Running head: MOTIVATION AND INTELLIGENCE TESTING

What Intelligence Tests Test: Individual Differences in Test Motivation and IQ

Angela Lee Duckworth

University of Pennsylvania

Patrick D. Quinn

The University of Texas at Austin

Donald Lynam

Purdue University

Rolf Loeber, Magda Stouthamer-Loeber

University of Pittsburgh

Terrie E. Moffitt and Avshalom Caspi

Duke University and Institute of Psychiatry, King's College London

Abstract

The terms *IQ* and *intelligence* are often used synonymously because intelligence tests are widely assumed to measure maximal intellectual performance. The current investigation shows that this assumption is incorrect and suggests that individual differences in test-taking motivation reflect traits that predict the same important life outcomes as intelligence. In Study 1, a meta-analysis of random-assignment studies (total $N = 2{,}008$) testing the effects of material incentives on intelligence test performance demonstrated that incentives increase IQ scores by an average of 0.64 standard deviations. In Study 2, trained observers rated test motivation during an intelligence test for 251 adolescent boys who were later interviewed in adulthood. Test motivation provided incremental predictive validity over and beyond IQ for academic and non-academic outcomes. Because test motivation was moderately associated with IQ and predicted the same outcomes as did IQ, we tested and found evidence that test motivation partially accounted for IQ-outcome relations. Test motivation was related to parent ratings of Big Five Conscientiousness, Agreeableness and Openness to Experience, but these factors only partially explained the effect of test motivation on life outcomes.

What Intelligence Tests Test: Individual Differences in Test Motivation and IQ

IQ scores predict a range of life outcomes, including academic performance, years of education, and job performance (Hogan, 2005; Jensen, 1998; Judge, Colbert, & Ilies, 2004; Neisser et al., 1996; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). Gottfredson (1997) has argued that intelligence causally determines performance across diverse domains because nearly every task in life requires processing and using information of some complexity. But what is intelligence? Boring's (1923) 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" (p. 35). Boring's early comment augured the now widespread and unfortunate conflation of the terms *IQ* and *intelligence.* We suggest that the distinction between the manifest variable of IQ and the unobserved latent construct of intelligence[1] is of critical importance because IQ scores are determined not only by intelligence but also by test motivation.

The question of whether IQ scores are partly determined by test motivation is a neglected issue in contemporary psychology: "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" (Revelle, 1993, pp. 352-353). This stance is surprising given that the earliest intelligence researchers explicitly acknowledged the problem. Thorndike (1904), for instance, conceded that whereas "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" (p. 228).

Intelligence tests are designed to measure maximal performance, but test design alone cannot guarantee that all examinees put forth their best possible performance. Indeed, the need for explicit directions to reduce outside distractions and to buoy interest and engagement give lie to the assumption that examinees invariably try their best. For instance, directions from the WISC-III manual imply that active intervention might be necessary to sustain test motivation in many examinees: "If the child says that he or she cannot perform as task or cannot answer a question, encourage the child by saying, "Just try it" or "I think you can do it. Try again." (p. 37). Similarly, the deliberate sequencing of items from easy to difficult is an explicit strategy for sustaining morale (MacNicol, 1960).

The gap between tested and maximal performance on any task can be substantial (Cronbach, 1960; Sackett, Zedeck, & Fogli, 1988), and conflating observed with maximal performance can lead to erroneous conclusions. For example, a longitudinal study by Zigler and Butterfield (1968) examined the effects of an intervention (i.e., nursery school) on measured IQ in a sample of low-income children. By administering the same intelligence test under typical and incentivized conditions, Zigler and Butterfield discovered that the intervention had a beneficial effect on motivation but not on intelligence. Specifically, the observed benefits of intervention in comparison to a no-treatment control group were large (nearly a standard deviation in magnitude) when the Stanford-Binet intelligence test was given using standard instructions, but non-existent when motivation to perform well was maximized by giving children a series of very easy questions prior to more difficult ones. Zigler and Butterfield concluded, "Once one recognizes that the performance deficit in disadvantaged children, instead

of being invariably due to a cognitive deficit, may be due to a variety of motives, attitudes, general approaches to tasks, and even to psychological defenses that tend to be defeating in the school setting but are generally adaptive to the child's life, then one is ready to seriously entertain the proposition that disadvantaged children are brighter than their test scores indicate" (p. 302).

Theoretically, there are four scenarios in which test motivation is unimportant. See **Figure 1** for a decision tree of these possibilities (Scenarios I through IV). First, test motivation could be close to maximal for all subjects taking an IQ test. Second, test motivation could be constant (i.e., homogenous) across the sample yet below maximum (e.g., all of the participants are exerting 70% of maximal effort). Third, test motivation could be completely stochastic (i.e., randomly distributed in the sample and uncorrelated within individuals across time). Fourth, test motivation variance could differ across sample members (i.e., be heterogeneous) but be unrelated to any traits that predict important outcomes. That is, some individuals might try harder than others on intelligence tests, but the traits that determine effort on tests may be unimportant to any consequential life outcomes. In Scenario V, test motivation is systematic, heterogeneous in the sample, and determined by traits that predict important life outcomes. It is this fifth possibility that we consider the most likely and for which we find empirical support in the present investigation.

Which personal characteristics might influence test motivation? Over 1,000 psychologists and educational specialists with expertise in intelligence testing rated the importance of six traits to performance on intelligence tests (Snyderman & Rothman, 1987). Using a 4-point scale where 1 was *of little importance* and 4 was *very important*, three traits related to the Big Five personality factor of Conscientiousness received the highest ratings: attentiveness (M = 3.39, SD = .74), persistence (M = 2.96, SD = .87), and achievement motivation (M = 2.87, SD = .96). Big

Five Conscientiousness has in many other studies demonstrated predictive validity over and beyond measured intelligence for academic achievement (Conard, 2005; Noftle & Robins, 2007) and job performance (Mount, Barrick, & Strauss, 1999). IQ experts also rated two facets of Big Five Neuroticism as at least somewhat important to test performance: anxiety (M = 2.68, SD = .90) and emotional lability (M = 2.52, SD = .94). Big Five Neuroticism also predicts life outcomes, though in general not as strongly as does Conscientiousness (Barrick & Mount, 1991; Roberts et al., 2007). Of the six rated personal characteristics, physical health was considered the least relevant to intelligence test performance, M = 2.34, SD = .89. The survey did not include aspects of Big Five Agreeableness (e.g., compliance with authority, trust, and cooperativeness) or Big Five Openness to Experience (e.g., creativity, intellectual curiosity), though these might also be expected to encourage effort on a test where no extrinsic incentives are provided.

In Study 1, we tested the hypothesis that motivation during intelligence testing is heterogeneous and systematic (Scenarios III, IV, and V) rather than maximal or sub-maximal but homogenous (Scenario II). Specifically, we conducted a meta-analysis of studies using random-assignment to measure the effect of material incentives on intelligence test performance. In Study 2, we tested whether test motivation predicts the same outcomes predicted by intelligence and whether test motivation accounts for IQ-outcome associations. In a longitudinal study of 251 boys followed from adolescence to early adulthood, we compared the predictive validity of observer ratings of test motivation to that of measured IQ for academic (i.e., school performance in adolescence and total years of education) and non-academic (i.e., employment and criminal behavior) outcomes. In addition, we examined relations between test motivation and concurrent parent ratings of personality to determine the extent to which test motivation reflects Big Five factors.

Study 1

The assumption that test motivation is uniformly maximal among all test takers predicts that material incentives, such as cash rewards, should have no effect on intelligence test performance. Study 1 was a meta-analysis of studies using random-assignment designs to test the effect of material incentives on intelligence test performance. We tested baseline IQ, incentive size, age, and study design as potential moderators and examined remaining heterogeneity in effect size once moderators were accounted for.

Method

Sample of Studies

In January 2008, we conducted a search of the PsycInfo database for articles examining the effects of incentives on intelligence test performance. Specifically, we searched for articles containing at least one keyword from both of the following two lists: (a) *intelligence*, *IQ*, *test performance*, or *cognitive ability* and (b) *reinforcement* or *incentive*. This search resulted in 1,015 articles and dissertations. We examined the abstracts of these publications using the following inclusion criteria: (a) the article described an empirical study, (b) the article used a between-subjects design with control and experimental groups,[2] (c) the experimental groups were rewarded with material incentives (e.g., money, tokens, candy) contingent on their intelligence test performance, (d) study participants did not meet diagnostic criteria for schizophrenia or other serious mental illness requiring inpatient care, and (e) study participants did not meet diagnostic criteria for mental retardation (i.e., study participants did not score below 70 on intelligence tests without incentives). The final sample comprised 19 published articles and 6 dissertations with 46 distinct samples and 2,008 total participants.

Coded Variables

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. The second author coded all articles. The first author coded a random sample of 10% of the articles; inter-rater reliability was 100%.

Age. Participant age was frequently reported as a range of grade levels or years, so we treated age as a categorical variable where 1 = *5 years old or younger*, 2 = *age 6 to 11*, 3 = *age 12 to 18*, and 4 = *19 or older*.

Article type. We recorded type of article (dissertation vs. published article).

IQ score. We recorded control and experimental group intelligence test means and standard deviations. Where available, baseline scores were also recorded.

IQ measure. Measures of intelligence included the Lorge-Thorndike Intelligence Test (LTIT), the McCarthy Scales of Children's Abilities (MSCA), the Otis-Lennon Mental Ability Test (OLMAT), the Otis Self-Administering Test (OSAT), the Peabody Picture Vocabulary Test (PPVT), Raven's Progressive Matrices (RPM), the Stanford-Binet Intelligence Scale, Third Revision (SBIS-III), the Wechsler Adult Intelligence Scale (WAIS), the Wechsler Intelligence Scale for Children (WISC), the Wechsler Intelligence Scale for Children, Revised (WISC-R), and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI). Whereas most measures of intelligence were administered individually, the OSAT, LTIT, and OLMAT were administered in group settings.

Incentive size. In order to take into account the effect of inflation on the value of incentives, we converted incentives to 2007 dollar equivalents. We treated the value of

incentives participants received for their intelligence test performance as a categorical variable where 1 = *less than $1*, 2 = *between $1 and $9*, and 3 = *$10 or more*.

Study design. We recorded whether the study design included baseline and post-test IQ measures or only a post-test measure.

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 Comprehensive Meta-Analysis (2005). Hedge's *g* is interpreted similarly to Cohen's *d* but is corrected for bias due to small sample size (Borenstein, Hedges, Higgins, & Rothstein, 2008). Where reported, means and standard deviations were used to calculate *g*. When these were not available, we calculated effect sizes from *t*-scores or from raw data. When pre- and post-test scores were available, we calculated *g* as the difference between control and incentive groups on mean change scores. When only post-test scores were reported, we calculated *g* as the difference between control and incentive group means.

Of the 25 articles included in the meta-analysis, 14 presented results for more than 1 sample. We treated individual samples as the basic unit of our analyses, giving us a *k* of 46 samples and an aggregated sample size of $N = 2{,}008$. When individual samples were tested on multiple measures of IQ, we computed a mean effect across measures within sample for use in our analyses. To compute the sample-size adjusted mean effect size, we used a random-effects model, which assumes that there is not one true population effect and allows for random between-sample variance in addition to error variance (Borenstein et al., 2008). See **Table 1** for the raw effect sizes.

Moderator analyses. We used mixed-effects models to independently test for the presence of four moderators: level of baseline IQ, incentive size, study design, and age. Unfortunately, we were unable to test for moderation by type of IQ measure because of an insufficient number of studies using each type of test. We were also unable to test for moderation by administration type (individual vs. group) because of an insufficient number of group-administered tests. The mixed-effects model relies upon the same assumptions as the random-effects model within each level of the moderator. That is, at each level of the moderator, there is random variation in the distribution of effect sizes. However, in comparing samples across levels of the moderator, the mixed-effects model assumes that the moderator is associated with systematic differences among effect sizes (Borenstein et al., 2008).

Results

Random-Effects Analyses

Material incentives raise IQ scores. In 46 samples (N = 2,008), the average effect of incentives on IQ was medium-to-large, g = 0.64 (95% CI: 0.39, 0.89), $p < .001$. An examination of **Table 1**, which lists the raw effect size from each sample, 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), $p < .001$.

Moderator Analyses

A test of heterogeneity among all 46 samples indicated that between-study variance accounted for 85% of the variance in effect sizes, $Q(45) = 303.68$, $p < .001$, $I^2 = 85.18$.

Moderator analyses indicated that baseline IQ score and incentive size, but not age nor study design, explained significant portions of this heterogeneity.

Because baseline IQ scores were reported as ranges in some samples, we created a binary variable where 1 = *below average* (i.e., IQ < 100) and 2 = *above average* (i.e., IQ ≥ 100). In studies that did not report baseline IQ, we used control group scores to estimate baseline intelligence. The effect of incentives was greater for individuals of below-average baseline IQ, $Q_{between}(1) = 9.76$, $p = .002$. 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). Moderation by baseline IQ did not account for all heterogeneity in effect size among low-IQ samples, $Q(22) = 226.23$, $p < .001$, $I^2 = 90.28$. In contrast, heterogeneity in effect size among high-IQ samples was not significantly different from zero, $Q(22) = 24.37$, $p = .33$, $I^2 = 9.71$. See **Table 2**.

As predicted, a systematic dose-response relationship was observed between incentive size and IQ score gain, $Q_{between}(2) = 28.95$, $p < .001$. Excluding 3 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, $Q_{between}(3) = 6.16$, $p = .10$ and $Q_{between}(1) = 2.14$, $p = .14$, respectively.

Publication Bias

Three of four tests indicated no publication bias. First, we followed Borenstein and colleagues (2008) in testing the relationship between sample size and effect size. Studies of

smaller sample size, and therefore larger standard error, should have more effect size variability among them. However, if studies of larger standard error and smaller effect size are not published (i.e., are hidden in the file drawer), there will be a weaker relationship between large standard error and variability. Egger's (1997) regression intercept was not significant (0.01, p = .99), suggesting a lack of publication bias. Second, we conducted two fail-safe N analyses. Rosenthal's (1979) fail-safe N indicated that an additional 1,885 samples of average effect size g = 0 would be required to eliminate the significance of the mean effect. Similarly, according to Orwin's (1983) fail-safe N, an additional 101 samples of effect size g = 0 would be needed to reduce the medium-to-large effect we found to a small effect of g = 0.2. Third, we conducted Duval's and Tweedie's (2000) trim and fill. The trim and fill, which adjusts the distribution of effect sizes to account for bias, found that no adjustment was necessary.

Only one of four analyses suggested a bias in the included samples. Specifically, there was evidence that article type (published article vs. dissertation) moderated the effect of incentives on IQ scores. In a mixed-effects analysis, the mean effect size in 33 published samples, g = 0.76 (95% CI: 0.46, 1.05), was significantly larger than that in 13 dissertation samples, g = 0.21 (95% CI: -0.07, 0.49), Q_{between} (1) = 7.02, p = .01. However, because we were unable to control for the simultaneous effects of other moderators, this test was of limited utility. In particular, it is possible that the lower effect size in dissertation samples can be explained by the higher proportion of high-IQ samples acquired from dissertations (69%) than from published articles (42%).

Discussion

The assumption that test motivation is always maximal during intelligence tests is incorrect. Rewarding subjects according to their performance with material incentives (e.g.,

cash) reliably increases performance by an average of 0.64 standard deviations. A systematic dose-response relationship between reward size and IQ-score improvement suggests that larger incentives than the nominal amounts typically employed in research studies might reveal even larger disparities between observed and maximal performance. The effect of incentives was moderated by IQ score: Incentives increased IQ scores by 0.96 standard deviations among individuals with above-average IQs at baseline and by only 0.26 standard deviations among individuals with below-average IQs at baseline. Further, homogeneity in the effect of incentives among individuals of above-average IQ suggests that at baseline (i.e., in the absence of incentives) they perform closer to maximal potential than do individuals of below-average IQ.

Study 2

In Study 1, we found that material incentives can substantially improve test scores, particularly among individuals with lower IQs at baseline. This finding argues against Scenarios I, II, and III in Figure 1 and suggests that test motivation is less than maximal, heterogeneous in the population, systematic, and either unrelated (Scenarios IV) or related (Scenario V) to life outcomes. To distinguish between these latter two possibilities, we used in Study 2 longitudinal data from the Pittsburgh Youth Study. We tested whether test motivation predicts the same life outcomes as does IQ and whether inferred associations between intelligence and outcomes are therefore spuriously inflated when test motivation is not measured and controlled. In addition, we examined the relation of test motivation to Big Five personality ratings provided by caregivers and assessed whether the predictive validity of test motivation for outcomes could be explained by associations with Big Five traits. Finally, we sought confirmation of the finding in Study 1 that test motivation is more heterogeneous and lower on average among individuals of lower IQ.

Method

Participants and Procedure

Subjects were boys in the middle sample of the Pittsburgh Youth Study. Details on the initial recruitment and screening of this sample in 1987-1988 when children (all male) were aged 10 are given in Loeber, Farrington, Stouthamer-Loeber, and van Kammen (1998). Briefly, the sample includes boys randomly selected from public schools in Pittsburgh, Pennsylvania. Of families contacted, 85% of the boys and their parents agreed to participate. About 50% of the final sample ($N = 508$) was identified as at-risk based on prior evidence of disruptive behavior problems. Fifty-four percent of the sample was Black, 43% was White, and the remaining 3% were Hispanic, Asian, or of mixed ethnicity.

The sample was followed from age 10 to 13 years; at age 12.5 years, about 80% of these boys completed an individually-administered intelligence test during which their behavior was videotaped for later coding. In terms of measured IQ, the sample was representative of the general population (mean IQ = 101.80, $SD = 15.77$). About 60% of these participants ($N = 251$) were interviewed in young adulthood (average age at follow-up was 24.0 years, $SD = 0.91$). The men who participated in these follow-up interviews did not differ from those who did not in IQ, $t(427) = 1.73$, $p = .09$, $d = .17$. At follow-up, participants did not differ from non-participants on years of education, $t(295) = 0.05$, $p = .96$, $d = .01$, current employment, $\chi^2(1) = 1.79$, $p = .18$, 12-month history of unemployment ($OR = 0.65$, $p = .16$), or additional arrests, $OR = 0.99$, $p = .98$. However, participants were rated higher in test motivation, $t(418) = 3.03$, $p = .003$, $d = .30$; they also performed better in school during adolescence, $t(505) = 3.03$, $p = .003$, $d = .27$, were higher on the Hollingshead (1975) two-factor socioeconomic status (SES) index at age 12.5 years, $t(481) = 2.91$, $p = .003$, $d = .27$, and were more likely to be Caucasian, $\chi^2(1) = 17.33$, $p < .001$, and

from two-parent homes, $\chi^2(1) = 19.01$, $p < .001$.

Measures

IQ. Participants completed a short form of the WISC-R (Wechsler, 1974). In this version, all 12 subtests were administered, but individual subtests were shortened by administering every other item. This procedure follows those used by Hobby (1980) and Yudin (1966) who reported correlations between the short form and full intelligence test of $r = .97$ for Full-Scale IQ scores. Trained testers administered the following set of instructions to each participant individually: "I'll be asking you to try a lot of different questions and puzzles. Some are like school work, most are not. Each task will only last about 3 minutes, so if you don't enjoy a task, don't worry, we will be switching to a different task soon. Each task asks you to do something different, because everybody has things they do well and things they don't do so well, and we want everyone to have a chance to succeed and have fun. Each task starts out easy and gets harder, and the questions go all the way up to college level for kids much older than you, so don't be surprised when you get some wrong. It is important to do your very best, the very best you can."

Test motivation. Three different raters coded 15 minutes of videotaped behavior for "impatience/impersistence" during the intelligence test (Lynam, Moffitt, & Stouthamer-Loeber, 1993). Raters were blind to both the boys' risk status and the hypotheses of the study. The raters were trained to consensus (20 hours) to identify behaviors such as refusing to attempt tasks, forcing examiners to work hard to get them to try a task, wanting the testing session to end as quickly as possibly, or responding rapidly with "I don't know" responses. Raters gave each boy a single rating using a standardized coding system where 3 = *severe*, 2 = *moderate*, 1 = *mild*, and 0 = *absent*. Scores were standardized within rater and then averaged across all three raters. Intraclass correlations for each set of raters ranged from .85 to .89. We reverse scored test

motivation so that higher scores indicated more motivation, and we used a natural log transformation to correct for right-skew before completing analyses.

Demographics. Five demographic variables were included as covariates in all analyses: race (1 = *Black* versus 0 = *White or other ethnicity*), family structure (0 = *two-parent* versus 1 = *not*), family SES, and age at the follow-up interview. Family SES was assessed using Hollingshead's two-factor index. If a boy had both a male and female parent or caretaker, scores were averaged; if he only had one caretaker, that score was used.

Big Five personality ratings. On the same day that participants completed an Intelligence test, caregivers completed the Common Language version of the California Child Q-Set (CCQ: Caspi et al., 1992). The CCQ uses a q-sorting procedure involving a set of rules for assigning scores to a set of 100 items describing a wide range of behaviors. John et al. (1994) constructed Big Five scales and provided evidence for their validity.

School performance in adolescence. Every fall and spring from age 10 to 13, teachers of participants completed the Teacher Report Form (TRF), the teacher version of the Child Behavior Checklist (CBCL: Achenbach & Edelbrock, 1983). Four items on the TRF inquired about the boy's performance in reading, writing, spelling, and math using a 5-point scale where 1 = *far below his grade level* and 5 = *far above his grade level*. At each assessment point, a summary score was computed as the average of these four items. The reliability of each summary score was high, α's ranged from .93 to .96. The average correlation among these summary scores over time was $r = .62$, suggesting reasonable cross-time stability. For each participant, a composite school performance score was computed as the mean of summary scores from age 10 to 13.

Years of education. At follow-up interviews in young adulthood, participants reported the highest grade level of education completed.

Employment. We measured employment in two ways. First, participants reported whether they were working for pay at the time of the follow-up interviews in young adulthood. Second, in follow-up interviews, each participant was also asked questions whose responses were recorded in a Life History Calendar (LHC: Caspi, Moffitt, Thornton, & Freedman, 1996). The interviewer asked the participant to detail life events since the age of 16. The calendar was divided into three-month periods corresponding to winter, spring, summer, and fall over the prior six years. To help participants remember what was happening at those times, the interviewer began by listing where the participant was living during these periods and then their best friend during each of these times. For each three-month period, participants were asked if they were "unemployed and looking for work and/or registered with unemployment office and getting unemployment checks." We recorded the number of seasons in the past 12 months that each participant reported being unemployed, which ranged from 0-4 seasons.

Number of arrests. In the LHC, each participant was also asked for each three-month period over the prior six years if he had been "arrested by a policeman, even if it did not result in formal charges or conviction."

Results

Summary statistics and zero-order correlations for test motivation, IQ, and demographic and outcome variables are given in **Table 3**. As predicted, test motivation and measured IQ were associated, even when controlling for demographic variables of race, family structure, and family SES, partial $r = 0.25$, $p < .001$.

Test Motivation Predicted Life Outcomes Over and Beyond IQ

To assess the effect of test motivation and IQ on life outcomes, both independently and in combination, we fit a series of three simultaneous multiple regression models for each of four measured outcomes. All regression models controlled for the demographic variables of race, family structure, and family SES; prospective models also controlled for age at follow-up interview. For each outcome, two separate models were fit including either IQ or test motivation, and a third model included both IQ and test motivation.

IQ and test motivation independently predicted all four life outcomes, and the predictive validity of test motivation remained significant when controlling for IQ. Specifically, when controlling for demographic variables, school performance in adolescence was associated with test motivation ($\beta = 0.27$, $p = .001$), and, separately, with IQ ($\beta = 0.71$, $p < .001$). When both predictors were entered simultaneously, test motivation and IQ both remained significant, $\beta = 0.12$, $p = .01$ and $\beta = 0.68$, $p < .001$, respectively. **See Table 4.** Similarly, cumulative years of education at follow-up were predicted by test motivation ($\beta = 0.22$, $p < .001$) and, separately, by IQ, $\beta = 0.43$, $p < .001$. When entered simultaneously, both test motivation ($\beta = 0.13$, $p = .04$) and IQ ($\beta = 0.39$, $p < .001$) remained significant predictors. See **Table 5**.

Current employment in adulthood was a binary variable, and prior to fitting a binary logistic regression model, we standardized all continuous predictor variables to facilitate interpretation of odds ratios. Excluding participants who were in college at the time of the interview and controlling for demographic variables, test motivation predicted current employment at follow-up ($OR = 1.69$, $p < .001$), as did IQ, $OR = 1.87$, $p = .002$. When entered simultaneously, both test motivation ($OR = 1.54$, $p = .007$) and IQ ($OR = 1.63$, $p = .02$) remained significant predictors. See **Table 6**.

Twelve-month unemployment history was a positively skewed ordinal variable. We fit an ordinal logistic regression model and again standardized all continuous predictors to ease interpretation of odds ratios. Controlling for demographic variables, test motivation predicted 12-month unemployment history ($OR = 0.71$, $p = .02$), as did IQ, $OR = 0.64$, $p = .02$. When entered simultaneously, both test motivation ($OR = 0.77$, $p = .08$) and IQ ($OR = 0.69$, $p = .07$) marginally predicted unemployment history. See **Table 7**.

Number of arrests was both positively skewed and zero-inflated (i.e., 61% of men had never been arrested). We therefore fit a zero-inflated Poisson (ZIP) regression model predicting number of arrests. The ZIP model is a mixture model, which approximates the distribution of the dependent variable by combining two distributions: a logistic regression distribution and a Poisson distribution (Atkins & Gallop, 2007). Regression coefficients for both the logistic (testing zero vs. any arrests) and count (testing increase in risk for additional arrests given one arrest) portions of three ZIP models are presented in **Table 8**. We again standardized all continuous predictors to facilitate the interpretation of odds ratios. No measured variables, including demographic characteristics, predicted zero vs. any arrests in any ZIP model. However, test motivation ($OR = 0.68$, $p < .001$) and IQ ($OR = 0.65$, $p = .001$) separately predicted fewer additional arrests. When entered simultaneously, test motivation ($OR = 0.71$, $p < .001$) and IQ ($OR = 0.76$, $p = .04$) remained significant predictors.

Predictive Validities of Test Motivation and IQ Were Comparable for Non-Academic Outcomes

The effects of measured IQ and test motivation on the non-academic outcomes of employment and additional arrests were all small-to-medium in size: IQ explained 3% of the variance in current employment, 2% of the variance in history of unemployment, and 1% of the variance in additional arrests when controlling for test motivation. Test motivation explained 4 to

7% of the variance in current employment, 1 to 3% of the variance in history of unemployment, and approximately 2% of the variance in additional arrests.[3] See **Table 9**. In contrast, for both academic outcomes, the estimated effect sizes for the predictive validity of measured IQ, even when controlling for test motivation, were medium-to-large, whereas the effects of test motivation were only small-to-medium. Measured IQ accounted for 34% of the variance in academic achievement and 10% of the variance in years of education when controlling for test motivation. In contrast, test motivation accounted for only 1 to 7% of the variance in academic achievement and 1 to 5% of the variance in years of education.

Test Motivation Partially Accounted for the Relationship Between IQ and Life Outcomes

Third-variable confounding among linear or binary variables can be tested with the same techniques used to test mediation (i.e., MacKinnon, Krull, & Lockwood, 2000). Using the Sobel (1982) test, we found that test motivation partially accounted for the relationships between IQ and school performance, years of education, and employment. The magnitude of the confound was larger for non-academic life outcomes. When test motivation was included in the regression model predicting school performance, the standardized regression coefficient for IQ decreased 4% from $\beta = 0.71$ to $\beta = 0.68$, $Z = 2.19$, $p = .03$. When test motivation was included in the model predicting years of education, the standardized regression coefficient for IQ decreased 9% from $\beta = 0.43$ to $\beta = 0.39$, $Z = 1.91$, $p = .056$. When test motivation was included in the binary logistic regression model predicting employment, the unstandardized regression coefficient for IQ decreased 21% from $\beta = 0.62$ to $\beta = 0.49$, $Z = 2.34$, $p = .02$. When test motivation was included in the ordinal logistic regression model predicting 12-month unemployment, the unstandardized regression coefficient for IQ decreased 18% from $\beta = -0.45$ to $\beta = -0.37$, $Z = 1.62$, $p = .10$.

When test motivation was included in the count portion of the ZIP model predicting

additional arrest, the unstandardized regression coefficient for IQ decreased 19% from β = -0.42 to β = -0.34. Thus, in magnitude the confounding effect of test motivation seemed largest for additional arrests, but since the Sobel test is not appropriate for relationships estimated using the ZIP model, we could not perform a formal test of statistical significance (see Pederson & McCarthy, 2008).

Big Five Personality Ratings Did Not Fully Explain the Prediction of Outcomes by Test Motivation

As shown in **Table 10**, children who tried harder on the intelligence test were more conscientious ($r = .15$, $p < .05$), more open to experience ($r = .15$, $p < .05$) and more agreeable, $r = .13$, $p < .05$. When controlling for IQ, however, only the association between test motivation and agreeableness reached statistical significance, partial $r = .14$, $p < .05$. Thus, regardless of measured IQ, boys who were more agreeable, as rated by their mothers, tried harder on the intelligence test. The causal relation between test motivation and the personality factors of conscientiousness and openness to experience is less clear, however, given that measured IQ accounts for a substantial proportion of their shared variance.

To test whether the associations between test motivation and life outcomes were attributable to variance in the three Big Five personality factors associated with test motivation, we regressed test motivation on Agreeableness, Conscientiousness, and Openness to Experience and saved the standardized residuals; this served to create a test motivation variable free from its overlap with the three Big Five factors. We then replaced test motivation with the residuals in the above regression models. The overall pattern of findings suggests that the effects of test motivation on outcomes were only partially explained by these Big Five personality measures. The test motivation residuals significantly predicted fewer additional arrests ($OR = 0.74$, $p <$

.001) and employment (*OR* = 1.45, *p* = .02), marginally predicted school performance in adolescence (β = 0.08, *p* = .07), and did not significantly predict 12-month unemployment (*OR* = 0.78, *p* = .11) or years of education, β = 0.09, *p* = .14. See **Table 9**.

Test Motivation Is Lower and More Heterogeneous Among Boys of Below-Average IQ

Study 1 demonstrated that material incentives were less effective at raising IQ scores among above-average-IQ individuals, suggesting that test motivation is closer to maximal among those individuals. Consistent with this finding, boys of below-average IQ in Study 2 were rated lower in test motivation than were boys of above-average IQ, $t(249) = 3.41$, $p < .001$, $d = .43$. There was significantly more variance in test motivation among boys of below-average IQ than among boys of above-average IQ, Levene's $F = 11.47$, $p < .001$. Unfortunately, our sample size ($N = 251$) did not allow sufficient power to test whether test motivation served as a differential confound across the range of IQ scores (see Preacher, Rucker, & Hayes, 2007).

Discussion

In Study 2, observer ratings of test motivation provided incremental predictive validity over and beyond measured IQ for important life outcomes. Because children who tried harder on the IQ test also earned higher IQ scores, we tested and found evidence that test motivation partially accounted for associations between IQ and outcomes (Scenario V in Figure 1). The magnitude of this confound was larger for the non-academic outcomes of employment and arrests than for the academic outcomes of school performance in adolescence and cumulative years of education. Consistent with the meta-analysis in Study 1, test motivation was lower and more heterogeneous among boys of below-average IQ. Big Five Conscientiousness, Agreeableness, and Openness to experience were associated with test motivation, but these Big Five factors did not fully account for the effect of test motivation on outcomes.

General Discussion

Wechsler (1940) was among the first to recognize that intelligence is not all that intelligence tests test: "from 30% to 50% of the total factorial variance" in test scores cannot be accounted for by factors that represent recognizable intellectual factors…this residual variance is largely contributed by such factors as drive, energy, impulsiveness, etc…" (p. 444). Our findings affirm that test motivation contributes to intelligence test performance and, moreover, constitutes a consequential individual difference in its own right (i.e., our data support Scenario V in Figure 1). In Study 1, a meta-analysis of random assignment studies assessing the impact of material incentives on test performance, we found that boosting extrinsic motivation can increase IQ scores more than half a standard deviation. Both Study 1 and Study 2 suggest that departures from maximal effort are greater and more heterogeneous among individuals of below-average IQ. Study 2 further demonstrated that test motivation variance is not only heterogeneous and systematic but also predictive of the same important life outcomes as IQ and partially accounts for IQ-outcome relations. The seriousness of this confound was more profound (i.e., decrements in regression coefficients of 18 to 21%) for the non-academic outcomes of employment and crime than for the academic outcomes of school achievement in adolescence and years of education (i.e., decrements in regression coefficients of 4 to 9%).

Individuals who try harder on intelligence tests tend to be higher in Big Five Conscientiousness (e.g., responsible, productive, thorough), Agreeableness (e.g., rated highly by their parents on traits such as trustful and likeable), and Openness to Experience (e.g., intellectual, imaginative, and curious). However, parent ratings of these Big Five factors did not fully account for the variance explained in life outcomes by test motivation. We suspect that specific *facets* of Conscientiousness (e.g., achievement motivation), Agreeableness (e.g.,

compliance with authority), and Openness (e.g., curiosity about puzzles and other analytic problems) not measured in our investigation might have explained better than did the available Big Five measures why some boys tried harder in the IQ test session. Generally, more narrowly defined facets of personality than those captured in Big Five measures can predict specific criteria substantially better than can broad factor measures (Paunonen & Ashton, 2001a; 2001b). Consistent with this possibility are the findings of Borghans, Meijers, and ter weel (2008) who conducted a within-subject study using an untimed IQ test and found that undergraduates higher in achievement motivation, curiosity, and internal locus of control were more likely to answer correctly and to invest more time solving each test item. Moreover, participants higher in these traits were less affected, in terms of time invested solving each item, by cash incentives to increase motivation. More research is needed to determine which facets of Big Five Conscientiousness, Agreeableness, and Openness—and perhaps other individual differences related to motivation and interest—determine effort on tests with no obvious external incentive.

The current investigation underscores the distinction between the latent construct of intelligence (i.e., the ability to process and use information) and the manifest variable IQ (i.e., the observed scores on intelligence tests which are subject to test motivation and error variance). However, it is important not to overstate our conclusions. For all measured outcomes in Study 2, the predictive validity of IQ remained statistically significant when controlling for test motivation. Moreover, the predictive validity of measured IQ was stronger for both measured academic outcomes than was the predictive validity of test motivation. These findings suggest that intelligence is, as Boring intimated, largely (albeit not exclusively) what intelligence tests test, and that intelligence is indeed important to academic success. Moreover, test motivation is likely less variable when extrinsic motivation is high (e.g., when financial incentives are

offered), when intrinsic motivation is high (e.g., among participants high in openness to experience who find the test problems interesting and fun), or when participants are above-average in measured IQ (e.g., college undergraduates).

Our investigation suggests that test motivation is likely a more serious confound in samples that include participants who are below-average in IQ and who lack external incentives to perform at their maximal potential. One such study is the National Longitudinal Survey of Youth (NLSY), a nationally representative sample of over 12,000 individuals first surveyed in 1979. Participants in the NLSY were administered the ASVAB in 1980. The Armed Forces Qualifying Test (AFQT) is a subset of four tests from the ASVAB. As is typical in social science research, participants were not rewarded in any way for higher scores, and maximal motivation was assumed rather than tested directly. The possible confound of test motivation was ignored by Herrnstein and Murray (1994) in their analysis of the AFQT data from the NLSY, published in *The Bell Curve*. In subsequent analysis of the same data, Segal (2006) argued that performance on the coding speed test is a good proxy for test motivation. Indeed, common sense suggests individuals who try harder will perform better on this 7-minute timed test in which the objective is simply to match 4-digit numbers to different words using a key of 10 words and their matching numbers. Given its simplicity and time limitation, performance on the coding speed test would seem more susceptible to decrements in test motivation than, say, the untimed ASVAB subtests for arithmetic reasoning or word knowledge. Consistent with this view is the fact that the coding speed test is the ASVAB subtest that is *least* correlated with other subtests (Hernnstein and Murray, 1994).

Segal (2006) found that performance on the coding speed test performance was highly correlated with earnings 23 years later, even after controlling for AFQT scores. Specifically, a

standard deviation increase in coding speed test scores predicted a 10% increase in 2003 earnings of male workers. Consistent with our finding that test motivation is more heterogeneous and below maximum (and, therefore, a more serious confound) among individuals of lower measured IQ, Segal found that coding speed best predicted income among the least educated individuals in the NLSY sample. Thus, as the economists Bowles, Gintis, and Osborne (2001) have suggested, "Because some of the noncognitive determinants of cognitive test performance may also influence performance on the job and hence subsequent earnings, covariation of test scores and earnings may overstate the importance of cognitive abilities in the earnings generation process" (p. 1158-59).

Apart from limitations in the measurement of personality in Study 2, the current investigation faced several methodological limitations that should be addressed in future research. Like any meta-analysis, Study 1 included articles varying widely in the detail of the described procedures. All included studies employed random assignment designs and previously validated intelligence tests, but we cannot affirm that the studies were uniformly well-executed (e.g., that experimenters administering IQ tests were blind to condition). In Study 2, the Pittsburgh Youth Study sample was socioeconomically and ethnically diverse but included only boys. While 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 statistically the prediction that test motivation was a stronger confound of IQ-outcome relations among participants of below-average IQ. Further, no data were available in Study 2 beyond early adulthood. An untested but intriguing possibility is that the predictive validity of test motivation might be even stronger for non-academic outcomes measured later in life (e.g., job performance,

age of retirement, divorce). Finally, as this investigation demonstrates, measured variables are necessarily imperfect observations of latent constructs. Accordingly, it may be that more precise and accurate measures of test motivation and personality may have explained more variance in life outcomes than those used in Study 2.

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 (Boe, May, & Boruch, 2002). 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 twelfth 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 student background questionnaire, accounted for 53% of between-nation variability in math achievement, 22% of between-classroom variability within nations, and 7% of between-student variability within classrooms (Boe, May & Boruch, 2002). 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.

The importance of the current investigation is two-fold. First, these findings constitute an existence proof that the motivation to perform well under standard testing conditions cannot be assumed to be maximal across all individuals. Some individuals try harder than others, and heterogeneity in test motivation constitutes systematic variance rather than random error. To amend Boring's original quip, *IQ* is what intelligence tests measure, and *both* intelligence and

test motivation determine IQ. Second, test motivation on low-stakes intelligence tests can partially account for observed IQ-outcome associations, particularly for non-academic outcomes. In part, the reason for this confound is that more conscientious, agreeable, and intellectual individuals both try harder on intelligence tests and do better in life.

These conclusions may come as no surprise to contemporary clinicians who appreciate the importance of test motivation as a determinant of IQ scores (Haywood, 1992). Indeed, the most recent manuals for the WAIS and the WISC intelligence tests explicitly acknowledge that certain subtests are sensitive to individual differences in attention and concentration. If prediction is the sole concern, the "pollution" of test scores by test motivation may in fact be welcome, insofar as it provides incremental predictive validity over and beyond "pure" intelligence. Where the problem lies, in our view, is in the most common theoretical interpretation of observed IQ scores. A research subject with a low IQ score is typically assumed to be lacking in intelligence rather than in motivation to perform well. Our simple point is that this is not necessarily true.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Achenbach, T. M., & Edelbrock, C. R. (1983). *Manual for the Child Behavior Profile and Child Behavior Checklist*. Burlington, VT: Author.

Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology, 21*, 726-735.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

* Benton, A. L. (1936). Influence of incentives upon intelligence test scores of school children. *Journal of Genetic Psychology, 49*, 494-497.

* Bergan, A., McManis, D. L., & Melchert, P. A. (1971). Effects of social and token performance on WISC Block Design performance. *Perceptual and Motor Skills, 32*, 871-880.

* Blanding, K. M., Richards, J., Bradley-Johnson, S., & Johnson, C. M. (1994). The effect of token reinforcement on McCarthy Scale performance for White preschoolers of low and high social position. *Journal of Behavioral Education, 4*, 33-39.

Boe, E. E., May, H., & Boruch, R. F. (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. *Research Report 2002-TIMSS1*.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive Meta-analysis (Version 2). Englewood, NJ: Biostat.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2008). Introduction to Meta-Analysis. Unpublished manuscript in preparation.

Borghans, L., Duckworth, A. L., Heckman, J. J., & Ter Weel, B. (in press). The economics and psychology of personality traits. *Journal of Human Resources*.

Borghans, L., Meijers, H., & Ter Weel, B. (2008). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry, 46*, 2-12.

Boring, E. G. (1923). Intelligence as the tests test it. *New Republic, 35*, 35-37.

Bowles, S., Gintis, H., & Osborne, M. (2001). The determinants of earnings: A behavioral approach. *Journal of Economic Literature, 39*, 1137-1176.

* Bradley-Johnson, S., Graham, D. P., & Johnson, C. M. (1986). Token reinforcement on WISC--R performance for White, low-socioeconomic, upper and lower elementary-school-age students. *Journal of School Psychology, 24*, 73-79.

* Bradley-Johnson, S., Johnson, C. M., Shanahan, R. H., Rickert, V. I., & Tardona, D. R. (1984). Effects of token reinforcement on WISC-R performance of Black and White, low socioeconomic second graders. *Behavioral Assessment, 6*, 365-373.

* Breuning, S. E., & Zella, W. F. (1978). Effects of individualized incentives on norm-referenced IQ test performance of high school students in special education classes. *Journal of School Psychology, 16*, 220-226.

Caspi, A., Block, J., Block, J. H., Klopp, B., Lynam, D., Moffit, T. E., et al. (1992). A "common-language" version of the California Child Q-Set for personality assessment. *Psychological Assessment, 4*, 512-523.

Caspi, A., Moffitt, T. E., Thornton, A., & Freedman, D. (1996). The life history calendar: A research and clinical assessment method for collecting retrospective event-history data. *International Journal of Methods in Psychiatric Research, 6*, 101-114.

* Clingman, J., & Fowler, R. L. (1976). The effects of primary reward on the I.Q. performance of grade-school children as a function of initial I.Q. level. *Journal of Applied Behavior Analysis, 9*, 19-23.

Conard, M. A. (2005). Aptitude is not enough: how personality and behavior predict academic performance. *Journal of Research in Personality, 40*, 339-346.

Cronbach, L. J. (1960). Is there rest for the test weary? *American Psychologist, 15*, 665-666.

* Devers, R., & Bradley-Johnson, S. (1994). The effect of token reinforcement on WISC-R performance for fifth- through ninth-grade American Indians. *Psychological Record, 44*, 441-449.

* Dickstein, L. S., & Ayers, J. (1973). Effect of an incentive upon intelligence test performance. *Psychological Reports, 33*, 127-130.

Duckworth, A. L. (in press). (Over and Beyond) High Stakes Testing. *American Psychologist*.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455-463.

* Edlund, C. V. (1972). The effect on the behavior of children, as reflected in the IQ scores, when reinforced after each correct response. *Journal of Applied Behavior Analysis, 5*, 317-319.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634.

* Ferguson, H. H. (1937). Incentives and an intelligence test. *Australasian Journal of Psychology & Philosophy, 15*, 39-53.

* Galbraith, G., Ott, J., & Johnson, C. M. (1986). The effects of token reinforcement on WISC-R performance of low-socioeconomic Hispanic second-graders. *Behavioral Assessment, 8*, 191-194.

* Gerwell, E. L. (1981). Tangible and verbal reinforcement effects on fluid and crystallized intelligence in the aged. Unpublished dissertation. Hofstra University.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24*, 79-132.

* Graham, G. A. (1971). The effects of material and social incentives on the performance on intelligence test tasks by lower class and middle class Negro preschool children. Unpublished dissertation. George Washington University.

Haywood, H. C. (1992). The strange and wonderful symbiosis of motivation and cognition. *International Journal of Cognitive Education and Mediated Learning, 2*, 186-197.

Herrnstein, R. J., & Murray, C. A. (1994). *The bell curve: Intelligence and class struggle in american life*. New York: Free Press.

Hobby, K. L. (1980). *WISC-R split-half short form manual*. Los Angeles: Western Psychological Services.

Hogan, R. (2005). In defense of personality measurement: New wine for old whiners. *Human Performance, 18*, 331-341.

Hollingshead, A. B. (1975). Four factor index of social status. Unpublished manuscript. Yale University Department of Sociology.

* Holt, M. M., & Hobbs, T. R. (1979). The effects of token reinforcement, feedback and response cost on standardized test performance. *Behaviour Research and Therapy, 17*, 81-83.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger Publishers/Greenwood Publishing Group.

John, O. P., Caspi, A., Robins, R. W., Moffitt, T. E., & Stouthamer-Loeber, M. (1994). The "little five": Exploring the nomological network of the five-factor model of personality in adolescent boys. *Child Development, 65*, 160-178.

Judge, T. A., Colbert, A. E., & Ilies, R. (2004). Intelligence and leadership: A quantitative review and test of theoretical propositions. *Journal of Applied Psychology, 89*, 542-552.

* Kapenis, J. T. (1979). The differential effects of various reinforcements and socioeconomic status upon Peabody Picture Vocabulary Test performance. Unpublished dissertation. University of South Dakota.

* Kieffer, D. A., & Goh, D. S. (1981). The effect of individually contracted incentives on intelligence test performance of middle- and low-SES children. *Journal of Clinical Psychology, 37*, 175-179.

* Lloyd, M. E., & Zylla, T. M. (1988). Effect of incentives delivered for correctly answered items on the measured IQs of children of low and high IQ. *Psychological Reports, 63*, 555-561.

Loeber, R., Farrington, D. P., Stouthamer-Loeber, M., & Van Kammen, W. B. (1998). *Antisocial behavior and mental health problems: Explanatory factors in childhood and adolescence*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lynam, D. R., Caspi, A., Moffit, T. E., Wikstrom, P.-O., Loeber, R., & Novak, S. (2000). The interaction between impulsivity and neighborhood context on offending: The effects of impulsivity are stronger in poorer neighborhoods. *Journal of Abnormal Psychology, 109*, 563-574.

Lynam, D., Moffitt, T. E., & Stouthamer-Loeber, M. (1993). Explaining the relation between IQ and delinquency: Class, race, test motivation, school failure, or self-control? *Journal of Abnormal Psychology, 102*, 187-196.

MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science, 1*, 173-181.

MacNicol, K. (1960). Effects of varying order of item difficulty in an unspeeded verbal test. Unpublished manuscript. Educational Testing Services.

Mount, M. K., Barrick, M. R., & Strauss, J. P. (1999). The joint relationship of conscientiousness and ability with performance: Test of the interaction hypothesis. *Journal of Management, 25*, 707-721.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., et al. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101.

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology, 93*, 116-130.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*, 157-159.

Paunonen, S. V., & Ashton, M. C. (2001a). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81*, 524-539.

Paunonen, S. V., & Ashton, M. C. (2001b). Big Five predictors of academic achievement. *Journal of Research in Personality, 35*, 78-90.

Pederson, S. L., & McCarthy, D. (2008). Person–Environment transactions in youth drinking and driving. *Psychology of Addictive Behaviors, 22*, 340-348.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42*, 185-227.

Revelle, W. (1993). Individual differences in personality and motivation: 'Non-cognitive' determinants of cognitive performance. In A. D. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness, and control* (pp. 346-373). New York: Oxford University Press.

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2*, 313-345.

Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Sackett, P. R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. *Journal of Applied Psychology, 73*, 482-486.

* Saigh, P. A., & Antoun, F. T. (1983). WISC-R incentives and the academic achievement of conduct disordered adolescent females: A validity study. *Journal of Clinical Psychology, 39*, 771-773.

Segal, C. (2006). Motivation, test scores, and economic success. Unpublished manuscript. Harvard Business School.

Snyderman, M., & Rothman, S. (1987). Survey of expert opinion on intelligence and aptitude testing. *American Psychologist, 42*, 137-144.

Sobel, M. E. (1982). Asymptotic intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology 1982* (pp. 290-312). San Francisco: Jossey-Bass.

* Steinweg, S. B. (1979). A comparison of the effects of reinforcement on intelligence test performance of normal and retarded children. Unpublished dissertation.

* Sweet, R. C., & Ringness, T. A. (1971). Variations in the intelligence test performance of referred boys of differing racial and socioeconomic backgrounds as a function of feedback or monetary reinforcement. *Journal of School Psychology, 9*, 399-409.

* Terrell, F., Terrell, S. L., & Taylor, J. (1980). Effects of race of examiner and type of reinforcement on the intelligence test performance of lower-class Black children. *Psychology in the Schools, 17*, 270-272.

Thorndike, E. L. (1904). *An Introduction to the theory of mental and social measurements*. Oxford: Science Press.

* Tiber, N. (1963). The effects of incentives on intelligence test performance. Unpublished dissertation. Florida State University.

Wechsler, D. (1940). Nonintellective factors in general intelligence. *Psychological Bulletin, 37*, 444-445.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children* (Third ed.). San Antonio, TX: The Psychological Corporation.

* Weiss, R. H. (1981). Effects of reinforcement on the IQ scores of preschool children as a function of initial IQ. Unpublished dissertation. Utah State University.

* Willis, J., & Shibata, B. (1978). A comparison of tangible reinforcement and feedback effects on the WPPSI I. Q. scores of nursery school children. *Education & Treatment of Children, 1*, 31-45.

Yudin, L. W. (1966). An abbreviated form of the WISC for use with emotionally disturbed children. *Journal of Consulting Psychology, 30*, 272-275.

Zigler, E., & Butterfield, E. C. (1968). Motivational aspects of changes in IQ test performance of culturally deprived nursery school children. *Child Development, 39*, 1-14.

Author Note

This research was supported by grant R01 MH45070 from the National Institute on Mental Health, grant R01 AG032282 from the National Institute on Aging, and the John Templeton Foundation.

Notes

[1] In the wake of Herrnstein and Murray's (1994) controversial treatise *The Bell Curve*, an APA taskforce offered the following consensual definition of intelligence: "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" (Neisser et al., 1996, p. 77). This definition emphasizes the ability to solve novel problems (i.e., fluid intelligence), but because it is not possible to create a content-free test, to varying degrees all intelligence tests also assess existing knowledge (i.e., crystallized intelligence).

[2] We did not find any within-subject studies that counterbalanced incentive and control conditions and also met the other inclusion criteria. Therefore, studies that lacked a between-subjects condition were excluded because the effect of incentives was not separable from the effect of practice.

[3] Because both test motivation and intelligence jointly determine measured IQ, we reasoned that estimates of the predictive validity of test motivation are likely underestimated when measured IQ is included as a covariate. Such estimates arguably constitute a lower bound on the effect of test motivation on outcomes, whereas upper bound estimates do not include measured IQ as a covariate.

Figure 1

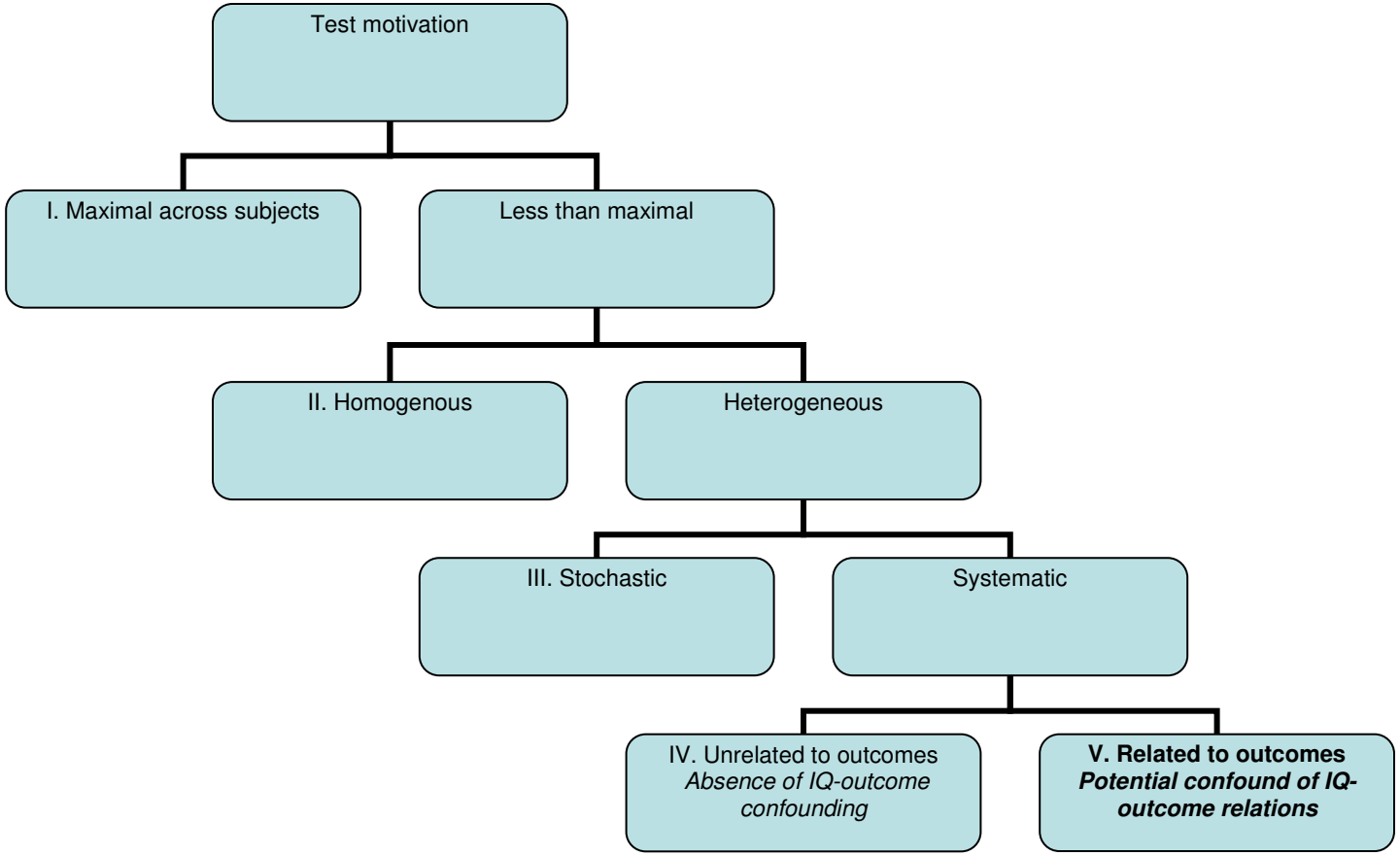
Test motivation
I. Maximal across subjects
Less than maximal
II. Homogenous
Heterogeneous
III. Stochastic
Systematic
IV. Unrelated to outcomes
Absence of IQ-outcome confounding
V. Related to outcomes
Potential confound of IQ-outcome relations

Table 1

Between-Subjects Studies of the Effect of Incentives on IQ Scores in Study 1

Study	Study Type	Sample	*N*	Intelligence Test	Raw Effect Size, *g* (*SE*)	Baseline IQ	Age in years	Incentive Value	Study Design
Benton (1936)	Article	-	50	OSAT	-0.09 (0.28)	Low	12-18	-[a]	Baseline and post-test
Bergan et al. (1971)	Article	Male	16	WISC	0.19 (0.47)	High	6-11	<$1	Baseline and post-test
		Female	16	WISC	-0.06 (0.48)	High	6-11	<$1	Baseline and post-test
Blanding et al. (1994)	Article	Study 1	29	MSCA	0.82 (0.38)	High	0-5	$1-$9	Baseline and post-test
		Study 2 low SES	23	MSCA	1.33 (0.45)	Low	0-5	$1-$9	Post-test only
		Study 2 high SES	20	MSCA	0.44 (0.43)	High	0-5	$1-$9	Post-test only
Bradley-Johnson et al. (1984)	Article	Study 1	22	WISC-R	0.17 (0.41)	Low	6-11	$1-$9	Baseline and post-test
		Study 2	22	WISC-R	0.67 (0.42)	High	6-11	$1-$9	Baseline and post-test
Bradley-Johnson et al. (1986)	Article	Early elementary	20	WISC-R	0.85 (0.45)	High	6-11	$10+	Baseline and post-test
		Late elementary	20	WISC-R	0.80 (0.45)	High	6-11	$10+	Baseline and post-test
Breuning & Zella (1978)	Article	Group 1	147	LTIT	2.12 (0.21)	Low	12-18	$10+	Baseline and post-test
		Group 2	129	OLMAT	2.38 (0.23)	Low	12-18	$10+	Baseline and post-test
		Group 3	209	WISC-R	10.94 (0.17)	Low	12-18	$10+	Baseline and post-test
Clingman & Fowler (1976)	Article	High IQ	16	PPVT	-0.06 (0.47)	High	6-11	$1-$9	Baseline and post-test
		Medium IQ	16	PPVT	-0.40 (0.48)	High	6-11	$1-$9	Baseline and post-test
		Low IQ	16	PPVT	1.42 (0.54)	Low	6-11	$1-$9	Baseline and post-test
Devers & Bradley-Johnson (1994)	Article	-	25	WISC-R	0.91 (0.41)	Low	12-18	$10+	Baseline and post-test
Dickstein & Ayers (1973)	Article	-	32	RPM, WAIS	0.49 (0.35)	High	19+	$1-$9	Post-test only
Edlund (1972)	Article	-	22	SBIS-III	0.89 (0.43)	Low	6-11	$1-$9	Baseline and post-test
Ferguson (1937)	Article	-	156	OSAT	0.03 (0.16)	High	12-18	<$1	Baseline and post-test
Galbraith et al. (1986)	Article	-	30	WISC-R	0.73 (0.37)	Low	6-11	$1-$9	Baseline and post-test
Gerwell (1981)	Dissertation	-	64	WAIS	0.58 (0.25)	High	19+	$1-$9	Post-test only
Graham (1971)	Dissertation	-	128	WPPSI	-0.14 (0.18)	Low	0-5	$1-$9	Baseline and post-test
Holt & Hobbs	Article	-	40	WISC	1.03 (0.33)	Low	12-18	-[a]	Post-test only

(1979)									
Kapenis (1979)	Dissertation	Low SES	28	PPVT	0.51 (0.37)	High	6-11	<$1	Post-test only
		Middle SES	28	PPVT	-0.17 (0.37)	High	6-11	<$1	Post-test only
		High SES	28	PPVT	0.09 (0.37)	High	6-11	<$1	Post-test only
Kieffer & Goh (1981)	Article	Low SES	32	WISC-R	0.66 (0.35)	Low	6-11	$1-$9	Post-test only
		Middle SES	32	WISC-R	-00.26 (0.35)	High	6-11	$1-$9	Post-test only
Lloyd & Zylla (1988)	Article	High IQ	16	WPPSI	1.04 (0.51)	High	0-5	$1-$9	Baseline and post-test
		Low IQ	16	WPPSI	0.62 (0.49)	Low	0-5	$1-$9	Baseline and post-test
Saigh & Antoun (1983)	Article	-	34	WISC-R	0.93 (0.36)	Low	12-18	-[a]	Post-test only
Steinweg (1979)	Dissertation	Group 1	10	SBIS-III	0.17 (0.57)	High	6-11	$1-$9	Post-test only
		Group 2	10	WISC-R	0.20 (0.57)	High	6-11	$1-$9	Post-test only
Sweet & Ringness (1971)	Article	Low SES black	36	WISC	0.39 (0.33)	Low	6-11	$1-$9	Post-test only
		Low SES white	48	WISC	1.17 (0.31)	Low	6-11	$1-$9	Post-test only
		Middle SES white	72	WISC	0.13 (0.23)	High	6-11	$1-$9	Post-test only
Terrell et al. (1980)	Article	Black examiner	30	WISC-R	1.18 (0.39)	Low	6-11	<$1	Post-test only
		White examiner	30	WISC-R	1.40 (0.40)	Low	6-11	<$1	Post-test only
Tiber (1963)	Dissertation	Low SES black	80	SBIS-III	0.11 (0.22)	Low	6-11	<$1	Post-test only
		Low SES white	80	SBIS-III	-0.34 (0.22)	Low	6-11	<$1	Post-test only
		Middle SES white	80	SBIS-III	0.00 (0.22)	High	6-11	<$1	Post-test only
Weiss (1981)	Dissertation	High IQ	10	PPVT	0.57 (0.59)	High	0-5	$1-$9	Baseline and post-test
		Medium IQ	10	PPVT	1.25 (0.64)	High	0-5	$1-$9	Baseline and post-test
		Low IQ	10	PPVT	3.64 (1.00)	Low	0-5	$1-$9	Baseline and post-test
Willis & Shibata (1978)	Article	-	20	WPPSI	0.59 (0.44)	Low	0-5	$1-$9	Baseline and post-test

[a] There was insufficient information presented to determine the value of the incentive for these samples.

Table 2

Moderation of Effect of Incentives by Intelligence Level in Study 1

Intelligence Level	*g*	95% CI of *g*	*k*	*n*	I^2	$Q_{between}$ *(df)*
Low	0.94**	0.54, 1.35	23	1257	90.28**	
High	0.26**	0.10, 0.41	23	751	9.71	
Between groups comparison						9.76* (1)

* $p < .01$. ** $p < .001$.

Table 3

Summary Statistics and Zero-Order Correlations in Study 2

Variable	*M*	*SD*	2	3	4	5	6	7	8	9	10[c]	11[d]	12[d]
1. Test motivation[a]	0.09	0.84	.28***	-.01	-.04	-.08	.15*	.31***	.24***	.21**	-.16*	-.01	-.11***
2. IQ	101.80	15.77	-	-.47***	-.34***	-.21**	.37***	.70***	.47***	.21***	-.21**	-.13	-.18***
3. Age at follow-up interview	24.02	0.91		-	.19**	.17**	-.27***	-.29***	-.24***	-.10	.13*	.04	.08*
4. Black	44%				-	.35***	-.17**	-.21***	-.13	-.14	.09	.06	.12**
5. Single parent home	56%					-	-.22***	-.30***	-.23***	-.14*	.10	.11	.17**
6. Family SES	38.13	11.67					-	.25***	.28***	.02	-.05	-.16*	-.05
7. Academic performance	2.76	0.81						-	.52***	.21***	-.19**	-.17*	-.12**
8. Years of education	12.36	2.02							-	.28***	-.34***	-.75***	-.06**
9. Employed at follow-up	72%									-	-.71***	-.08	-.19***
10. Unemployment	0.57	1.15									-	.27**	.04
12. Ever arrested	39%											-	-
13. Additional arrests[b]	2.10	2.65											-

Note. *ns* range from 223 to 251.
* $p < .05$. ** $p < .01$. *** $p < .001$.
[a] Correlations are with natural log transformed test motivation. [b] Number of arrests among participants with at least one arrest.
[c] Spearman's rho correlation coefficients. [d] Converted from odds ratios.

Table 4

Summary of Simultaneous Regression Models Predicting School Performance in Adolescence in Study 2

Variable	*1*	*2*	3
Race	0.10*	-0.10	0.09
Single-parent home	-0.20***	-0.22***	-0.20***
Family SES	-0.04	0.15*	-0.05
IQ	0.71***		0.68***
Test motivation		0.27***	0.12*
R^2	.53***	.21***	.54***

Note. Values are standardized regression coefficients. $n = 251$.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 5

Summary of Simultaneous Linear Regression Models Predicting Years of Education in Study 2

Variable	*1*	*2*	3
Race	0.09	-0.003	0.08
Single-parent home	-0.16*	-0.16*	-0.16**
Family SES	0.10	0.17**	0.09
Age at follow-up	-0.01	-0.17*	-0.03
IQ	0.43***		0.39***
Test motivation		0.22***	0.13*
R^2	.26***	.18***	.28***

Note. Values are standardized regression coefficients. $n = 223$.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 6

Summary of Simultaneous Binary Logistic Regression Models Predicting Current Employment in Study 2

Variable	*1*	*2*	3
Race	0.87	0.78	0.86
Single-parent home	0.79	0.80	0.80
Family SES	0.79	0.83	0.75
Age at follow-up	0.98	0.76	0.90
IQ	1.87**		1.63*
Test motivation		1.69***	1.54**
R^2	.12***	.13***	.16***

Note. Values are odds ratios. $n = 236$.
* $p < .05$. ** $p < .01$. ** $p < .001$.
[a] Nagelkerke R^2.

Table 7

Summary of Simultaneous Ordinal Logistic Regression Models Predicting Unemployment History in Study 2

Variable	*1*	*2*	3
Race	1.04	1.12	1.05
Single-parent home	1.14	1.12	1.13
Family SES	1.11	1.06	1.13
Age at follow-up	1.13	1.35	1.20
IQ	0.64*		0.69^
Test motivation		0.71*	0.77^
R^2	.06*	.06*	.07*

Note. Values are odds ratios. $n = 239$.
^ $p < .10$. * $p < .05$.
[a] Nagelkerke R^2.

Table 8

Summary of Zero-Inflated Poisson Regression Models Predicting Number of Arrests in Study 2

Variable	Count 1	Logistic 1	Count 2	Logistic 2	Count 3	Logistic 3
Race	1.15	0.86	1.28*	0.84	1.21	0.74
Single-parent home	1.34*	1.42	1.43**	1.25	1.45**	1.28
Family SES	0.99	0.70	0.99	0.65	1.03	0.68
Age at follow-up	0.97	0.89	1.13	0.97	1.03	0.82
IQ	0.65**	0.82			0.76*	0.66
Test motivation			0.68***	1.25	0.71***	1.46

Note. Values are odds ratios. Odds ratios in columns marked "logistic" reflect likelihood of having at least one arrest. Odds ratios in columns marked "count" reflect likelihood of having additional arrests given at least one arrest. $n = 239$.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Table 9

Percentage of Variance Explained in Life Outcomes in Study 2

Predictor	Covariates	Academic performance in adolescence	Total years of education	Employment in adulthood[a]	Unemployment History[a]	Additional arrests in adulthood[b]
IQ	Demographics	39.34***	12.95***	5.84**	2.86*	2.72**
IQ	Demographics, test motivation	33.50***	9.81***	3.09*	1.75	1.17*
IQ	Demographics, test motivation residuals[c]	35.93***	11.16***	3.90**	1.97*	1.52*
Test motivation	Demographics	7.10***	4.60***	6.87***	2.53*	2.34***
Test motivation	Demographics, IQ	1.26*	1.46*	4.00**	1.42	1.74***
Test motivation residuals[c]	Demographics, IQ	0.65	0.76	3.04*	1.58	1.42***

Note. * $p < .05$. ** $p < .01$. ** $p < .001$.
[a] Nagelkerke R^2.
[b] Converted from Odds Ratio.
[c] Remaining variance in test motivation after removing variance explained by Big Five Agreeableness, Conscientiousness, and Openness to Experience.

Table 10
Correlations among Test Motivation, IQ, and Parent-Rated Big Five Dimensions in Study 2

Variable	1	2	3	4	5	6	7
1. Test Motivation	-	.13*	.15*	.02	-.05	.15*	.28***
2. Agreeableness	.14*	-	.41***	-.04	-.12	.16**	.01
3. Conscientiousness	.10	.42***	-	.08	-.35***	.22***	.19**
4. Extraversion	.01	-.04	.08	-	-.37***	.05	.02
5. Neuroticism	-.01	-.12	-.34***	-.37***	-	-.24***	-.12
6. Openness to Experience	.05	.17**	.17**	.04	-.21***	-	.34***
7. IQ	-	-	-	-	-	-	-

Note. Zero-order correlation coefficients are above the diagonal (n = 249). Partial correlation coefficients controlling for measured IQ are below the diagonal.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Figure Captions

Figure 1. Potential scenarios and consequences of variation in test motivation.