differences in either educational attainment (see, for example, Hansen et al. 2004, Cascio and Lewis 2006) or in maturity. Indeed, the ASVAB scores increase with age. Thus, to compare test scores of individuals of different age groups, an adjustment is needed. I adjust the test scores used in the analysis in §5 by school-year cohorts, where a school-year cohort includes all individuals born between October 1 of one calendar year and September 30 of the subsequent one. Specifically, the residuals from the regressions of the test scores' variables on school-year cohort dummies for the restricted sample were normalized, using the ASVAB sampling weights, to have weighted mean zero and standard deviation one. 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 this normalization is that it excludes from the analysis participants who were born after September 30, 1964, who are believed to be a nonrandom sample.[42]

[40] In additions to these assumptions, one needs to assume that $d^2TS/dxd\theta \geq 0$ to get that stochastic dominance in test scores implies stochastic dominance in cognitive skills levels.

[41] Hispanic participants who did not understand English were excluded from the analysis.

[42] The NLSY sample includes "too few" participants born after September 30, 1964, compared to the general population (Center for Human Resource Research 2001, pp. 19–20).

Table B.1 The ASVAB Subtests

Subtest	Time (min)	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
Coding speed	7	84	Measures ability to use a key in assigning code numbers to words
Auto and shop information	11	25	Measures knowledge of automobiles, tools, and shop terminology and practices
Mathematics knowledge	24	25	Measures knowledge of high school 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

The variable *years of schooling completed* used in §5 was constructed using the variables *years of schooling completed as of May 1 of the survey year* and *highest degree ever received*. For those who reported receiving no high school diploma, the actual years of schooling were used. Degrees were converted to years of schooling in the usual manner (i.e., 12 years for a high school diploma, 14 for an A.A. degree, 16 for a B.A. or B.S. degree, 18 for finishing professional school or an M.S. or M.A. degree, and 20 for a Ph.D.); 13 years were assigned to persons who received no degree but reported completing at least a year of post secondary schooling.[43]

In the regressions reported in §5, I used the 2004 cross-sectional weights.

B.2. Armed Forces Data: Variable Construction for Figure 2

The test score distribution of the 1984 male applicants for enlistment (IOT&E 1984), reported in Figure 2, was

[43] The NLSY variable *years of schooling completed as of May 1* takes the value 16 years if a person received B.A. or B.S. However, the value 17 years of schooling is assigned to those who continued to graduate school, but also those who did not graduate from college within four years. The coding above maintains the link that higher educational attainment corresponds to more years of completed schooling.

RIGHTSLINK

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

constructed using the data provided by Maier and Hiatt (1986). The authors provided, in Table A–4 (pp. A9–A10), a conversion between the coding speed scores of the NLSY participants born before January 1, 1963, and the IOT&E 1984 scores. This conversion was done by setting the raw scores of individuals from the two groups that had the same cumulative frequency, conditional on measure of ability, equal to one another. The ability measure used was the health, social, and technology (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 NLSY participants (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. First, Maier and Hiatt (1986) did 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 because it will make the IOT&E 1984 population look worse. Second, Maier and Hiatt (1986) reported a range for the coding speed test score of two (2–10); I have again taken the lowest value (2).

Appendix C. Additional NLSY Regression Results

Table C.1 **Earnings of Men (Dependent Variable: ln of *Earnings 2003*)**

	(1)	(2)	(3)	(4)
Black	−0.202	−0.286	−0.276	−0.276
	(0.069)***	(0.069)***	(0.070)***	(0.070)***
Hispanic	−0.123	−0.150	−0.151	−0.151
	(0.071)*	(0.069)**	(0.069)**	(0.069)**
AFQT scores	0.278	0.123		
	(0.033)***	(0.040)***		
Coding speed scores	0.092	0.064	0.069	
	(0.027)***	(0.026)**	(0.026)***	
AFQT scores: *College grads*			0.262	0.262
			(0.067)***	(0.067)***
AFQT scores: *Less than college*			0.095	0.095
			(0.044)**	(0.044)**
Coding speed scores: *College grads*				0.069
				(0.051)
Coding speed scores: *Less than college*				0.069
				(0.031)**
Years of schooling (*2003*)		0.104	0.084	0.084
		(0.014)***	(0.016)***	(0.016)***
Age in 2003	0.015	0.025	0.023	0.023
	(0.029)	(0.029)	(0.029)	(0.029)
Constant	10.062	8.251	8.588	8.588
	(1.188)***	(1.223)***	(1.186)***	(1.184)***
Observations	1,187	1,187	1,187	1,187
R^2	0.18	0.23	0.24	0.24

Notes. See notes to Table 1. Robust standard errors are in parentheses.
*Significant at 10%; **significant at 5%; ***significant at 1%.

References

Angrist, J., V. Lavy. 2009. The effect of high stakes high school achievement awards: Evidence from a school-centered randomized trial. *Amer. Econom. Rev.* **99**(4) 1384–1414.

Ariely, D., U. Gneezy, G. Loewenstein, N. Mazar. 2009. Large stakes and big mistakes. *Rev. Econom. Stud.* **76**(2) 451–469.

Ayres, I., B. Nalebuff. 2007. For the love of the game. *Forbes Magazine* (March 12), http://www.law.yale.yale.edu/news/4741.html.

Benjamin, D. J., S. A. Brown, J. M. Shapiro. 2005. Who is "behavioral"? Cognitive ability and anomalous preferences. Working paper, Cornell University, Ithaca, NY.

Borghans, L., H. Meijers, B. ter Weel. 2008. The role of noncognitive skills in explaining cognitive test scores. *Econom. Inquiry* **46**(1) 2–12.

Borghans, L., B. Golsteyn, J. Heckman, J. E. Humphries. 2009. IQ, achievement, and personality. Working paper, University of Maastricht, Maastricht, The Netherlands.

Bowles, S., H. Gintis, M. Osborne. 2001. The determinants of earnings: A behavioral approach. *J. Econom. Literature* **39**(4) 1137–1176.

Cascio, E. U., E. G. Lewis. 2006. Schooling and the armed forces qualifying test. *J. Human Resources* **41**(2) 294–318.

Cawley, J., K. Conneely, J. Heckman, E. Vytlacil. 1997. Cognitive ability, wages, and meritocracy. B. Devlin, S. E. Fienberg, D. P. Resnick, K. Roeder, eds. *Intelligence, Genes, and Success: Scientists Respond to the Bell Curve*. Springer-Verlag, New York, 179–192.

Center for Human Resource Research. 2001. *NLSY79 User's Guide: A Guide to the 1979–2000 National Longitudinal Survey of Youth Data.* CHRR, The Ohio State University, Columbus.

Coles, P., A. Kushnir, M. Niederle. 2011. Preference signaling in matching markets. Working paper, Harvard Business School, Cambridge, MA.

Coles, P., J. Cawley, P. B. Levine, M. Niederle, A. E. Roth, J. J. Siegfried. 2010. The job market for new economists: A market design perspective. *J. Econom. Perspect.* **24**(4) 187–206.

Dohmen, T. J., A. Falk, D. Huffman, U. Sunde. 2010. Are risk aversion and impatience related to cognitive ability? *Amer. Econom. Rev.* **100**(3) 1238–1260.

Duckworth, A. L., P. D. Quinn, D. Lynam, R. Loeber, M. Stouthamer-Loeber, T. E. Moffit, A. Caspi. 2009. What intelligence tests test: Individual differences in test motivation and IQ. Working paper, University of Pennsylvania, Philadelphia.

Fehr, E., A. Falk. 1999. Wage rigidity in a competitive incomplete contract market. *J. Political Econom.* **107**(1) 106–134.

Gneezy, U., A. Rustichini. 2000. Pay enough or don't pay at all. *Quart. J. Econom.* **115**(3) 791–810.

Hansen, K. T., H. James, K. J. Mullen. 2004. The effect of schooling and ability on achievement test scores. *J. Econometrics* **121**(1) 39–98.

Heckman, J. 1995. Lessons from the bell curve. *J. Political Econom.* **103**(5) 1091–1120.

Heckman, J., Y. Rubinstein. 2001. The importance of noncognitive skills: Lessons from the ged testing program. *Amer. Econom. Rev.* **91**(2) 145–149.

Heckman, J., J. Stixrud, S. Urzua. 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *J. Labor Econom.* **24**(3) 411–482.

Heckman, J., L. Malofeeva, R. Pinto, P. A. Savelyev. 2009. The effect of the Perry Preschool program on the cognitive and noncognitive skills of its participants. Working paper, University of Chicago, Chicago.

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

Hunter, J. E. 1986. Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *J. Vocational Behav.* **29**(3) 340–362.

Johnson, W. R., D. A. Neal. 1998. Basic skills and the black-white earnings gap. C. Jencks, M. Phillips, eds. *The Black-White Test Score Gap*. The Brookings Institute, Washington, DC, 480–498.

RIGHTSLINK

Judge, T. A., C. A. Higgins, C. J. Thoresen, M. R. Barrick. 1999. The big five personality traits, general mental ability, and career success across the life span. *Personnel Psych.* **52**(3) 621–652.

Judge, T. A., C. L. Jackson, J. C. Shaw, B. A. Scott, B. L. Rich. 2007. Self-efficacy and work-related performance: The integral role of individual differences. *J. Appl. Psych.* **92**(1) 107–127.

Kremer, M., E. Miguel, R. Thorton. 2009. Incentives to learn. *Rev. Econom. Statist.* **91**(3) 437–456.

Kuhn, P., C. Weinberger. 2005. Leadership skills and wages. *J. Labor Econom.* **23**(3) 395–434.

Maier, M. H., C. M. Hiatt. 1986. Evaluating the appropriateness of the numerical operations and math knowledge subtests in the AFQT. CNA Memorandum 86-228, Center for Naval Analyses, Alexandria, VA.

Maier, M. H., W. Sims. 1983. The appropriateness for military applicants of the ASVAB subtests and score scale in the new 1980 reference population. CNA Memorandum 83-3102, Center for Naval Analyses, Alexandria, VA.

Maier, M. H., W. Sims. 1986. The ASVAB score scales: 1980 and World War II. CNA Report 116, Center for Naval Analyses, Alexandria, VA.

McFadden, D. L. 1989. Testing for stochastic dominance. T. Fomby, T. K. Seo, eds. *Studies in the Economics of Uncertainty*. Springer, New York, 113–134.

Mulligan, C. B., Y. Rubinstein. 2008. Selection, investment, and women's relative wages over time. *Quart. J. Econom.* **123**(3) 1061–1110.

Nagin, D., J. Rebitzer, S. Sanders, L. Taylor. 2005. Monitoring, motivation and management: The determinants of opportunistic behavior in a field experiment. *Amer. Econom. Rev.* **92**(4) 850–873.

Neal, D. A., W. R. Johnson. 1996. The role of premarket factors in black-white wage differences. *J. Political Econom.* **104**(5) 869–895.

Persico, N., A. Postlewaite, D. Silverman. 2004. The effect of adolescent experience on labor market outcomes: The case of height. *J. Policital Econom.* **112**(5) 1019–1053.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.

Revelle, W. 1993. Individual differences in personality and motivation: "Non-cognitive" determinants of cognitive performance. A. Baddeley, L. Weiskrantz, eds. *Attention: Selection, Awareness, and Control: A tribute to Donald Broadbent*. Oxford University Press, Oxford, UK, 346–373.

Roberts, B. W., A. Caspi, T. E. Moffitt. 2003. Work experiences and personality development in young adulthood. *J. Personality Soc. Psych.* **84**(3) 582–593.

Roberts, B. W., P. D. Harms, A. Caspi, T. E. Moffitt. 2007. Predicting the counterproductive employee in a child-to-adult prospective study. *J. Appl. Psych.* **92**(3) 1427–1436.

Rohde, T. E., L. A. Thompson. 2007. Predicting academic achievement with cognitive ability. *Intelligence* **35**(1) 83–92.

Segal, C. 2012. Misbehavior, education, and labor market outcomes. *J. Eur. Econom. Assoc.* Forthcoming.

Segall, D. O. 2006. Equating the CAT-ASVAB. Technical bulletin, Personnel Testing Division, Defense Manpower Data Center, Washington, DC, 9-1–9-43.

Srivastava, S., O. P. John, S. D. Gosling, J. Potter. 2003. Reconsidering the use of personality tests in personnel selection contexts: Development of personality in early and middle adulthood: Set like plaster or persistent change? *J. Personality Soc. Psych.* **84**(5) 1041–1053.

Steele, C. M., J. Aronson. 1998. Stereotype threat and the test performance of academically successful african americans. C. Jencks, M. Phillips, eds. *The Black-White Test Score Gap*. The Brookings Institute, Washington, DC, 401–427.

Tversky, A., D. Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* **211**(4481) 453–458.

U.S. Department of Defense. 1982. Profile of American youth: 1980 nationwide administration of the armed services vocational aptitude battery. Office of the Assistant Secretary of Defense, U.S. Department of Defense, Washington, DC.

Wolfe, J. H., K. E. Moreno, D. O. Segall. 2006. Evaluating the predictive validity of CAT-ASVAB. Technical bulletin, Personnel Testing Division, Defense Manpower Data Center, Washington, DC, 8-1–8-120.

Downloaded from informs.org by [155.246.103.35] on 06 April 2017, at 14:56 . For personal use only, all rights reserved.

RIGHTSLINK