# Web Appendix for “Identification Problems in Personality Psychology”


Lex Borghans
Maastricht University

Bart H. H. Golsteyn
Maastricht University
SOFI, Stockholm University

James Heckman
University of Chicago
University College Dublin
American Bar Foundation

John Eric Humphries
University of Chicago

# Contents

# 1. Incentives and Performance on Intelligence Tests

Table 1. Incentives and Performance on Intelligence Tests

| Study | Sample and Study Design | Experimental Group | Effect size of incentive (in standard deviations) | Summary |
|---|---|---|---|---|
| Edlund [1972] | Between subjects study. 11 matched pairs of low SES children; children were about one standard deviation below average in IQ at baseline | M&M candies given for each right answer | Experimental group scored 12 points higher than control group during a second testing on an alternative form of the Stanford Binet (about 0.8 standard deviations). | "…a carefully chosen consequence, candy, given contingent on each occurrence of correct responses to an IQ test, can result in a significantly higher IQ score."(p. 319) |
| Ayllon and Kelly [1972] Sample 1 | Within subjects study. 12 mentally retarded children (avg IQ 46.8) | Tokens given in experimental condition for right answers exchangeable for prizes | 6.25 points out of a possible 51 points on Metropolitan Readiness Test. $t$ = 4.03 | "…test scores often reflect poor academic skills, but they may also reflect lack of motivation to do well in the criterion test…These results, obtained from both a population typically limited in skills and ability as well as from a group of normal children (Experiment II), demonstrate that the use of reinforcement procedures applied to a behavior that is tacitly regarded as "at its peak" can significantly alter the level of performance of that behavior." (p. 483) |
| Ayllon and Kelly [1972] Sample 2 | Within subjects study 34 urban fourth graders (avg IQ = 92.8) | Tokens given in experimental condition for right answers exchangeable for prizes | $t$ = 5.9 | |
| Ayllon and Kelly [1972] Sample 3 | Within subjects study of 12 matched pairs of mentally retarded children | Six weeks of token reinforcement for good academic performance | Experimental group scored 3.67 points out of possible 51 points on a post-test given under standard conditions higher than at baseline; control group dropped 2.75 points. On a second post-test with incentives, exp and control groups increased 7.17 and 6.25 points, respectively. | |
| Clingman and Fowler [1976] | Within subjects study of 72 first- and second-graders assigned randomly to contingent reward, noncontingent reward, or no reward conditions. | M&Ms given for right answers in contingent cdtn; M&Ms given regardless of correctness in noncontingent condition | Only among low-IQ (<100) subjects was there an effect of the incentive. Contingent reward group scored about 0.33 standard deviations higher on the Peabody Picture Vocabulary test than did no reward group. | "…contingent candy increased the I.Q. scores of only the 'low I.Q.' children. This result suggests that the high and medium I.Q. groups were already functioning at a higher motivational level than children in the low I.Q. group." (p. 22) |

(Table 1. Incentives and Performance on Intelligence Tests Continued …)

| Study | Sample and Study Design | Experimental Group | Effect size of incentive (in standard deviations) | Summary |
|---|---|---|---|---|
| Zigler and Butterfield [1968] | Within and between subjects study of 52 low SES children who did or did not attend nursery school were tested at the beginning and end of the year on Stanford-Binet Intelligence Test under either optimized or standard conditions. | Motivation was optimized without giving test-relevant information. Gentle encouragement, easier items after items were missed, and so on. | At baseline (in the fall), there was a full standard deviation difference (10.6 points and SD was about 9.5 in this sample) between scores of children in the optimized vs standard conditions The nursery group improved their scores, but only in the standard condition. | "…performance on an intelligence test is best conceptualized as reflecting three distinct factors: (a) formal cognitive processes; (b) informational achievements which reflect the content rather than the formal properties of cognition, and (c) motivational factors which involve a wide range of personality variables. (p. 2)<br>"…the significant difference in improvement in standard IQ performance found between the nursery and non-nursery groups was attributable solely to motivational factors…" (p. 10) |
| Breuning and Zella [1978] | Within and between subjects study of 485 *special education* high school students. All took IQ tests, then were randomly assigned to control or incentive groups to retake tests. Subjects were below-average in IQ. | Incentives such as record albums, radios (<$25) given for improvement in test performance | Scores increased by about 17 points. Results were consistent across the Otis-Lennon, WISC-R, and Lorge-Thorndike tests. | "In summary, the promise of individualized incentives contingent on an increase in IQ test performance (as compared with pretest performance) resulted in an approximate 17-point increase in IQ test scores. These increases were equally spread across subtests… The incentive condition effects were much less pronounced for students having pretest IQs between 98 and 120 and did not occur for students having pretest IQs between 121 and 140." (p. 225) |

(Table 1. Incentives and Performance on Intelligence Tests Continued …)

| Study | Sample and Study Design | Experimental Group | Effect size of incentive (in standard deviations) | Summary |
|---|---|---|---|---|
| Holt and Hobbs [1979] | Between and within subjects study of 80 delinquent boys randomly assigned to three experimental groups and one control group. Each exp group received a standard and modified administration of the WISC-verbal section. | Exp 1-Token reinforcement for correct responses; Exp 2 – Tokens forfeited for incorrect responses (punishment), Exp 3-feedback on correct/incorrect responses | 1.06 standard deviation difference between the token reinforcement and control groups (inferred from $t$= 3.31 for 39 degrees of freedom) | "Knowledge of results does not appear to be a sufficient incentive to significantly improve test performance among below-average I.Q. subjects…Immediate rewards or response cost may be more effective with below-average I.Q. subjects while other conditions may be more effective with average or above-average subjects." (p. 83) |
| Larson, Saccuzzo and Brown [1994] | Between subjects study of 109 San Diego State University psychology students | Up to $20 for improvement over baseline performance on cognitive speed tests | "While both groups improved with practice, the incentive group improved slightly more." (p.34) $F(1,93) = 2.76,\ p < .05$ | 2 reasons why incentive did not produce dramatic increase: 1) few or no unmotivated subjects among college volunteers, 2) information processing tasks are too simple for 'trying harder' to matter |
| Duckworth [2007] | Within subjects study of 61 urban low-achieving high school students tested with a group-administered Otis-Lennon IQ test during their freshman year, then again 2 years later with a one-on-one (WASI) test | Standard directions for encouraging effort were followed for the WASI brief test. Performance was expected to be higher because of the one-on-one environment. | Performance on the WASI as juniors was about 16 points higher than on the group-administered test as freshmen. Notably, on the WASI, this population looks almost "average" in IQ, whereas by Otis-Lennon standards they are low IQ. $t(60) = 10.67$, $p < 0.001$ | The increase in IQ scores could be attributed to any combination of the following 1) an increase in "*g*" due to schooling at an intensive charter school, 2) an increase in knowledge or crystallized intelligence, 3) an increase in motivation due to the change in IQ test format, and/or 4) an increase in motivation due to experience at high performing school |

Source: Almlund, Duckworth, Heckman et al. [2011]

## 2. Data

We use two complementary data sets. The Stella Maris high school data contains an achievement test (the Differential Aptitude Test), a measure of IQ which is often considered to have the highest load on *g* (the Raven test), various measures of personality traits (Big 5, Grit) and measures of performance (grades). We complement our analyses using the NLSY79 data set. This includes the AFQT and the DAT, many IQ tests, two measures of personality traits (self-esteem and self-efficacy) and a large set of outcomes later in life.

### 2.A. *Stella Maris data set*

We combine baseline data from an experiment conducted at Stella Maris high school near Maastricht in the Netherlands in June 2008 (Borghans, Golsteyn, Heckman et al. [2009]) with administrative sources. The data are for diverse students who attain different levels of education. There are three academic tracks at Dutch high school.[1] Our data include students from the middle track (HAVO) which prepares students for professional colleges schools and the upper track (VWO) which prepares students for university.[2]

The students in our sample are 15 and 16 years of age at the time of the experiment in 2008. Participation in the experiment was compulsory. Some of the students had valid reasons not to participate. Of an initial sample of 374 students, 347 students (93.1%) actually participated.

[1] Dutch primary schools educate all students at the same level. In the last year in primary school - the students are around 12 years of age then - a decision is made whether the student should continue education in the lower high school track or a higher (i.e. middle or upper) track. This decision is based on how the student performed in primary school and on a score on the CITO achievement test, taken in the last year in primary school. The separation between middle and upper level is made in the first year in high school, and is based on first year high school grades.

[2] We do not have data on students who attend the lowest track (VMBO) that prepares students for trade schools. Approximately 50% of all Dutch high school students attend this lower track. The NLSY data contains students over the full education spectrum.

In the baseline for the experiment, we collect several measures of IQ, personality and outcomes. We use the following measures of personality: 50 items to measure the Big 5 (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) from Goldberg [1992] and 17 questions to measure Grit, a measure of perseverance and passion for long term goals, from Duckworth, Peterson, Matthews et al. [2007]. All measured traits have high Cronbach's Alphas, a measure of inter-correlation among scores.

We use the principal component of 8 Raven Progressive Matrices as a measure of IQ (alpha = 0.62). The Raven matrices are often considered to have the highest loading on *g*. The test is a measure of fluid intelligence, i.e. the ability to solve novel problems in which advanced elements of the collective intelligence of the culture are not required for solution.[3]

It is conceivable that even pure tests of cognition are to some extent related to personality skills. Our data show that the Raven test correlates with personality traits but the *R*-squared is very low (0.02).

From administrative records, we obtain scores on the Dutch Differential Aptitude Test (DAT) comparable to the American DAT, an achievement test taken at age 15. In the third year of high school (age 15), students have to choose in which of four fields[4] they will specialize during the last three high school years. The fields are prerequisites for entry into specific majors in college. Choosing the right field is important since it implies excluding certain college options. Students at this node of the decision tree have to formulate provisional plans about the major they want to pursue in college.[5] At this stage, students take a test to receive additional information about which major fits their abilities and interest. The DAT is taken as part of this

[3] Crystallized intelligence, in contrast, measures the available knowledge a person possesses.
[4] These fields are: culture and society; economics and society; nature and health; nature and technology.
[5] College educations in the Netherlands are highly specialized. For example, when students enter university (age 18), they have to choose between studying economics or econometrics.

test. It measures abilities in nine subfields.[6] The outcomes on the test do not restrict the choices of students to pick a field.  Performing well on the DAT helps in choosing a subfield, but there are other ways to receive information on abilities. We use the principal component of the DAT scores (except "speed and accuracy" and "practical insight") in our analyses (Alpha = 0.68).

The DAT and the AFQT are similar in terms of components and – as we will show below – the DAT and AFQT correlate highly. Therefore, conclusions we draw based on the DAT will be instructive about the AFQT as well.

We also received data from the school's administrative records containing the students' grades in all years they attended the school. We use the grades in the first year the students attended high school because only in this year all students were taught at the same level and had the same combination of courses.[7]

### 2.B. ***NLSY data***

We use a sub-sample from the NLSY79 with valid IQ scores included in their high school transcripts to decompose AFQT scores into the portion explained through IQ, the portion explained through personality traits proxied by the Rosenberg Self Esteem test and Rotter Locus of Control, and background characteristics.

The personality traits in the NLSY include the 4 item Rotter Locus of Control (Alpha = 0.359) and the 10 item and the Rosenberg Self Esteem Scale (Alpha = 0.831). Both tests were administered in 1979.

---

[6] These subfields include Word list (synonyms), Word image (spelling), Language use (sentences), Thinking with words (analogies), Thinking with numbers (sequences of numbers), Speed and accuracy (choosing between combinations), Thinking with figures (sequences of figures), Three-dimensional (folded patterns), Practical insight (how can a practical problem be solved).

[7] These courses included Dutch, English, French, Math, Biology, Technology, Computer science, Geography, and History.

The NLSY includes many IQ tests collected from school transcript data for subgroups (the number of respondents is reported in parentheses): CTMM – California Test of Mental Maturity (599), OLMAT – Otis Lenon Mental Ability Test (1191), LTIT – Lorge-Thorndike Intelligence Test (691), HNTMM – Henmon-Nelson Test of Mental Maturity (201), KAIT – Kuhlmann-Anderson Intelligence Test (176), SBIS – Stanford-Binet Intelligence Scale (101), and WISC – Wechsler Intelligence Scale for Children (120). The date at which these tests are administered ranges from early childhood to the 12th grade. We use IQ percentiles as IQ scores are highly non-linear. When using standard IQ scores, including IQ scores squared, cubed, and raised to the fourth power (while all significant), the model fit is still inferior to using IQ percentile scores. IQ percentiles are reported for fewer people, but by matching percentiles across those with the same standard score, most scores are recovered. This imputation filled in missing percentile scores with the mode percentile score of individuals who received the same score on the same IQ test. The advantage of percentile scores is that - in theory - they should be comparable across tests, allowing us to pool test scores from SBIS, WCIS, OLMAT, LTIT, HNTMM, KAIT, and CTMM IQ tests for a much larger sample of test takers.

The scores on the IQ tests are related to some extent to locus of control and self esteem. The R-squared of a regression of IQ on these personality traits is 0.07. This is much higher than the R-squared of the regression of Raven on Big five and Grit in the Stella Maris data (0.02), indicating that our measure for IQ in the NLSY loads higher on personality than our measure for IQ in the Stella Maris data set.

The achievement tests included in the NLSY are the AFQT and the DAT. The Armed Forces Qualifying Test (AFQT) was administered in 1980 as part of the Armed Services Vocational Aptitude Battery (ASVAB). The AFQT is a combination of four subtests of the

Armed Services Aptitude Battery (ASVAB): arithmetic reasoning, numeric operations, word knowledge, and paragraph comprehension. Coding speed, numeric operations are additional ASVAB subtests which are not included in the AFQT. Our data contain 11,878 valid AFQT scores for the entire NLSY79 once we restrict to the non-military sample. The Differential Aptitude Test (DAT) was administered between 7th and 12th grade. It was collected from high school transcripts. We obtain 569 valid DAT scores for the nonmilitary NLSY79 sample. A subsample of the NLSY did both the DAT and AFQT (289 respondents). The correlation of percentile scores is very high, ranging from 0.76 to .80.

## References

Almlund, Mathilde, Duckworth, Angela L., Heckman, James J. and Kautz, Tim (2011). "Personality Psychology and Economics." in *Handbook of the Economics of Education*. E. A. Hanushek, S. Machin and L. Wößmann, eds. Amsterdam, Elsevier**:** Forthcoming.

Ayllon, Teodoro and Kelly, Kathy (1972). "Effects of Reinforcement on Standardized Test Performance." *Journal of applied behavior analysis* **5**(4): 477.

Borghans, Lex , Golsteyn, Bart H. H., Heckman, James J. and Meijers, Huub (2009). "Gender Differences in Risk Aversion and Ambiguity Aversion." *Journal of the European Economic Association* **7**(2-3): 649-658.

Breuning, Stephen E. and Zella, William 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**(3): 220.

Clingman, Joy and Fowler, Robert 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**(1): 19-23.

Duckworth, Angela L. (2007). "Unpublished Dataset." University of Pennsylvania, Department of Psychology.

Duckworth, Angela L., Peterson, Christopher, Matthews, Michael D. and Kelly, Dennis R. (2007). "Grit: Perseverance and Passion for Long-Term Goals." *Journal of Personality and Social Psychology* **92**(6): 1087-1101.

Edlund, Calvin 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**(3): 317.

Goldberg, Lewis R. (1992). "The Development of Markers for the Big-Five Factor Structure." *Psychological Assessment* **4**(1): 26-42.

Holt, Michael M. and Hobbs, Tom R. (1979). "The Effects of Token Reinforcement, Feedback and Response Cost on Standardized Test Performance." *Behaviour Research and Therapy* **17**(1): 81-83.

Larson, Gerald E., Saccuzzo, Dennis P. and Brown, James (1994). "Motivation: Cause or Confound in Information Processing/Intelligence Correlations?" *Acta Psychologica* **85**(1): 25-37.

Zigler, Edward F. and Butterfield, Earl C. (1968). "Motivational Aspects of Changes in IQ Test Performance of Culturally Deprived Nursery School Children." *Child Development* **39**(1): 1-14.