Role of test motivation in intelligence testing

Angela Lee Duckworth[a,1], Patrick D. Quinn[b], Donald R. Lynam[c], Rolf Loeber[d], and Magda Stouthamer-Loeber[d]

[a]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104; [b]Department of Psychology, University of Texas at Austin, Austin, TX 78712; [c]Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907; and [d]Department of Psychiatry, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213

Edited by Edward E. Smith, Columbia University, New York, NY, and approved March 25, 2011 (received for review December 14, 2010)

Intelligence tests are widely assumed to measure maximal intellectual performance, and predictive associations between intelligence quotient (IQ) scores and later-life outcomes are typically interpreted as unbiased estimates of the effect of intellectual ability on academic, professional, and social life outcomes. The current investigation critically examines these assumptions and finds evidence against both. First, we examined whether motivation is less than maximal on intelligence tests administered in the context of low-stakes research situations. Specifically, we completed a meta-analysis of random-assignment experiments testing the effects of material incentives on intelligence-test performance on a collective 2,008 participants. Incentives increased IQ scores by an average of 0.64 SD, with larger effects for individuals with lower baseline IQ scores. Second, we tested whether individual differences in motivation during IQ testing can spuriously inflate the predictive validity of intelligence for life outcomes. Trained observers rated test motivation among 251 adolescent boys completing intelligence tests using a 15-min "thin-slice" video sample. IQ score predicted life outcomes, including academic performance in adolescence and criminal convictions, employment, and years of education in early adulthood. After adjusting for the influence of test motivation, however, the predictive validity of intelligence for life outcomes was significantly diminished, particularly for nonacademic outcomes. Collectively, our findings suggest that, under low-stakes research conditions, some individuals try harder than others, and, in this context, test motivation can act as a third-variable confound that inflates estimates of the predictive validity of intelligence for life outcomes.

One of the most robust social science findings of the 20th century is that intelligence quotient (IQ) scores predict a broad range of life outcomes, including academic performance, years of education, physical health and longevity, and job performance (1–7). The predictive power of IQ for such diverse outcomes suggests intelligence as a parsimonious explanation for individual and group differences in overall competence.

However, what is intelligence? Boring's now famous reply to this question was that "intelligence as a measurable capacity must at the start be defined as the capacity to do well in an intelligence test. Intelligence is what the tests test." (ref. 8, p. 35). This early comment augured the now widespread conflation of the terms "IQ" and "intelligence," an unfortunate confusion we aim to illuminate in the current investigation.

Intelligence has more recently—and more usefully—been defined as the "ability to understand complex ideas, to adapt effectively to the environment, to learn from experience, to engage in various forms of reasoning, to overcome obstacles by taking thought" (ref. 5, p. 77). IQ scores, in contrast, measure the performance of individuals on tests designed to assess intelligence. That is, IQ is an observed, manifest variable, whereas intelligence is an unobserved, latent variable.

That IQ scores do not perfectly capture latent intelligence is well known. However, to the extent that IQ scores are affected by systematic biases and not just random measurement error, there is the worrisome possibility that IQ–outcome associations are also systematically biased. The direction of bias depends on the relation between test-taking motivation and life outcomes: If test motivation does not derive from relatively stable and adaptive traits, then the influence of test motivation will erode IQ–outcome associations, indicating that current inferences about the effects of intelligence on success in life are spuriously low. If, on the other hand, the tendency to try hard on low-stakes intelligence tests derives from what Wechsler called "nonintellective" traits (9) (e.g., competitiveness, compliance with authority) that *also* predict life outcomes, then test motivation will inflate IQ–outcome associations, resulting in an overestimation of the predictive power of intelligence (Fig. 1).

In the current investigation, we hypothesize that individual differences in low-stakes test motivation are, in fact, much greater than currently assumed in the social science literature. Further, we hypothesize that test motivation is a third-variable confound that tends to inflate, rather than erode, the predictive power of IQ scores for later-life outcomes.

Eliciting Maximal Intellectual Performance. Intelligence-test procedures are designed to maximize the motivation of test takers (10). For instance, directions from the Third Edition of the Wechsler Intelligence Scale for Children (WISC-III) manual suggest, "If the child says that he or she cannot perform a task or cannot answer a question, encourage the child by saying, 'Just try it' or 'I think you can do it. Try again.'" (ref. 11, p. 37). Similarly, the deliberate sequencing of items from easy to difficult is an explicit strategy for sustaining morale (12). We submit that the motivation-maximizing design features of intelligence tests do not always succeed in maximizing effort, particularly in the context of research studies in which test takers face no consequences for good or bad performance. As Revelle has pointed out, "A common assumption when studying human performance is that subjects are alert and optimally motivated. It is also assumed that the experimenter's task at hand is by far the most important thing the subject has to do at that time. *Thus, although individual differences in cognitive ability are assumed to exist, differences in motivation are ignored*." (emphasis added; ref. 13, pp. 352–353).

Prior Research on Test Motivation and Intelligence. Prior studies have found that self-reported motivation is higher on employment tests among job applicants than incumbents (14) ($d = 0.92$) and among eighth grade students paid for correctly answering questions compared with students not paid for correct answers (15) ($d = 0.42$). However, because ratings of motivation are typically self-reported post hoc, it is possible that they reflect how well test takers think they performed as opposed to how hard they tried. Thus, the direction of causality in studies using self-report measures of test motivation is unclear. Moreover, prior studies have not directly examined whether variance in test

Author contributions: A.L.D., D.R.L., R.L., and M.S.-L. designed research; A.L.D., P.D.Q., D.R.L., R.L., and M.S.-L. performed research; A.L.D. and P.D.Q. analyzed data; and A.L.D. and P.D.Q. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: duckworth@psych.upenn.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1018601108/-/DCSupplemental.

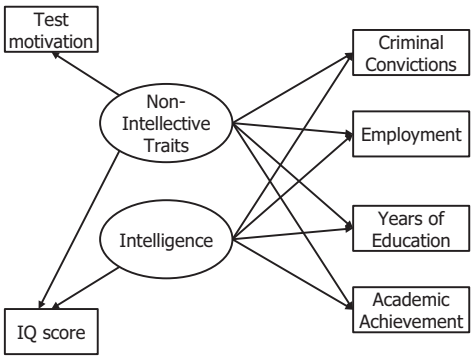

Fig. 1. Hypothesized associations among IQ, test motivation, life outcomes, and latent variables.

motivation is associated with life outcomes and, as a consequence, either inflates or erodes IQ–outcome associations.

Current Investigation. If indeed test takers are less than maximally motivated in low-stakes research settings, material incentives should substantially improve their performance. We tested this hypothesis in Study 1, a random-effects meta-analysis of random-assignment laboratory experiments comparing IQ scores under incentivized and standard testing conditions in 46 independent samples. In Study 2, we tested whether objectively measured test motivation confounds the predictive association between intelligence, as indexed by IQ scores, and later-life outcomes. Specifically, in a longitudinal study of 251 boys followed from adolescence to early adulthood, we tested whether the nonintellective traits underlying test motivation predict the same academic (i.e., school performance in adolescence and total years of education) and nonacademic (i.e., employment and criminal behavior) outcomes as does IQ and whether the predictive validity of intelligence for outcomes is reduced when test motivation is measured and controlled.

Results

Study 1. In 46 independent samples (n = 2,008), the mean effect of material incentives on IQ was medium to large: g = 0.64 [95% confidence interval (CI) = 0.39, 0.89], $P < 0.001$. An examination of Table S1, which lists the raw effect size from each sample (16–40), reveals that a small number of samples with very large effect sizes may have exerted undue influence on the mean effect size. To ensure that these samples did not account for the significance of the effect, we excluded the three samples with raw effect sizes greater than g = 2.00 and recomputed the mean effect. In the remaining 43 samples, the effect was still medium in size and statistically significant: g = 0.51 (95% CI = 0.31, 0.72), $P < 0.001$. Three of four tests suggested no publication bias on effect-size estimates (*SI Materials and Methods*).

A test of heterogeneity among all 46 samples indicated that between-study variance accounted for 85% of the variance in effect sizes: $Q(45) = 303.68$, $P < 0.001$, $I^2 = 85.18$. We therefore tested whether baseline IQ, incentive size, age, or study design accounted for heterogeneity in effect sizes.

Because exact baseline IQ scores were not reported in some samples, we created a binary variable where 1 = below average (i.e., IQ < 100) and 2 = above average (i.e., IQ ≥ 100). The effect of incentives was greater for individuals of below-average baseline IQ: $Q_{\text{between}}(1) = 9.76$, $P = 0.002$. In 23 samples with IQ scores below the mean, the effect size was large: g = 0.94 (95% CI = 0.54, 1.35). In contrast, in 23 samples of above-average IQ, the effect was small: g = 0.26 (95% CI = 0.10, 0.41). A similar analysis in which baseline IQ scores (available for 43 of 46 samples) were treated as a continuous moderator indicated that a 1 SD increase in IQ is associated with about two-thirds of an SD decrease in the effect of incentives: $b = -0.04$, $P < 0.001$. Moderation by baseline IQ did not account for all heterogeneity in effect size among low-IQ samples: $Q(22) = 226.23$, $P < 0.001$, $I^2 = 90.28$. In contrast, heterogeneity in effect size among high-IQ samples was not significantly different from zero: $Q(22) = 24.37$, $P = 0.33$, $I^2 = 9.71$ (Table S2).

As predicted, a systematic dose–response relationship was observed between incentive size and IQ score gain: $Q_{\text{between}}(2) = 28.95$, $P < 0.001$. Excluding three samples for which incentive size was not reported, large incentives produced a very large effect [g = 1.63 (95% CI = 1.15, 2.10)], whereas medium [g = 0.58 (95% CI = 0.37, 0.79)] and small [g = 0.16 (95% CI = −0.09, 0.41)] incentives produced smaller effects.

Neither sample age nor study design were significant moderators of the effect of incentive on IQ score change: $Q_{\text{between}}(3) = 6.16$, $P = 0.10$ and $Q_{\text{between}}(1) = 2.14$, $P = 0.14$, respectively.

Study 2. Consistent with Study 1, test motivation, measured by observer ratings of "thin-slice" video footage of boys taking a full-scale intelligence test, was lower among boys with lower IQ, even when controlling for the demographic variables of race, family structure, and family socioeconomic status (SES): partial $r = 0.25$, $P < 0.001$. Test motivation was also more variable among boys of below-average IQ: Levene's $F = 11.47$, $P < 0.001$ (Table S3).

We estimated a series of structural equation models to test whether the nonintellective traits contributing to test motivation confound the associations between IQ and important outcomes. As a preliminary step, we tested a model of the predictive validity of IQ without considering test motivation. We specified a model in which observed IQ scores were a function of a latent intelligence factor and measurement error (estimated as 1 minus the published reliability of the IQ scores), which adjusted effect sizes for attenuation attributable to measurement error. We set the latent intelligence variance to 1 for identification and regressed each outcome onto the intelligence factor. This model fit the data well: $\chi^2(1) = 0.002$, $P = 0.96$, comparative fit index (CFI) = 1.00, RMS error of approximation (RMSEA) = 0.00. As shown in Fig. 2, after controlling for demographic variables and adjusting for reliability, intelligence as assessed by IQ predicted academic performance in adolescence and cumulative years of education, current employment, and fewer criminal convictions in early adulthood.

We next tested for the confounding effect of test motivation by using a series of nested models. First, we tested the full

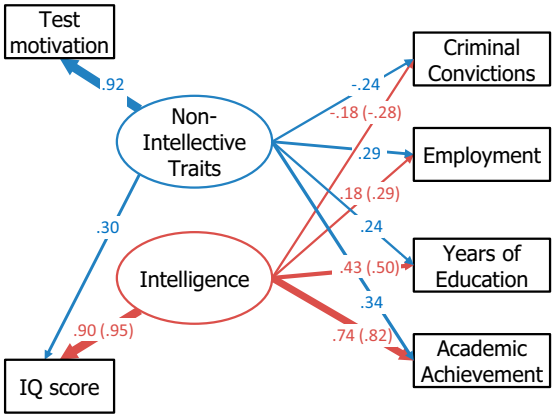

Fig. 2. Model tested in Study 2. Values are standardized regression coefficients, and parenthetical values are from the IQ-only model, which did not include test motivation or its associations with IQ scores and outcomes. The thickness of regression paths are proportional to standardized coefficients. All paths are significantly different from zero; $P < 0.05$. Demographic covariates and covariances among life outcomes are not shown.

hypothesized model: test motivation ratings were specified as a function of measurement error (estimated as 1 minus its interrater reliability) and a nonintellective latent factor, and IQ scores were specified as a function of both latent intelligence and the nonintellective traits that contribute to test motivation, in addition to error. The latent factor variances were again set to 1 for identification. The full model fit well: $\chi^2(1) = 0.03$, $P = 0.86$, CFI = 1.00, RMSEA = 0.00. Standardized regression coefficients are shown in Fig. 2 and Table S4. Both latent intelligence and nonintellective traits significantly predicted all four life outcomes, and nonintellective traits were significantly associated with IQ scores. As a test of the influence of the latent traits underlying test motivation, we compared the full model with a model in which paths from nonintellective traits to outcomes were constrained to equal zero. As expected, this model fit significantly worse than did the full model: $\Delta\chi^2(3) = 55.07$, $P < 0.001$, CFI = 0.83, RMSEA = 0.23. In sum, the nonintellective traits assessed by test motivation predicted both IQ scores and life outcomes associated with IQ scores. As summarized in Table 1, failing to account for the influence of motivation on IQ scores resulted in an overestimation of the association between latent intelligence and all four life outcomes.

The standardized regression coefficients presented in Fig. 2 suggested that intelligence was more strongly associated with academic outcomes, whereas nonintellective traits appeared more strongly associated with nonacademic outcomes (i.e., employment and fewer criminal convictions). A model in which the intelligence and nonintellective path coefficients were constrained to be equal showed that intelligence predicted academic achievement significantly better than did nonintellective traits, $\Delta\chi^2(1) = 12.12$, $P = 0.002$, CFI = 0.98, RMSEA = 0.12. In contrast, coefficients for intelligence and nonintellective traits could be constrained to be equal for the three other outcomes without worsening model fit, suggesting that the predictive validities for intelligence and nonintellective traits did not significantly differ: $\Delta\chi^2(1) \leq 2.52$, $P > 0.11$, CFI = 1.00, RMSEA ≤ 0.03.

Discussion

In Study 1, material incentives in random-assignment studies increased IQ scores by an average of 0.64 SD, suggesting that test motivation can deviate substantially from maximal under low-stakes research conditions. The effect of incentives was moderated by IQ score: Incentives increased IQ scores by 0.96 SD among individuals with below-average IQs at baseline and by only 0.26 SD among individuals with above-average IQs at baseline. Because no samples exceeded 120 in average baseline IQ, it is unlikely that this result can be explained by a ceiling effect on IQ-test performance. Further, homogeneity in the effect of incentives among samples of above-average IQ suggests that, in the absence of incentives, individuals perform closer to maximal potential than do individuals of below-average IQ.

In Study 2, observer ratings of test motivation were associated with both IQ scores and important life outcomes. Because children who tried harder on the low-stakes test earned higher IQ scores and also had more positive life outcomes, we tested for and found evidence that relying on IQ scores as a measure of intelligence may overestimate the predictive validity of intelligence. That is, nonintellective traits partially accounted for associations between IQ and outcomes. The seriousness of this confound was more profound (i.e., reductions in the proportion of variance explained by 68–84%) for the nonacademic outcomes of employment and crime than for the academic outcomes of school achievement in adolescence and years of education (i.e., reductions in the proportion of variance explained of 23–27%).

Despite efforts to "encourage in order that every one may do his best" on intelligence tests (ref. 41, p. 122), pioneers in intelligence testing took seriously the possibility that test takers might not, in fact, exert maximal effort. Thorndike, for instance, pointed out that although "all our measurements assume that the individual in question tries as hard as he can to make as high a score as possible . . . we rarely know the relation of any person's effort to his maximum possible effort" (ref. 42, p. 228). Likewise, Wechsler recognized that intelligence is not all that intelligence tests test: "from 30% to 50% of the total factorial variance [in intelligence test scores remains] unaccounted for. . . .this residual variance is largely contributed by such factors as drive, energy, impulsiveness, etc. . . ." (ref. 9, p. 444).

It is important not to overstate our conclusions. For all measured outcomes in Study 2, the predictive validity of intelligence remained statistically significant when controlling for the nonintellective traits underlying test motivation. Moreover, the predictive validity of intelligence was significantly stronger than was the predictive validity of test motivation for academic achievement. In addition, both Studies 1 and 2 indicate that test motivation is higher and less variable among participants who are above-average in measured IQ. These findings imply that earning a high IQ score requires high intelligence in addition to high motivation. Lower IQ scores, however, might result from either lower intelligence or lack of motivation. Thus, given closer-to-maximal performance, test motivation poses a less serious threat to the internal validity of studies using higher-IQ samples, such as college undergraduates, a popular convenience sample for social science research (43). Test motivation as a third-variable confound is also less likely when experimenters provide substantial performance-contingent incentives or when test results directly affect test takers (e.g., intelligence tests used for employment or admissions decisions).

On the other hand, test motivation may be a serious confound in studies including participants who are below-average in IQ and who lack external incentives to perform at their maximal potential. Consider, for example, the National Longitudinal Survey of Youth (NLSY), a nationally representative sample of more than 12,000 adolescents who completed an intelligence test called the Armed Forces Qualifying Test (AFQT). As is typical in social science research, NLSY participants were not rewarded in any way for higher scores. The NLSY data were analyzed in *The Bell Curve*, in which Herrnstein and Murray (44) summarily dismissed test motivation as a potential confound in their analysis of black–white IQ disparities.

Segal (45) subsequently reanalyzed the NLSY data, presenting evidence that performance on the coding speed subtest, the objective of which is to match 4-digit numbers to words by using a key

Table 1. Percentage of variance explained by IQ, intelligence, and nonintellective traits

Predictor	Academic performance in adolescence, %	Total years of education, %	Employment in adulthood, %	Lifetime convictions, age 26 y, %
IQ[a]	40.1	15.2	5.0	3.2
Intelligence[b]	31.0	11.1	1.6	0.5
Nonintellective traits	9.2	4.9	8.1	5.2

[a]IQ scores include variance attributable to intelligence and the nonintellective traits that influence test motivation.
[b]Intelligence is defined as variance in IQ scores independent of variance explained by the nonintellective traits that influence test motivation.

of number–word pairs, administered in the same session as the AFQT, is a good proxy for test motivation. Performance on the coding speed test demonstrates the lowest correlations with the AFQT and other IQ subtests administered in the same session (44) yet predicts earnings in adulthood over and beyond AFQT scores. Furthermore, in line with our finding that test motivation is lower and more heterogeneous among individuals of lower IQ, coding speed best predicts income among the least educated individuals in the NLSY sample.

Limitations and Future Directions. Limitations of the current investigation suggest profitable directions for future research. First, the Pittsburgh Youth Study sample used in Study 2 was socioeconomically and ethnically diverse but included only boys. Although we have no theoretical reason to suspect that test motivation is an important individual difference among males but not females, this assumption should be tested empirically. A second limitation of Study 2 is that its sample size did not allow sufficient power to test whether test motivation was a stronger confound of IQ–outcome relations among participants of below-average IQ. Finally, more precise measures of test motivation may have explained more variance in life outcomes—and pointed to an even larger confounding effect than was observed than the available thin-slice video ratings.

Future studies are needed to identify the traits that determine effort on low-stakes tests. More than 1,000 psychologists and educational specialists with expertise in intelligence testing rated the importance of six "personal characteristics" to performance on intelligence tests (46). On a 4-point scale where 1 = of little importance and 4 = very important, experts' ratings were as follows: attentiveness (M = 3.39, SD = 0.74), persistence (M = 2.96, SD = 0.87), achievement motivation (M = 2.87,SD = 0.96), anxiety (M = 2.68, SD = 0.90), emotional ability (M = 2.52, SD = 0.94), and physical health (M = 2.34, SD = 0.89). Consistent with these ratings, Borghans, Meijers, and Ter Weel (47) found that individuals higher in achievement motivation tend to think longer and are less sensitive to financial incentives when answering questions on an untimed intelligence test.

Future research should also examine cross-cultural differences in test motivation. A recent analysis of the Third International Mathematics and Science Study (TIMSS) suggested that test motivation accounts for a significant proportion of achievement differences between countries (48). The TIMSS examined math and science achievement among a half-million students from 41 nations in 1995. The relatively disappointing performance of the more than 33,000 American students who participated in the TIMSS has been widely publicized (e.g., American 12th graders earned among the lowest scores in both science and mathematics). A little-publicized TIMSS report revealed that test motivation, indexed as the proportion of optional self-report questions answered in the accompanying student background questionnaire, accounts for 53% of between-nation variability in math achievement, 22% of between-classroom variability within nations, and 7% of between-student variability within classrooms (48). These findings are consistent with the current investigation and further suggest that cross-cultural differences in test motivation may be even greater than individual differences among students within a particular culture.

Conclusion

The current investigation supports the hypothesized relations in Fig. 1. What do intelligence tests test? Both intelligence and test motivation. Why is this a problem? Because test motivation on low-stakes intelligence tests can partially confound IQ outcome associations.

Our conclusions may come as no surprise to psychologists who administer intelligence tests themselves (49). Where the problem lies, in our view, is in the interpretation of IQ scores by economists, sociologists, and research psychologists who have not witnessed variation in test motivation firsthand. These social scientists might erringly assume that a low IQ score invariably indicates low intelligence. As pioneers in intelligence testing pointed out long ago, this is not necessarily true.

Materials and Methods

Study 1. ***Sample of studies.*** In January 2008, we conducted a search of the PsycInfo database for articles containing at least one keyword from both of the following two lists: (*i*) intelligence, IQ, test performance, or cognitive ability and (*ii*) reinforcement or incentive. This search resulted in 1,015 articles and dissertations. We examined the abstracts of these publications by using the following inclusion criteria: (*i*) the article described an empirical study, (*ii*) the article used a between-subjects design with control and experimental groups, (*iii*) the experimental groups were rewarded with material incentives (e.g., money, tokens, candy) contingent on their intelligence-test performance, (*iv*) study participants did not meet diagnostic criteria for schizophrenia or other serious mental illness requiring inpatient care, and (*v*) study participants did not meet diagnostic criteria for mental retardation (i.e., study participants did not score below 70 on intelligence tests without incentives). No within-subject studies that counterbalanced incentive and control conditions met the other inclusion criteria. Therefore, within-subjects studies were excluded because the effect of incentives was not separable from the effect of practice.

The final sample comprised 19 published articles and 6 dissertations with 46 independent samples and 2,008 total participants. The included articles ranged in publication date from 1936 to 1994, and descriptions of study characteristics varied widely in level of detail. Consequently, participant age, baseline level of intelligence, and incentive size were coded as categorical variables (*SI Materials and Methods*). P.D.Q. coded all articles, and A.L.D. coded a random sample of 10% of the articles; interrater reliability was 100%.

Effect-size analyses. In all samples, the difference between intelligence-test scores for control (i.e., no incentive) and material incentive groups was the effect size of interest. We computed Hedge's g, the bias-corrected standardized mean difference, using random-effects models in Comprehensive Meta-Analysis (50). Hedge's g is interpreted similarly to Cohen's d but is corrected for bias attributable to small sample sizes (51). To compute the sample-size adjusted mean effect size, we used a random-effects model, which assumes that there is no single true population effect and allows for random between-sample variance in addition to error variance (51). We used mixed-effects models to independently test for the effect of four categorical moderators: level of baseline IQ, incentive size, study design, and age. We used control group scores to estimate baseline IQ in studies that did not report scores at baseline. See Table S1 for the raw effect sizes.

The full Study 1 data are available in Dataset S1.

Study 2. Participants were drawn from 508 boys in the middle sample of the Pittsburgh Youth Study (52) (*SI Materials and Methods*). At average age 12.5 y, ∼80% of these boys completed a short form of the Wechsler Intelligence Scale for Children–Revised (WISC-R) (11, 53, 54), the reliability of which has been estimated at 0.91 (55). During testing, a 15-min video sample of the boys' behavior was recorded to be coded by three different raters who were blind to the hypotheses of the study and to the IQ scores of the boys. Coders were trained to consensus (20 h) to observe and identify behaviors that indicated low motivation, including refusing to attempt tasks, forcing examiners to work hard to get them to try a task, expressing the desire for the testing session to end as quickly as possible, or responding very rapidly with "I don't know" responses (56). Scores were standardized within each rater and then averaged across all three raters (*SI Materials and Methods*). Intraclass correlations for each set of raters ranged from 0.85 to 0.89.

In early adulthood, ∼60% of these participants (n = 251) completed structured interviews assessing educational attainment, employment, and other outcomes and were therefore included in our analyses (*SI Materials and Methods*). Average age at follow-up was 24.0 y, SD = 0.91. In addition, lifetime criminal history data were obtained from government records for all participants when they reached age 26 y. In terms of measured IQ, the longitudinal sample was representative of the general population (mean IQ = 101.80, SD = 15.77). The men who participated in the follow-up interviews did not differ from those who did not on most study variables, but they were significantly higher in test motivation, performed better academically during adolescence, had fewer criminal convictions by age 26, came from higher-SES families, and were more likely to be Caucasian and from two-parent homes. These effects were small to moderate in size (*SI Materials and Methods*).

References 16–40 were studies included in the meta-analysis in Study 1.

ACKNOWLEDGMENTS. We gratefully acknowledge Elliot Tucker-Drob for guidance on statistical analyses and Rebecca Stallings for her extremely helpful guidance in using the data for Study 2. We also thank Avshalom Caspi, Terri Moffitt, and two anonymous reviewers for thoughtful comments on earlier drafts of the manuscript. This research was supported by National Institute of Mental Health Grant R01 MH45070 and National Institute on Aging Grants R01 AG032282 and K01-AG033182.

1. Gottfredson LS (2004) Intelligence: Is it the epidemiologists' elusive "fundamental cause" of social class inequalities in health? *J Pers Soc Psychol* 86:174–199.
2. Hogan R (2005) In defense of personality measurement: New wine for old whiners. *Hum Perform* 18:331–341.
3. Jensen AR (1998) *The G Factor: The Science of Mental Ability* (Praeger/ Greenwood, Westport, CT).
4. Judge TA, Colbert AE, Ilies R (2004) Intelligence and leadership: A quantitative review and test of theoretical propositions. *J Appl Psychol* 89:542–552.
5. Neisser U, et al. (1996) Intelligence: Knowns and unknowns. *Am Psychol* 51:77–101.
6. Roberts BW, Kuncel NR, Shiner R, Caspi A, Goldberg LR (2007) The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspect Psychol Sci* 2:313–345.
7. Sternberg RJ, Grigorenko EL, Bundy DA (2001) The predictive value of IQ. *Merrill-Palmer Q* 47:1–41.
8. Boring EG (1923) Intelligence as the tests test it. *New Repub* 35:35–37.
9. Wechsler D (1940) Nonintellective factors in general intelligence. *Psychol Bull* 37: 444–445.
10. Ackerman PL, Heggestad ED (1997) Intelligence, personality, and interests: Evidence for overlapping traits. *Psychol Bull* 121:219–245.
11. Wechsler D (1974) *Wechsler Intelligence Scale for Children–Revised* (The Psychological Corporation, San Antonio, TX).
12. MacNicol K (1960) *Effects of Varying Order of Item Difficulty in an Unspeeded Verbal Test* (Educational Testing Services, Princeton, NJ).
13. Revelle W (1993) Individual differences in personality and motivation: "non-cognitive" determinants of cognitive performance. *Attention: Selection, Awareness, and Control.A Tribute to Donald Broadbent*, eds Baddeley AD, Weiskrantz AD (Oxford Univ Press, Oxford), pp 346–373.
14. Arvey RD, Strickland W, Drauden G, Martin C (1990) Motivational components of test taking. *Person Psychol* 43:695–716.
15. O'Neil HF, Jr., Sugrue B, Baker EL (1995/1996) Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educ Assess* 3:135–157.
16. Benton AL (1936) Influence of incentives upon intelligence test scores of school children. *J Genet Psychol* 49:494–497.
17. Bergan A, McManis DL, Melchert PA (1971) Effects of social and token reinforcement on WISC block design performance. *Percept Mot Skills* 32:871–880.
18. Blanding KM, Richards J, Bradley-Johnson S, Johnson CM (1994) The effect of token reinforcement on McCarthy Scale performance for White preschoolers of low and high social position. *J Behav Educ* 4:33–39.
19. Bradley-Johnson S, Graham DP, Johnson CM (1986) Token reinforcement on WISC–R performance for White, low-socioeconomic, upper and lower elementary-school-age students. *J Sch Psychol* 24:73–79.
20. Bradley-Johnson S, Johnson CM, Shanahan RH, Rickert VI, Tardona DR (1984) Effects of token reinforcement on WISC-R performance of Black and White, low socioeconomic second graders. *Behav Assess* 6:365–373.
21. Breuning SE, Zella WF (1978) Effects of individualized incentives on norm-referenced IQ test performance of high school students in special education classes. *J Sch Psychol* 16:220–226.
22. Clingman J, Fowler RL (1976) The effects of primary reward on the I.Q. performance of grade-school children as a function of initial I.Q. level. *J Appl Behav Anal* 9:19–23.
23. Devers R, Bradley-Johnson S (1994) The effect of token reinforcement on WISC-R performance for fifth- through ninth-grade American Indians. *Psychol Rec* 44: 441–449.
24. Dickstein LS, Ayers J (1973) Effect of an incentive upon intelligence test performance. *Psychol Rep* 33:127–130.
25. Edlund CV (1972) The effect on the test behavior of children, as reflected in the I.Q. scores, when reinforced after each correct response. *J Appl Behav Anal* 5:317–319.
26. Ferguson HH (1937) Incentives and an intelligence test. *Australasian J Psychol Philos* 15:39–53.
27. Galbraith G, Ott J, Johnson CM (1986) The effects of token reinforcement on WISC-R performance of low-socioeconomic Hispanic second-graders. *Behav Assess* 8:191–194.
28. Gerwell EL (1981) Tangible and verbal reinforcement effects on fluid and crystallized intelligence in the aged. PhD dissertation (Hofstra University, Hempstead, NY).
29. Graham GA (1971) The effects of material and social incentives on the performance on intelligence test tasks by lower class and middle class Negro preschool children. PhD dissertation (George Washington University, Washington, DC).
30. Holt MM, Hobbs TR (1979) The effects of token reinforcement, feedback and response cost on standardized test performance. *Behav Res Ther* 17:81–83.
31. Kapenis JT (1979) The differential effects of various reinforcements and socioeconomic status upon Peabody Picture Vocabulary Test performance. PhD dissertation (University of South Dakota, Vermillion, SD).
32. Kieffer DA, Goh DS (1981) The effect of individually contracted incentives on intelligence test performance of middle- and low-SES children. *J Clin Psychol* 37: 175–179.
33. Lloyd ME, Zylla TM (1988) Effect of incentives delivered for correctly answered items on the measured IQs of children of low and high IQ. *Psychol Rep* 63:555–561.
34. Saigh PA, Antoun FT (1983) WISC-R incentives and the academic achievement of conduct disordered adolescent females: A validity study. *J Clin Psychol* 39:771–774.
35. Steinweg SB (1979) A comparison of the effects of reinforcement on intelligence test performance of normal and retarded children. PhD dissertation (University of North Carolina, Chapel Hill, NC).
36. Sweet RC, Ringness TA (1971) Variations in the intelligence test performance of referred boys of differing racial and socioeconomic backgrounds as a function of feedback or monetary reinforcement. *J Sch Psychol* 9:399–409.
37. Terrell F, Terrell SL, Taylor J (1980) Effects of race of examiner and type of reinforcement on the intelligence test performance of lower-class Black children. *Psychol Sch* 17:270–272.
38. Tiber N (1963) The effects of incentives on intelligence test performance. PhD dissertation (Florida State University, Tallahassee, FL).
39. Weiss RH (1981) Effects of reinforcement on the IQ scores of preschool children as a function of initial IQ. PhD dissertation (Utah State University, Logan, UT).
40. Willis J, Shibata B (1978) A comparison of tangible reinforcement and feedback effects on the WPPSI I.Q. scores of nursery school children. *Educ Treat Child* 1:31–45.
41. Binet A, Simon T, Kite ES (1916) *The Development of Intelligence in Children (The Binet Simon Scale)* (Williams & Wilkins, Baltimore).
42. Thorndike EL (1904) *An Introduction to the Theory of Mental and Social Measurements* (Science Press, Oxford).
43. Wintre MG, North C, Sugar LA (2001) Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Can Psychol* 42:216–225.
44. Herrnstein RJ, Murray CA (1994) *The Bell Curve: Intelligence and Class Struggle in American Life* (Free Press, New York).
45. Segal C (2006) Motivation, test scores, and economic success. Working Paper no. 1124 (Universitat Pompeu Fabra, Barcelona). Available at www.econ.upf.edu/docs/papers/downloads/1124.pdf.
46. Snyderman M, Rothman S (1987) Survey of expert opinion on intelligence and aptitude testing. *Am Psychol* 42:137–144.
47. Borghans L, Meijers H, Ter Weel B (2008) The role of noncognitive skills in explaining cognitive test scores. *Econ Inq* 46:2–12.
48. Boe EE, May H, Boruch RF (2002) *Student Task Persistence in the Third International Mathematics and Science Study: A Major Source of Achievement Differences at the National, Classroom, and Student Levels* (Center for Research and Evaluation in Social Policy, University of Pennsylvania, Philadelphia), Research Rep 2002-TIMSS1.
49. Haywood HC (1992) The strange and wonderful symbiosis of motivation and cognition. *Int J Cogn Ed Mediated Learn* 2:186–197.
50. Borenstein M, Hedges L, Higgins J, Rothstein H (2005) *Comprehensive Meta-Analysis* (Biostat, Englewood, NJ), Version 2.
51. Borenstein M, Hedges L, Higgins J, Rothstein H (2009) *Introduction to Meta-Analysis* (Wiley, West Sussex, UK).
52. Loeber R, Farrington DP, Stouthamer-Loeber M, Van Kammen WB (1998) *Antisocial Behavior and Mental Health Problems: Explanatory Factors in Childhood and Adolescence* (Lawrence Erlbaum Associates, Mahwah, NJ).
53. Hobby KL (1980) *WISC-R Split-Half Short Form Manual* (Western Psychological Services, Los Angeles).
54. Yudin LW (1966) An abbreviated form of the WISC for use with emotionally disturbed children. *J Consult Psychol* 30:272–275.
55. Silverstein AB (1990) Notes on the reliability of Wechsler short forms. *J Clin Psychol* 46: 194–196.
56. Lynam D, Moffitt TE, Stouthamer-Loeber M (1993) Explaining the relation between IQ and delinquency: Class, race, test motivation, school failure, or self-control? *J Abnorm Psychol* 102:187–196.