Supporting online material for

Secular Rise in Economically Valuable Personality Traits

Markus Jokela, Tuomas Pekkarinen, Matti Sarvimäki*, Marko Terviö, Roope Uusitalo

*Corresponding author. E-mail: matti.sarvimaki@aalto.fi

Estimated Flynn effect

Table S1 reports the means and SDs of raw test scores by birth year. Our estimated “Flynn effect” is based on these numbers. Regressing the sum of cognitive test scores of the 1962-1976 cohorts on birth year and a constant yields a trend coefficient of 0.018, with standard error 0.005 (both measured in 1962 SDs per year). With all cohorts included the estimate is 0.012 (0.004).

Test score distributions

In Fig 1, we reported changes in average test scores across birth years. A limitation of these measures is that test scores are rank ordered statistics that have no natural scale. Thus any monotonic transformation of the test score is, in principle, an equally valid measure of performance and these arbitrary scaling decisions may affect conclusions about cohort differences (1,2,3).

We first examine the sensitivity of our conclusions by studying how the distributions of test scores evolve across cohorts. Figs S2 and S3 plot the cumulative distribution functions of the raw scores for three-year birth cohorts. They reveal that the shift in the test scores takes place over the entire distribution for all traits except *masculinity* (for which we do not find trends in averages, either). Figs S4 and S5 show the same distributional shifts using histograms instead of CDFs.

As is clear from these figures, with the exception of *masculinity*, the test score distributions of later years stochastically dominate the test score distributions of the earlier years. In particular, comparing the 1962 and 1976 birth cohorts reveals that the distributions of scores for the latter cohort dominate the earlier distributions, in the sense of first order stochastic dominance (FOSD). A formal test, proposed by (4), fails to reject the null hypothesis of FOSD at any conventional significance level for all subscores except masculinity, while yielding p-values below 10^{-7} for all subscores for the converse hypotheses.

Anchored test scores

In Figs 3 and 4 of the main paper, we reported results for anchored test scores, where we had scaled the raw test scores by average earnings associated with each combination of test score. We construct these measures using a similar approach as (5). We first estimate regressions:

$$y_i = \alpha + \sum_{s=1}^{N_s} P_{is}\beta_s + \sum_{c=1963}^{1976} C_{ic}\theta_c + \varepsilon_i \tag{1}$$

where y_i is the average annual earnings at ages 30-34 of individual i, P_{is} is his subscore s, C_{ic} is an indicator variable for being born in year c (using year 1962 as omitted category), and ε_i is an error term. In our baseline analysis, we estimate equation (1) using data for the 1962–1976 birth cohorts. The resulting estimates for personality test scores are reported in the first column of Table S2. We use these estimates to construct predicted earnings for each individual based on their personality test scores while holding the year of birth fixed-effects fixed at 1962 level. The second column of Table S2 reports corresponding estimates for cognitive ability test scores, which we use to construct anchored scores for cognitive ability in the same way.

Table S2 shows that most personality test scores predict higher later-life earnings also when we condition on other personality tests scores. The exceptions are *sociability* and *activity-energy*, which predict lower income conditional on other personality test scores. Furthermore, *masculinity* does not have statistically significant predicting power once we condition on other personality test scores.

The third column 3 of Table S2 reports results from a specification where we include both personality and cognitive ability tests scores in the anchoring regression. The estimates for personality test scores are quite robust to conditioning for cognitive ability test scores. The exceptions are that *sociability* and *dutifulness* now predict higher income, while *deliberation* predicts lower income.

Fig S6B reports trends corresponding to Fig 3 of the main paper, but now based on anchoring regressions that include simultaneously both personality and cognitive ability test scores (Fig S6A reproduces Fig 3 for reference). More precisely, the average anchored personality test scores are constructed using the estimates reported in Table S2, column 3, and holding cognitive ability test scores constant at 1962 birth cohort average and using the observed personality test scores to predict earnings. Similarly, we present average anchored test scores holding personality test scores at the 1962 average level while allowing cognitive ability test scores to evolve as they did. This approach yields an increase of 7.4% for anchored personality test scores and of 8.4% for anchored cognitive ability test scores.

For simplicity and comparability with studies that only have either cognitive or personality measures available we report in Figure 3 of the main paper trends based on regressions where cognitive skills and personality are anchored separately. However, the conclusions are similar when anchoring is done using both tests at the same time. The growth rate of both cognitive skills and personality test scores has been roughly equal over cohorts that we observe in data.

Alternative anchoring variables

Fig S7 shows that our results are not driven by the choice of model specification or the way we construct our earnings measure. We start by presenting a nonparametric version of Fig 3 in Fig S7A. We have constructed it by first regressing our main anchoring variable (average earnings at age 30-34) on a full set of indicator variables for each possible personality test score and then predicting income for all birth cohorts using the resulting coefficients. The results are closely correlated with those from a linear specification (correlation coefficient 0.92). For cognitive ability, the linear and the nonparametric anchored test scores are even more similar (correlation coefficient 0.99). Furthermore, the trends shown in Fig S7A are very similar to those in Fig 3. Anchored personality scores increased by 11.6% between the 1962 and 1976 birth cohorts according to the linear specification and by 12.1% according to the nonparametric specification. Thus we conclude that the linear specification is sufficient for anchoring the test scores in our context.

Fig S7B presents a version of Fig 3 where we anchor test scores to a broader income measure at age 30–34, which now also includes capital income and most government transfers. The resulting increase in anchored personality test scores between the 1962 and 1976 birth cohorts is 11.1%. The next three panels show similar patters when anchoring test scores to earnings at age 30 (Fig S7C), earnings percentile rank at age 30 (Fig S7D) and a logarithm of earnings at age 30–34 (Fig S7E).

Fig S7F reports results for the annualized discounted earnings from ages 28-48. As above, we do not drop zeros. We use discount rate 3%, and deflate all values to 2010 Euros using the Statistics Finland CPI. The advantage of this earnings measure is that it is a better proxy for lifetime income than our baseline measure of average earnings at age 30–34, but has the drawback of being available only for the 1962 birth cohort. This approach leads to a 9.5% increase in anchored personality test scores. Again, the trend in anchored cognitive ability tests scores is very similar.

Figs S7G and S7H report changes in test scores anchored to educational attainment. They show that changes in personality test scores predict 8.7 percentage points increase in the likelihood of obtaining lower tertiary degree or more (from a baseline of 30.1%). The predicted likelihood of completing an advanced degree increases by 4.8 percentage points (from a baseline of 8.3%). The trends in personality test scores and cognitive ability test scores are very similar to each other, as was the case with income-related anchoring variables.

Coefficients for alternative outcomes

The middle panel of Table S2 reports regression coefficients when using earnings percentile rank at age 30 as the outcome variable. The results are very similar to those above in the baseline specification, with the exception that *sociability* and *deliberation* do not have predictive power, while *masculinity* is now statistically significantly associated with higher income rank. The differences in specifications controlling for cognitive ability test scores are that *self-confidence* is not statistical significant, while *masculinity* is.

The last panel of Table S2 reports results when using an indicator for holding a lower tertiary degree or more. Again, the results are broadly similar to those for earnings. The main difference is that *deliberation* and *masculinity* now predict lower educational attainment in both specifications.

Changes over the distribution of anchored test scores

In addition to documenting changes in average anchored test scores, it is informative to examine whether some parts of the test score distribution change differently than others. Fig S8 plots the CDFs our baseline anchored scores and Table S3 corresponding estimates from quantile regressions. The results show that, while anchored personality test scores increased throughout the distribution, the changes are larger at the bottom of the distribution. For example, between the 1962 and 1976 birth cohorts, the 10th percentile of the test score distribution increased by €3,200, while the 90th percentile increased by €1,340. Similar pattern, though less pronounced, is also present for the cognitive ability test scores, where the corresponding estimates are €2,600 and €1,670 in the 10th and 90th percentiles, respectively.

Exploratory factor analysis

The FDF test score data contain eight personality trait scores and three cognitive skill scores. Table S4 shows that both the cognitive scores and the personality trait scores are strongly correlated within their domains, but the correlations across cognitive and personality domains are only modest.

We performed a simple explorative factor analysis to determine an appropriate way to reduce dimensionality of the test score data. In Fig S9, we plot the eigenvalues of the test score data. Only two first eigenvalues exceed one, suggesting that a two-factor model is a sufficient description of the data. The two first factors also already explain most of the variability in the test scores when principal factor analysis is used.

Table S5 reports factor loadings after an oblique rotation where the factors are allowed to be correlated. In a two-factor model, the cognitive test scores and the personality test scores load on distinct factors. Masculinity is only weakly related to other scores. It has large uniqueness and a factor loading of only 0.22.

As an alternative specification, we retained three factors. In a three-factor model, the consciousness-related scores “deliberation” and “dutifulness” load on a separate factor. The other five personality test scores and the three cognitive test scores still load on distinct factors, and masculinity has a low factor loading and large uniqueness.

Measurement error

The personality and cognitive ability tests are likely to measure the underlying traits with some error. This measurement error may stem from several sources. The test items may not fully capture the underlying personality traits. Some individuals may perform particularly poorly or particularly well in tests taken on a given day. Individuals also may make idiosyncratic errors in each test.

Measurement error causes a bias in the estimated coefficients of the anchoring regressions where the test scores are used as explanatory variables. In a simplest univariate case with classical measurement error, the regression coefficients would be attenuated towards zero. As a result also the differences across cohorts in the anchored test scores would be smaller than the differences in the underlying traits.

Furthermore, measurement error may be larger in personality tests than in cognitive tests (6) and therefore cause a larger downward bias in regressions where personality test scores are used as explanatory variables. Earlier work has shown that such bias may be large and substantially affect the comparisons between different demographic groups, particularly if the reliability of the test varies across groups (5).

To assess the likely direction and magnitude of bias caused by measurement error, we first simulate the effects of additional measurement error. We take i.i.d. random draws from a normal distribution with variance equal to 25, 50, 75 and 100 percent of the variance in the observed test scores, and add these additional errors to the observed test scores. We then re-estimate the anchoring equations. The results reported in Table S6 illustrate the that individual coefficients change to varying directions; adding error to all test scores increases some coefficients while decreasing others. As displayed in Fig S10 the aggregate effect of additional error is a reduction of cohort differences. For example, the difference between the youngest and the oldest cohort in anchored personality test scores declines from 11.6% when observed test scores are used to 9.6% when additional measurement error corresponding to 50% of original variance is added to each score. The corresponding decline in anchored cognitive test scores is from 10.4% to 8.2%.

In a univariate regression the effects of classical measurement error can be easily corrected if a ratio of variance of the true unobserved score and the variance of the observed erroneous score (reliability ratio) is known. The coefficient of erroneously measured variable is simply inflated by the reliability ratio. The method can be extended to a multivariate case as long as measurement errors are independent.

Unfortunately FDF has only reported a range of test-retest reliabilities in test scores rather than separate reliability ratios for each scale (7). Item-level data that would allow the estimation of scale-specific internal reliabilities are not available. However, as we discuss in detail below, instrumental variables and structural equation models can be used to adjust the estimates so that the effects of measurement error are taken into account (if the assumptions underlying these methods are valid). We emphasize that these adjustments only affect the estimates of the magnitude of cohort trends. The best evidence for the existence of cohort trends was shown above in section *Test score distributions* and in Figs S2-S5, where we demonstrate that test score distributions of later cohorts stochastically dominate those of earlier cohorts.

Instrumental variables estimates

One approach for correcting for measurement error is to combine multiple measurements using instrumental variables (IV) framework. In order for this approach to yield consistent estimates, we need instrumental variables that (i) are strongly correlated with the test scores (*first-stage*), and (ii) do not have an independent impact on the outcomes (*exclusion restriction*).

We use brother and twin test scores as instruments. 42 percent of men in our data have at least one brother (defined as a man born to the same mother) for whom we also observe the test scores. We also have 2,385 twin pairs (brothers born on same date) in data. We pick randomly one of the brothers or one of the twins to the estimation sample and use his brother's or his twin brother's scores as an instrument. In cases with more than two brothers in a family we only use one randomly chosen brother pair from each family.

Test scores within brother and twin pairs are highly correlated. The first eight columns of panels A, Tables S7 and S8, report the first-stage estimates i.e. regress each test score in turn on all the test scores of the brother. Panel B of Tables S7 and S8 report the corresponding results for twin data. The F-statistics in these regressions range between 181 and 957 in the brother sample and between 12 and 54 in the twin brother sample. It is also noteworthy that same trait coefficients are clearly larger than coefficients of other instruments.

Column 9 of Tables S7 and S8 report the IV-estimates. These coefficients are substantially larger than the OLS estimates using same brother or twin samples (column 10). Furthermore, OLS estimates from the brother and twin samples are quite similar to OLS estimates using the full sample (Column 11).

Panels A of Figs S11 and S12 report the resulting average anchored test scores by birth cohorts for personality traits. Point estimates based on OLS estimates from full data, brother data and twin data are rather similar, showing a €2,200–€2,500, or 10–12% increase in comparison to the 1962 baseline, in the anchored personality test scores. In comparison, anchored personality test scores using IV estimates suggest a €4,700, or 21%, increase in brother data and a €5100, or 22%, increase in the twin data. The results related to cognitive test scores are similar but the difference between the OLS and IV-estimates is smaller. Anchored cognitive test scores increase by €2200 when anchoring is based on OLS estimates, by €3300 when anchoring is based on IV estimates from the brother data and by €3400 when anchoring is based on IV estimates from the twin data.

The difference between OLS- and IV-based anchored test scores is consistent with measurement error leading to a substantial attenuation bias in the OLS estimates. However, it is also consistent with the exclusion restriction being violated. Brother's personality traits could have a direct impact on earnings or brother's personality could be correlated with unobserved factors that are shared by brothers and that have an effect on earnings.

Structural equations model

An alternative approach for examining the importance of measurement error is to use a structural equation model that combines a measurement model linking latent skills to test scores and a structural model linking latent skills to earnings.

Based on the exploratory factor analysis discussed above, we assume that there are two underlying unobserved latent factors, one related to cognitive skills and one related to personality. We treat the three cognitive test scores as error-ridden proxies of latent cognitive skills and the eight personality test scores as error-ridden proxies of latent non-cognitive skills. We allow for a possible correlation between these latent skills and assume that both the cognitive skills and personality are associated with earnings. We scale the latent variables by constraining the path from the latent variables on earnings to equal one.

More formally, the structural equation linking latent skills to earnings is

$$y_{it} = \alpha_t + \theta_{it}^k + \varepsilon_{it}$$

where y_{it} indicates later-life earnings of person i from cohort t, α_t is a cohort-specific constant and θ_{it}^k a latent index of trait k. Note that we are using a normalizing restriction and set the coefficients of latent traits in the structural equation to 1. The latent traits are related to observed test scores by measurement equations

$$P_{it}^s = \lambda_s^k \theta_{it}^k + \nu_{it}$$

where P_{it}^s is the sth observed test score related to latent trait k with s=1,2,3 for the cognitive test and s = 1, …, 8 for the personality test. The association between test scores and latent traits is described by factor loadings λ_s^k. ν_{it} is measurement error, i.e., variation in test scores not related to the variation in latent traits. We assume that these measurement errors are uncorrelated normal random variables.

Fig S13 describes the structure of the model in a path diagram and reports the estimated factor loadings as well as the estimates of correlation between latent factors.

We estimate the factor loadings (the effect of latent variables on observed test scores) and the error variances (variances of the observed test scores not explained by the latent variables) by fitting the model using the same data (men born between 1962–1976) and earnings measure (average earnings at ages 30–34) as for our regression-based analyses discussed above. The two-factor model provides a reasonably good fit to the data (CFI=0.83, RMSEA=0.11). The correlation between cognitive and non-cognitive latent factors is 0.41, suggesting that there are two correlated but distinct latent factors.

In columns 5-8 of Table S9 we report the differences in means of latent cognitive and personality factors by cohort. For comparison we also report, in columns 1-4, the corresponding cohort differences estimated using regression analysis.

Columns 5 and 6 of Table S9 report cohort trends from a model where personality test scores and cognitive test scores are anchored separately to later earnings. According to the estimates, mean of the latent personality factor increased between the 1962 and 1976 birth cohorts by an amount that corresponds to €2,546 higher earnings; the analogous figure for mean of latent cognitive factor is €2,328. These results are similar to our main results (Fig. 3), which are presented for ease of comparison in columns 1 and 2. The corresponding changes in our baseline anchored test scores are €2,474 for personality and €2,219 for cognitive skills.

The last two columns of Table S9 report results from a model corresponding to Fig S6, where both the cognitive skills and personality are anchored simultaneously to later earnings. The increase in the mean latent personality factor now corresponds to €1,481 and in the mean cognitive factor to €1,856. The corresponding regression-based results, reported in columns 3 and 4 (and in Fig S6B), are €1,586 and €1,793.

Selectivity in test score data

Finland is one of the few countries that have retained compulsory conscription system until present. All men are required to participate in either a military or a civilian service and roughly 80% choose the military service. Nevertheless, sample selectivity could affect the test scores, if selectivity into military service changes over time. We next analyze the effects of selectivity using data that cover the full population of men in these cohorts.

For those born between 1964 and 1976, we have test score data for 80% of men and this fraction remains roughly stable over time. For the earlier birth cohorts born in 1962 and 1963, we observe test scores for 66% and 76%, respectively. This smaller share in is due to men who started their service as "volunteers" (at an earlier age) before the test database was created in 1982.

Fig S14 reports the share of men serving in the military by their later income (measured as within-cohort earnings percentile rank at age 30). It shows that having served in the military is less common among the men who later appear in the bottom quintile of income distribution. However, apart from the early-1960s cohort, the selectivity pattern remains rather constant over time.

Table S10 reports results from two approaches examining the extent to which changes in selectivity into military service may affect the trends in the test scores. For reference, the first columns in panels A and B, report the changes between the 1962 and the 1976 birth cohorts without a selection correction. The following two columns report corresponding changes after reweighting the data so that observed characteristics remain constant over time. The two rightmost columns report lower and upper bounds for the change in the test scores allowing changes in selectivity also with respect to unobserved characteristics. We next describe both approaches in detail.

We use inverse probability reweighting (IPW) for constructing the results reported in columns 2 and 3. We denote the potential test score of the ith individual as r_i. The test scores are only observed for those men who served in the military. Let $z_i = 1$ if r_i is observed and $z_i = 0$ if r_i is not observed.

We first estimate the likelihood of having a non-missing test score $\hat{e} = P(z_i = 1|X_i)$ as a function of observed characteristics. We then reweight the data using these predicted probabilities, or propensity scores (8), yielding an estimator

$$L = \frac{1}{N}\sum_{i=1}^{N}\frac{z_i r_i}{\hat{e}_i}$$

This way the observations that are underrepresented in the available data due to larger than average fraction of missing observations in categories defined by observed characteristics are inflated by giving them a higher weight. As long as selectivity is based on observed characteristics only, this method produces unbiased estimates of population parameters.

In the current context, we first estimate $\hat{e}$ using a logit model separately in each cohort to predict whether a person has non-missing test score data. In the first specification, we use the total parental income, classified as deciles within each birth cohort, and father's and mother's education classified to four levels as explanatory variable. In the second specification we also add individual's own completed education and his earnings at age 30, again classified as deciles. We then use predicted values from these regressions to calculate weights for each person and calculate reweighted cohort averages as described above. We report these reweighted changes between birth cohorts 1962 and 1976, i.e. $\hat{L}_{1976} - \hat{L}_{1962}$, in the second and the third columns of panels A and B Table S10.

Overall, the baseline and selection corrected estimates are very similar to each other with the anchored personality test scores growing 4–10% slower and anchored cognitive ability test scores 7–15% slower in the selection corrected series than in the raw data. In terms of individual measures, selection correction has the largest impact on *deliberation* (6–16% slower growth), *dutifulness* (7–17%) and *verbal* (13–28%).

A limitation of the IPW approach is that it corrects for changes in selectivity that are due to characteristics observable in our data. It is naturally also possible that selectivity has changed in dimensions that are not included in our data and therefore cannot be corrected by reweighting by observed characteristics. Given that our data do not contain any variables that could be plausibly used as instruments to correct for changes in selection on unobservable characteristics, we adopt a bounding approach based on trimming the upper or lower part of the test score distribution as in (9) and (10).

The basic idea is the following. For the oldest 1962 birth cohort we have test score data for 66% of the male population. In comparison, for the 1976 birth cohort we have non-missing data for 80% of the male population. We construct a lower bound of changes in test scores by making an extreme assumption that the "additional" 14% of the population observed in the 1976 birth cohort are those at the top of the observed 1976 test score distribution. Hence by dropping the fraction corresponding to 14% of the population from the top of the 1976 test score distribution, we can calculate a conservative lower bound for the increase in the average scores. Similarly assuming that the additional 14% of population are at the bottom of the 1976 test score distribution and dropping this fraction from the bottom of the 1976 test score distribution yields a conservative upper bound for the increase in the test scores. The key assumption behind this bounding exercise is that the changes in the fraction of men serving in the military have a monotonous effect on the likelihood of any individual person to perform his military service.

The fourth and fifth columns of panels A and B, Table S10, report the results for this bounding exercise. The estimates suggest that anchored personality test scores increased between €1,426 and €3,990 and anchored cognitive ability test scores by €1,109 and €3,394.

We note that the 1962 birth cohort is a particularly challenging starting point, because we do not observe test scores for those who started service before 1982. As a robustness check, panels C and D, Table S10, report similar analysis as above, but using 1964 birth cohort as the starting point. The IPW approach now yields changes in anchored personality and cognitive ability test scores that are 1–5% and 3–8% smaller than in the raw data, respectively. Furthermore, changing the starting point by two years yields substantially tighter bounds suggesting that anchored personality test scores grew by €1,895–€2,232 and anchored cognitive ability scores by €1,403–€2,145.

Age at test

According to Conscription Act (452/1950) all male citizens of Finland were required to attend the military call-up during the year they turned 19. At the call-up they were assigned a date when they should report for service. Up until 1989 conscripts were assigned to service in the year following the call-up date, i.e., during the calendar year when they turn 20. It was also possible to apply to serve as a volunteer from age 17 onwards and to request postponing service up to age 30 due to reasons related to e.g. on-going education.

In 1988 the Conscription Act was amended and the call-up date moved to the year when the men turned 18. At the call-up the men were assigned to service within two years after the call-up date, i.e., in the years when they turned 19 or 20. As a result, the fraction of men entering military service at age 19 is higher starting from the 1971 birth cohort. In the government's proposal to the Parliament (HE 76/88) the amendment was motivated by the decrease in the size of draft cohorts and as an attempt to lessen disruptions to education by assigning men to service at an age when 75% on men finish their secondary education. Rules related to volunteering to early service and to postponing service remained essentially intact.

Table S11 shows that while the fraction tested at ages 18 and 19 increases at the time when the call-up date was moved, most men were tested at age 20 throughout the birth cohorts we examine. The table also reveals that postponing service by several years is rare: only 4% of men are 22 or older when taking the test.

Studies where the same test was given to same individuals at sparse intervals show that there are age effects on personality test scores (*3*). Therefore changes in the age of taking the test across cohorts could bias the estimated trends in cohort mean scores. Fig S15 presents the trends in anchored test scores by age at taking the test. Those taking the test at older age tend to get higher scores. However, these differences cannot be interpreted as age effects, because those deciding to take the military service at an unusual age are likely to differ from the rest of the population also in other dimensions. Nevertheless, Fig S15 show that the trends in the average test scores are unlikely to be driven by the slight decrease in the average age of taking the test, because trends are consistent across birth cohorts in each age category, with the only anomalies occurring in the under-19 category when their cohort share was below 10%.

Another way to see that changes in the test taking age are unlikely to drive our results is to estimate trends in personality traits, while keeping age at test constant. We do this using a simple regression adjustment, where we estimate

$$P_{is} = \alpha_s + \sum_{c=1963}^{1976} C_{ic}\beta_{cs} + \sum_{a=18}^{22} A_{ia}\gamma_{as} + \varepsilon_{is} \tag{2}$$

where P_{is} is subscore s of individual i, α_s is a constant, C_{ic} is an indicator variable taking value one if individual i was born in year c and zero otherwise (using birth cohort 1962 as omitted category) and A_{ia} is an indicator variable taking value one if he takes the test at age a and zero otherwise (categories are: "18 or less" (omitted category), "19", "20", "21" and "22 or more"). The parameters β_{cs} measure the difference in average test scores in trait s between birth cohort *c* and birth cohort 1962, while keeping the age-at-test distribution constant. We estimate equation 2 by running separate regressions for each personality trait *s*.

The results reported in Table S12, columns 2–3, show that the trends keeping test taking age constant are very similar as the baseline trends. The only large difference (in percentage terms) is for Masculinity, the only trait without a clear trend.

Validity of test responses

Another possible explanation for the secular increase in personality scores is that young men have become more adept at giving socially desirable answers. In this case the trends in personality traits could reflect systematic changes in measurement error.

A related concern is that as the same test is used for successive cohorts, test questions could be leaked and the content of the test could become more widely known over time. The test results are not published and generally not even revealed to the conscripts themselves. The test booklet is labeled as confidential and even sample questions are not publicly available. Yet, it is impossible to rule out the passing of information on test contents by earlier test takers to younger cohorts. However, incentives for gaming the test are not obvious. The conscripts are aware that the test is used as one of the criteria in selecting men to officer training but do not know how the test is scored. The scoring algorithm that FDF uses was published for the first time in (7).

One way of detecting such changes is to use the Lie-score from the Minnesota Multiphasic Personality Inventory (MMPI), which is also included in the FDF test. Lie-score measures attempts to give an overly favorable impression of one's conduct; high scores suggest that the person is attempting to "fake good".

As above, we use two approaches to examine whether the changes in Lie-scores are sufficiently large to explain the changes in the measured personality traits. First, Fig S16 reports anchored test scores by quintiles of the Lie-score. The quintiles are defined over all birth cohorts, i.e., the cutoff points for the underlying Lie-score remain constant, while the share of a birth cohort falling into each quintile changes over time. Those who score high in the Lie-score tend to have higher personality test scores and lower cognitive ability test scores. Importantly, however, we document clear upward trends in test scores within each Lie-score quintile.

Table S12, columns 4-5, reports results from similar regressions as those used above for keeping age at test constant over time. That is, we regress the personality test scores on a vector of year of birth indicator variables and a vector of Lie-score results (Lie-score of 20 and more are aggregated into one category). The trends are slightly less pronounced once we condition on Lie-scores. The largest difference between the adjusted and unadjusted trends are in *deliberation* and *dutifulness*, where adjusted increase between the 1962 and 1976 birth cohorts is 0.18–0.19 standard deviations in comparison to 0.26–0.27 standard deviations suggested by the unadjusted trends. For other personality measures, the adjusted changes in the measures are 5–10% smaller than unadjusted ones. Thus the trends in personality test scores do not appear to be driven by changes in the attempts of young men to give an overly favorable impression of themselves.

Cognitive ability and personality test scores

As an additional robustness check, we extend our analysis on the extent of which trends in personality traits are simply a reflection of a rise in cognitive ability. Above, we already reported results from anchoring personality and cognitive ability test scores jointly on later-life earnings. Table S12, columns 6–7, reports regressions estimates similar to those used for examining age at test and Lie-scores above. That is, we regress the personality test scores on a vector of year of birth indicator variables and a vector of cognitive ability test score results (40 indicator variables for each subtest). The adjusted trends in personality test scores are slightly less pronounced than the baseline trends, but remain economically and statistically significant.

Fig S17A reports trends in anchored personality scores by the quintiles of the anchored cognitive ability test scores. It shows an upward trend in test scores within each cognitive ability quintile. Thus we conclude that the trends in personality traits are a separate phenomenon from the trend in cognitive ability.

Trends in background variables

We now turn to the role of background variables in explaining trends in personality traits and cognitive ability. In order to understand the extent to which the trends in traits reflect changes in background variables, such as parental education, sibship size, or urbanization, we estimate a hypothetical distribution of test scores that would have prevailed if the 1962 cohort of conscripts had had the same distribution of background variables as the 1976 cohort of conscripts. This counterfactual distribution of test scores – when compared to the actual distribution of test scores of the 1976 birth cohort – provides a measure of how much of the between-cohort differences in traits can be attributed to differences in background variables.

Our decomposition follows the semi-parametric DFL methodology (11). More formally, let $f_t(p)$ denote the observed density of trait p for cohort t. We denote the full vector of observable characteristics with $\boldsymbol{X}$. Then the conditional density of p of the 1962 cohort, given its background characteristics, can be written as:

$$\begin{aligned} f_{62}(p) &= \int dF\big(p, \boldsymbol{X} \big| t_{p,x} = 62\big) \\ &= \int f\big(p \big| \boldsymbol{X}, t_p = 62\big) dF(\boldsymbol{X} | t_x = 62) \end{aligned} \qquad (7)$$

where $F\big(p, \boldsymbol{X} \big| t_{p,x} = 62\big)$ is the joint distribution of p and $\boldsymbol{X}$ of the cohort born in 1962. Following this notation, we can write the hypothetical, or counterfactual, density of the traits of the 1962 cohort with the distribution of $\boldsymbol{X}$ at their 1976 values as:

$$\begin{aligned} f\big(p; t_p = 62, t_x = 76\big) &= \int f\big(p \big| \boldsymbol{X}, t_p = 62\big) dF(\boldsymbol{X} | t_x = 76) \\ &= \int f\big(p \big| \boldsymbol{X}, t_p = 62\big) \Psi_x(\boldsymbol{X}) dF(\boldsymbol{X} | t_x = 62) \end{aligned} \qquad (8)$$

where the reweighting function, $\Psi_x(\boldsymbol{X})$, is defined as:

$$\Psi_x(\boldsymbol{X}) = \frac{dF(\boldsymbol{X}|t_x = 76)}{dF(\boldsymbol{X}|t_x = 62)} \tag{9}$$

This is simply the ratio of the probability mass at each point of $\boldsymbol{X}$ for the cohort born in 1976 relative to the cohort born in 1962. Applying Bayes' rule $\Psi_x(\boldsymbol{X})$ can be written as:

$$\Psi_x(\boldsymbol{X}) = \frac{P(t_x = 76|\boldsymbol{X})}{P(t_x = 62|\boldsymbol{X})}\frac{P(t_x = 62)}{P(t_x = 76)} \tag{10}$$

which implies that $\Psi_x(\boldsymbol{X})$ can be estimated using the pooled data of the 1962 and 1976 cohorts. The procedure starts by estimating a probit model where the probability of belonging to a cohort $t = 62{,}76$ is regressed on background characteristics $\boldsymbol{X}$**:**

$$P(t_x = t|\boldsymbol{X}) = P\big(\varepsilon > -\beta H(\boldsymbol{X})\big) = 1 - \Phi(-\beta H(\boldsymbol{X})) \tag{11}$$

where $\Phi(\cdot)$ is the cumulative normal distribution and $H(\boldsymbol{X})$ is a vector of background characteristics that is a function of $\boldsymbol{X}$. The unconditional probabilities $P(t_x = t)$ are equal to the weighted number of observations in the cohort t.

The set of background characteristics that we use in our analysis consist of indicator variables for age at test (18 or younger, 19, 20, or 21 or older), for the education level of the mother and the father of the conscript (secondary or less, lower tertiary, upper tertiary, unknown, or missing), for municipality type at childhood (rural, semi-urban, urban) and for sibship size (six or more siblings are aggregated into one category; we also include a dummy for the information on sibship size missing).

Fig S19 plots kernel estimates of the observed distributions of the personality and cognitive ability indices for the 1962 and 1976 cohorts. We use these distributions to construct the results reported in Tab 1 of the main paper for anchored personality test scores. First, we report the difference in the average tests scores between the *observed* test score distribution of the 1962 birth cohort (solid line) and the *observed* test score distribution of the 1976 birth cohort (dashed line). Next, we report the difference in the average tests scores between the *observed* 1962 test score distribution (solid line) and an average of the *counterfactual* distribution where we reweight the 1962 test score distribution to correspond to the 1976 distribution of background characteristics (dotted line). This comparison answers the question: how different would the average test scores for the 1962 birth cohort had been, if the 1962 birth cohort had had the same characteristics as the 1976 birth cohort and the association between background characteristics and test scores had remained at the level observed for the 1962 birth cohort. Finally, we report the ratio between the predicted and the observed change in average test scores, i.e. the share of the observed change in average test scores that can be attributed to changes in background characteristics.

Table S13 examines how much of the changes in average test scores can be attributed to each background characteristic. The first column shows the observed difference in the average test scores between the 1962 and 1976 birth cohorts. The remaining columns report results similar to those in Table 1, but now using only one background characteristic at a time. The results suggest that changes in test age, rural/urban status and sibship size explain quite little of the changes in average test scores, while much larger share can be attributed to changes parental education.

The Relation of the FDF test and the Five Factor Model

In modern personality psychology, the Five Factor Model of personality is one of the most robust and widely used models of personality structure (12). The five higher-order personality traits of the model include extraversion, neuroticism, agreeableness, conscientiousness, and openness to experience. Extraversion is related to sociability, assertiveness, and positive emotionality. Neuroticism is expressed as low emotional stability, low self-esteem, and heightened psychological vulnerability. Agreeableness reflects the person's cooperativeness, level of empathy, and general trust in other people. Conscientiousness characterizes the person's degree of self-discipline, self-efficacy, and orderliness. Openness to Experience can be observed in the person's intellectual adventurousness, curiosity, and artistic interests.

In order to see how the Finnish Defense Forces (FDF) personality traits relate to traits of the Five Factor Model (FFM), we administered online a short version of the FDF and a 60-item FFM personality test to a sample of 231 participants who were recruited via email lists of university students and people who had participated in open university courses at the University of Helsinki. The data were collected for the revision of our manuscript per a reviewer's request over a two-week period in January 2017. The mean age of the sample was 28.6 (SD=9.1), 87.5% were women, 75.2% were full-time students, and 19.5% had a full-time employment. We did not have access to the full FDF measure used in the main analysis but we had a shortened version that included 6 items per scale (48 items in total). The FFM traits were measured with a 60-item FFM measure used in previous Finnish studies (13-15). The participants rated each item on a 5-point scale ranging from 1=Strongly disagree to 5=Strongly agree. Of the 108 items, 46 items were reverse coded. Given the very small number of missing values in the data (n=125 missing responses of all the possible 108 items × 231 participants = 24,948 responses in total), all missing values in the items were imputed using the mean value of the item in the sample. We examined (i) the pairwise correlations between all the traits, (ii) how traits of the FFM predicted traits of the FDF, and (iii) how traits of the FDF predicted traits of the FFM.

Table S14 shows the pairwise correlations between all the traits. Table S15 shows multivariate models for each of the FDF traits. Sociability, leadership motivation, and activity–energy were strongly related to higher extraversion. Achievement striving, deliberation, and dutifulness were most strongly related to higher conscientiousness. Self-confidence correlated most strongly with lower neuroticism, and also with higher extraversion and higher conscientiousness. These associations provide convergent validity for the FDF traits, as the FDF traits match closely with the underlying contents of the Five Factor traits. The most marked correlations with agreeableness included negative correlations with leadership motivation and achievement striving, which may reflect the less considerate and cooperative tendencies associated with social dominance and competitiveness. Openness to experience had only moderate correlations with higher dutifulness and higher achievement striving.

Table S16 shows regression models in which each of the Five Factor traits is predicted by the FDF traits. As indicated by the proportions of variance explained, the FDF personality traits capture much of the variance in Extraversion, Conscientiousness, and Neuroticism, whereas variances in Agreeableness and Openness to Experience were less well captured.

In sum, the results from our test sample indicate that the FDF traits show convergent validity with standard measures of the Five Factor personality traits. As suggested by their labels, most of the FDF traits are related to extraversion and conscientiousness. In addition, self-confidence correlated strongly with lower neuroticism. Lower agreeableness was reflected in higher leadership motivation and achievement striving. It must be emphasized that we only used the short versions of the FDF traits, which may have weakened their psychometric properties, such as reliability, and even these results are based on a convenience sample.

References

1. T.N. Bond, K. Lang, The evolution of the Black-White test score gap in Grades K–3: The fragility of results. *Review of Economics and Statistics* 95, 1468-1479 (2013).

2. F. Cunha, J.J. Heckman, S.M. Schennach, Estimating the technology of cognitive and noncognitive skill formation. *Econometrica* 78, 883-931 (2010).

3. B. Jacob, J. Rothstein, The Measurement of Student Ability in Modern Assessment Systems. *Journal of Economic Perspectives* 30, 85-108 (2016).

4. G.F. Barrett. S.G Donald, Consistent Tests for Stochastic Dominance. *Econometrica* 71, 71–104 (2003).

5. T. N. Bond, K. Lang, The Black-White Education-Scaled Test-Score Gap in Grades K-7. *NBER Working Paper* 19243, (NBER, Cambridge, MA, 2013).

6. E. Grönqvist, B. Öckert, J. Vlachos, The Intergenerational Transmission of Cognitive and Non-Cognitive Abilities. *Journal of Human Resources*, forthcoming.

7. K. Nyman, *Varusmiesten johtajavalintojen luotettavuus*, National Defense University, Department of Behavioral Sciences, Publication Series 1, 1 (2007).

8. P.R. Rosenbaum, Model-based direct adjustment. *Journal of the American Statistical Association* 82, 387–394 (1987).

9. J. Angrist, E. Bettinger, M. Kremer, Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia. *American Economic Review* 96, 847-862 (2006).

10. D.S. Lee, Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies* 76, 1071-1102 (2009).

11. J. DiNardo, N.M. Fortin, T. Lemieux, Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach. *Econometrica* 64, 1001-1044 (1996).

12. J.J. Denissen, L. Penke, Motivational individual reaction norms underlying the Five-Factor model of personality: First steps towards a theory-based conceptual framework. *Journal of Research in Personality* 42, 1285–1302 (2008).

13. A. Pulver, J. Allik, L. Pulkkinen, M. Hämäläinen, A Big 5 personality inventory in 2 non-Indo-European languages. *European Journal of Personality* 9, 109–124 (1995).

14. J. Rantanen, R.L. Metsäpelto, T. Feldt, L. Pulkkinen, K. Kokko, Long-term stability in the Big Five personality traits in adulthood. *Scandinavian Journal of Psychology* 48, 511–518 (2007).

15. M. Jokela, L. Keltikangas-Järvinen, The association between low socioeconomic status and depressive symptoms depends on temperament and personality traits. *Personality and Individual Differences* 51, 302-308 (2011).

	Year of birth								
	1962	1963	1964	1965	1966	1967	1968	1969	1970
A: Personality									
Self-	19.9	20.4	20.8	21.2	21.4	21.6	21.9	21.8	22.2
confidence	(6.6)	(6.6)	(6.5)	(6.4)	(6.4)	(6.4)	(6.3)	(6.3)	(6.1)
Sociability	17.8	18.2	18.7	19.1	19.4	19.7	20.1	20.0	20.5
	(8.1)	(8.2)	(8.1)	(8.1)	(8.1)	(8.0)	(7.9)	(7.9)	(7.8)
Leadership	12.2	12.6	12.9	13.3	13.6	13.8	14.1	14.1	14.4
motivation	(7.6)	(7.7)	(7.7)	(7.8)	(7.9)	(7.9)	(7.9)	(7.9)	(7.8)
Activity-	14.5	14.8	15.0	15.2	15.4	15.5	15.7	15.7	16.0
energy	(5.4)	(5.4)	(5.4)	(5.4)	(5.5)	(5.5)	(5.4)	(5.4)	(5.3)
Achievement	12.3	12.5	12.7	12.7	12.9	13.0	13.1	13.1	13.2
striving	(5.0)	(5.1)	(5.1)	(5.1)	(5.1)	(5.1)	(5.1)	(5.1)	(5.0)
Dutifulness	10.4	10.5	10.4	10.4	10.5	10.5	10.7	10.6	10.8
	(3.7)	(3.7)	(3.7)	(3.7)	(3.8)	(3.8)	(3.8)	(3.8)	(3.7)
Deliberation	15.5	15.7	15.8	15.9	16.0	16.0	16.1	16.0	16.1
	(5.4)	(5.4)	(5.4)	(5.4)	(5.4)	(5.4)	(5.4)	(5.5)	(5.4)
Masculinity	18.4	18.4	18.5	18.5	18.5	18.5	18.5	18.5	18.5
	(2.9)	(3.0)	(2.9)	(3.0)	(3.0)	(3.0)	(3.0)	(3.0)	(2.9)
Anchored	21.4	21.7	21.9	22.1	22.3	22.4	22.5	22.5	22.7
test score	(4.5)	(4.6)	(4.6)	(4.6)	(4.6)	(4.6)	(4.6)	(4.6)	(4.6)
B: Cognitive ability									
Visuo-	23.3	23.7	23.8	23.9	24.3	24.2	24.4	24.5	24.8
spatial	(5.9)	(5.8)	(5.9)	(5.8)	(5.8)	(5.8)	(5.8)	(5.7)	(5.7)
Arithmetic	18.2	18.8	18.8	18.9	19.4	19.3	19.7	19.9	20.2
	(7.9)	(8.0)	(8.1)	(8.0)	(8.0)	(7.9)	(8.0)	(8.0)	(7.9)
Verbal	21.9	22.4	22.4	22.6	22.9	22.6	22.7	22.7	22.9
	(7.5)	(7.5)	(7.5)	(7.4)	(7.4)	(7.4)	(7.4)	(7.4)	(7.3)
General (sum)	42.6	50.0	52.2	52.9	48.0	50.6	53.8	54.2	53.8
	(33.6)	(32.0)	(31.0)	(30.8)	(33.9)	(32.5)	(31.4)	(31.3)	(32.1)
Anchored	21.4	21.8	21.8	21.9	22.2	22.1	22.3	22.4	22.7
test score	(4.9)	(4.9)	(5.0)	(4.9)	(4.9)	(4.9)	(4.9)	(4.9)	(4.8)
C: Lie-score	5.8	5.8	5.9	5.9	6.0	5.9	5.8	5.7	5.9
	(3.9)	(3.8)	(3.8)	(3.8)	(3.8)	(3.7)	(3.8)	(3.8)	(3.8)
Observations	26,676	30,781	31,774	31,243	31,194	30,211	29,767	27,693	26,570

Table S1: Means and standard deviations (in parentheses) of raw test scores. Anchored test scores are the predicted values of the regressions reported in the first and second column of Table S2.

	Year of birth								
	1971	1972	1973	1974	1975	1976	1977	1978	1979
A: Personality									
Self-	22.5	22.8	23.5	23.8	23.9	24.2	24.9	24.8	24.8
confidence	(5.9)	(5.7)	(5.5)	(5.3)	(5.2)	(5.0)	(4.2)	(4.2)	(4.3)
Sociability	20.8	21.2	21.9	22.1	22.2	22.5	22.8	22.9	22.8
	(7.7)	(7.5)	(7.3)	(7.3)	(7.2)	(7.1)	(6.7)	(6.7)	(6.8)
Leadership	14.8	15.1	15.7	15.8	16.1	16.3	17.2	17.2	17.2
motivation	(7.7)	(7.6)	(7.6)	(7.5)	(7.5)	(7.4)	(6.9)	(6.9)	(6.9)
Activity-	16.2	16.5	16.7	16.8	16.9	17.1	18.2	18.1	18.0
energy	(5.2)	(5.1)	(5.0)	(5.0)	(4.9)	(4.8)	(4.3)	(4.3)	(4.4)
Achievement	13.4	13.7	13.9	14.1	14.1	14.2	14.7	14.8	14.8
striving	(4.9)	(4.8)	(4.8)	(4.7)	(4.7)	(4.6)	(4.2)	(4.2)	(4.2)
Dutifulness	10.9	11.0	11.3	11.3	11.4	11.4	11.6	11.5	11.4
	(3.7)	(3.7)	(3.7)	(3.6)	(3.6)	(3.6)	(3.4)	(3.3)	(3.3)
Deliberation	16.3	16.4	16.8	16.9	16.9	17.0	17.6	17.4	17.4
	(5.4)	(5.3)	(5.2)	(5.2)	(5.2)	(5.1)	(4.8)	(4.8)	(4.8)
Masculinity	18.5	18.5	18.5	18.5	18.4	18.5	18.9	18.9	18.9
	(2.8)	(2.8)	(2.8)	(2.8)	(2.7)	(2.7)	(2.6)	(2.6)	(2.6)
Anchored	22.9	23.1	23.6	23.7	23.8	24.0	24.4	24.4	24.4
test score	(4.5)	(4.4)	(4.3)	(4.2)	(4.1)	(4.0)	(3.6)	(3.6)	(3.6)
B: Cognitive ability									
Visuo-	25.2	25.4	25.7	26.2	26.4	26.6	26.8	26.7	26.4
spatial	(5.5)	(5.5)	(5.5)	(5.4)	(5.3)	(5.3)	(5.3)	(5.4)	(5.6)
Arithmetic	20.7	20.9	20.9	21.2	21.2	21.3	21.4	21.2	20.7
	(7.9)	(7.8)	(7.8)	(7.7)	(7.6)	(7.7)	(7.6)	(7.6)	(7.6)
Verbal	23.1	23.2	22.9	23.2	23.4	23.5	23.4	23.0	22.4
	(7.1)	(7.1)	(7.0)	(6.9)	(6.9)	(6.8)	(6.9)	(6.8)	(6.9)
General (sum)	53.0	54.2	54.5	55.3	55.5	55.3	55.5	54.7	52.4
	(33.1)	(32.8)	(32.7)	(33.0)	(33.1)	(33.5)	(33.4)	(33.5)	(33.6)
Anchored	23.0	23.1	23.1	23.4	23.5	23.6	23.6	23.5	23.2
test score	(4.7)	(4.7)	(4.7)	(4.7)	(4.6)	(4.6)	(4.6)	(4.6)	(4.6)
C: Lie-score	5.8	5.9	6.4	6.5	6.4	6.5	6.9	6.8	6.7
	(3.8)	(3.8)	(4.0)	(4.0)	(4.1)	(4.1)	(4.1)	(4.0)	(4.0)
Observations	24,766	24,310	23,663	25,847	27,568	27,460	25,986	25,363	24,431

Table S1: (cont') Means and standard deviations (in parentheses) of raw test scores. Anchored test scores are the predicted values of the regressions reported in the first and second column of Table S2.

	Earnings at age 30–34			Earnings percentile rank at age 30			Lower tertiary degree or more		
A: Personality									
Self-confidence	1.15	.	0.75	0.007	.	0.001	0.068	.	0.049
	(0.05)		(0.05)	(0.001)		(0.001)	(0.001)		(0.001)
Sociability	-0.25	.	0.54	0.001	.	0.013	-0.087	.	-0.050
	(0.04)		(0.04)	(0.001)		(0.001)	(0.001)		(0.001)
Leadership	1.99	.	0.76	0.027	.	0.009	0.120	.	0.064
motivation	(0.04)		(0.04)	(0.001)		(0.001)	(0.001)		(0.001)
Activity-	1.92	.	0.30	0.031	.	0.007	0.097	.	0.024
energy	(0.04)		(0.04)	(0.001)		(0.001)	(0.001)		(0.001)
Achievement	0.78	.	1.09	0.015	.	0.019	0.030	.	0.045
striving	(0.03)		(0.03)	(0.001)		(0.001)	(0.001)		(0.001)
Dutifulness	-0.31	.	0.43	-0.004	.	0.007	-0.056	.	-0.023
	(0.04)		(0.04)	(0.001)		(0.001)	(0.001)		(0.001)
Deliberation	0.08	.	-0.26	0.000	.	-0.005	0.035	.	0.018
	(0.04)		(0.04)	(0.001)		(0.001)	(0.001)		(0.001)
Masculinity	0.03	.	-0.02	0.004	.	0.004	-0.034	.	-0.036
	(0.03)		(0.03)	(0.000)		(0.000)	(0.001)		(0.001)
B: Cognitive ability									
Visuospatial	.	1.39	1.11	.	0.02	0.02	.	0.02	0.01
		(0.04)	(0.04)		(0.00)	(0.00)		(0.00)	(0.00)
Verbal	.	0.96	0.45	.	0.01	0.00	.	0.09	0.07
		(0.04)	(0.04)		(0.00)	(0.00)		(0.00)	(0.00)
Arithmetic	.	3.18	2.86	.	0.04	0.04	.	0.13	0.12
		(0.04)	(0.04)		(0.00)	(0.00)		(0.00)	(0.00)
R^2	0.111	0.132	0.153	0.045	0.045	0.045	0.045	0.045	0.045
N	413,203	407,770	402,165	411,192	402,348	400,231	417,331	411,804	406,134

Table S2: Anchoring test scores. Regression coefficients and robust standard errors (in parentheses) from regressing later-life outcomes on personality test scores, cognitive ability test scores, and both. All regressions also control for year of birth fixed-effects. Test scores are scaled by the 1962 standard deviations. Earnings are in thousands of 2010 Euros. Earnings percentile is calculated within birth cohort of native-born men.

	A. *Personality*						B. *Cognitive ability*					
		Quantile						Quantile				
Coef.	Mean	.10	.25	.50	.75	.90	Mean	.10	.25	.50	.75	.90
1963	0.34	0.22	0.30	0.41	0.45	0.33	0.41	0.36	0.45	0.40	0.37	0.44
	(0.04)	(0.07)	(0.06)	(0.06)	(0.06)	(0.05)	(0.04)	(0.06)	(0.07)	(0.06)	(0.05)	(0.05)
1964	0.56	0.48	0.59	0.69	0.67	0.46	0.39	0.24	0.38	0.38	0.42	0.51
	(0.04)	(0.06)	(0.05)	(0.06)	(0.05)	(0.05)	(0.04)	(0.06)	(0.07)	(0.06)	(0.05)	(0.06)
1965	0.76	0.64	0.78	0.94	0.91	0.60	0.51	0.57	0.60	0.50	0.44	0.59
	(0.04)	(0.07)	(0.06)	(0.06)	(0.06)	(0.05)	(0.04)	(0.06)	(0.07)	(0.06)	(0.06)	(0.05)
1966	0.91	0.65	0.96	1.15	1.10	0.77	0.82	0.77	0.99	0.83	0.77	0.87
	(0.04)	(0.07)	(0.06)	(0.06)	(0.06)	(0.05)	(0.04)	(0.06)	(0.07)	(0.06)	(0.05)	(0.06)
1967	1.00	0.71	1.07	1.30	1.21	0.83	0.74	0.79	0.83	0.75	0.67	0.72
	(0.04)	(0.07)	(0.05)	(0.05)	(0.05)	(0.04)	(0.04)	(0.06)	(0.07)	(0.06)	(0.05)	(0.06)
1968	1.15	0.87	1.24	1.50	1.33	0.90	0.96	0.90	1.08	1.03	0.87	0.92
	(0.04)	(0.07)	(0.06)	(0.05)	(0.06)	(0.05)	(0.04)	(0.06)	(0.07)	(0.06)	(0.05)	(0.07)
1969	1.11	0.73	1.22	1.46	1.34	0.87	1.06	1.14	1.25	1.10	0.94	1.04
	(0.04)	(0.07)	(0.06)	(0.05)	(0.05)	(0.05)	(0.04)	(0.06)	(0.06)	(0.06)	(0.05)	(0.06)
1970	1.29	1.12	1.50	1.68	1.41	0.94	1.28	1.32	1.53	1.38	1.17	1.16
	(0.04)	(0.07)	(0.06)	(0.07)	(0.06)	(0.05)	(0.04)	(0.06)	(0.07)	(0.06)	(0.05)	(0.06)
1971	1.54	1.44	1.82	1.97	1.60	0.98	1.59	1.83	1.93	1.68	1.34	1.34
	(0.04)	(0.07)	(0.06)	(0.06)	(0.05)	(0.04)	(0.04)	(0.06)	(0.06)	(0.06)	(0.05)	(0.06)
1972	1.77	1.88	2.15	2.21	1.74	1.01	1.72	1.95	2.11	1.78	1.45	1.40
	(0.04)	(0.07)	(0.06)	(0.06)	(0.06)	(0.05)	(0.04)	(0.07)	(0.07)	(0.06)	(0.05)	(0.06)
1973	2.19	2.35	2.63	2.69	2.12	1.32	1.73	2.00	2.20	1.83	1.47	1.35
	(0.04)	(0.07)	(0.06)	(0.06)	(0.05)	(0.05)	(0.04)	(0.08)	(0.07)	(0.06)	(0.05)	(0.06)
1974	2.36	2.65	2.91	2.84	2.17	1.31	2.00	2.37	2.50	2.08	1.70	1.57
	(0.04)	(0.07)	(0.06)	(0.06)	(0.05)	(0.05)	(0.04)	(0.07)	(0.07)	(0.06)	(0.05)	(0.06)
1975	2.47	2.86	3.07	2.97	2.23	1.34	2.08	2.57	2.63	2.16	1.72	1.59
	(0.04)	(0.07)	(0.06)	(0.06)	(0.06)	(0.05)	(0.04)	(0.07)	(0.07)	(0.06)	(0.05)	(0.06)
1976	2.59	3.20	3.25	3.05	2.26	1.34	2.18	2.60	2.75	2.26	1.83	1.67
	(0.04)	(0.07)	(0.06)	(0.06)	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)	(0.06)	(0.05)	(0.06)
1977	3.04	4.25	4.03	3.42	2.35	1.30	2.22	2.77	2.78	2.26	1.82	1.72
	(0.03)	(0.07)	(0.05)	(0.05)	(0.05)	(0.04)	(0.04)	(0.07)	(0.07)	(0.06)	(0.05)	(0.06)
1978	3.03	4.20	4.04	3.41	2.25	1.29	2.11	2.65	2.65	2.18	1.77	1.62
	(0.04)	(0.06)	(0.05)	(0.05)	(0.05)	(0.04)	(0.04)	(0.07)	(0.07)	(0.06)	(0.05)	(0.06)
1979	2.97	4.16	3.95	3.35	2.24	1.24	1.77	2.25	2.23	1.81	1.45	1.37
	(0.04)	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)	(0.07)	(0.07)	(0.06)	(0.05)	(0.06)
Cons.	21.38	15.38	18.07	21.33	24.78	27.52	21.39	14.71	17.78	21.61	25.03	27.61
	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.05)	(0.04)	(0.04)	(0.04)

Table S3: Changes in anchored test scores. Estimates from OLS ("Mean") and quantile regressions, where anchored test scores are regressed on year of birth indicators (using 1962 as the omitted category) and a constant. Each entry measures changes in comparison to the 1962 birth cohort. Bootstrapped standard errors (in parentheses) are constructed using 250 replications.

	Self-confidence	Socia-bility	Leader-ship	Activity-energy	Achie-vement	Duti-fulness	Deli-beration	Mascu-linity	Anc-hored	Visuo-spatial	Arith-metic	Verbal	Anc-hored
A: Personality													
Self-confidence	1.00												
Sociability	0.71	1.00											
Leadership	0.66	0.76	1.00										
Activity-energy	0.63	0.60	0.68	1.00									
Achievement	0.50	0.48	0.70	0.61	1.00								
Dutifulness	0.51	0.44	0.56	0.50	0.55	1.00							
Deliberation	0.42	0.20	0.32	0.39	0.39	0.64	1.00						
Masculinity	0.21	0.12	0.10	0.28	0.11	0.03	0.10	1.00					
Anchored	0.82	0.67	0.84	0.74	0.87	0.70	0.59	0.19	1.00				
B: Cognitive ability													
Visuospatial	0.32	0.20	0.29	0.19	0.31	0.24	0.16	0.03	0.36	1.00			
Arithmetic	0.33	0.20	0.30	0.19	0.35	0.26	0.16	0.04	0.38	0.66	1.00		
Verbal	0.33	0.20	0.30	0.18	0.34	0.28	0.17	0.01	0.38	0.60	0.71	1.00	
Anchored	0.36	0.22	0.33	0.21	0.38	0.29	0.18	0.04	0.42	0.81	0.96	0.81	1.00

Table S4: Correlations of the test scores.

	Two factors			Three factors			
	Factor 1	Factor 2	Uniq.	Factor 1	Factor 2	Factor 3	Uniq.
Visuospatial	**0.74**	0.00	0.45	**0.75**	0.00	-0.02	0.45
Verbal	**0.79**	0.01	0.37	**0.80**	-0.01	0.00	0.37
Arithmetic	**0.84**	-0.02	0.31	**0.85**	-0.01	-0.03	0.31
Leadership motivation	0.02	**0.85**	0.26	0.03	**0.86**	0.00	0.22
Activity-Energy	-0.11	**0.84**	0.37	-0.11	**0.71**	0.17	0.37
Achievement	0.14	**0.68**	0.44	0.14	**0.45**	0.29	0.43
Self-Confidence	0.08	**0.76**	0.36	0.09	**0.72**	0.07	0.35
Deliberation	-0.01	**0.54**	0.71	-0.05	-0.14	**0.84**	0.45
Sociability	-0.09	**0.81**	0.40	-0.06	**1.00**	-0.23	0.29
Dutifulness	0.04	**0.69**	0.49	0.01	0.13	**0.71**	0.35
Masculinity	-0.05	0.22	0.96	-0.05	0.24	-0.02	0.96

Table S5: Factor loadings. Principle factor analysis, oblique rotation, loadings > 0.4 indicated with bold. See section *Exploratory factor analysis* for details.

	Added measurement error				
	0%	25%	50%	75%	100%
A: Personality					
Self-confidence	1.15	0.95	0.86	0.79	0.78
	(0.05)	(0.03)	(0.03)	(0.02)	(0.02)
Sociability	-0.25	0.19	0.33	0.36	0.40
	(0.04)	(0.03)	(0.03)	(0.02)	(0.02)
Leadership	1.99	1.48	1.25	1.09	0.96
motivation	(0.04)	(0.03)	(0.03)	(0.02)	(0.02)
Activity-	1.92	1.21	0.95	0.85	0.78
energy	(0.04)	(0.03)	(0.03)	(0.02)	(0.02)
Achievement	0.78	0.71	0.62	0.58	0.55
striving	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Dutifulness	-0.31	0.12	0.28	0.35	0.33
	(0.04)	(0.03)	(0.03)	(0.02)	(0.02)
Deliberation	0.08	0.38	0.45	0.48	0.48
	(0.04)	(0.03)	(0.02)	(0.02)	(0.02)
Masculinity	0.03	0.07	0.08	0.07	0.09
	(0.03)	(0.02)	(0.02)	(0.02)	(0.02)
R^2	0.111	0.106	0.102	0.099	0.097
N	413,203	413,203	413,203	413,203	413,203
B: Cognitive ability					
Visuospatial	1.39	1.39	1.27	1.22	1.16
	(0.04)	(0.03)	(0.02)	(0.02)	(0.02)
Verbal	0.96	1.20	1.23	1.13	1.11
	(0.04)	(0.03)	(0.03)	(0.02)	(0.02)
Arithmetic	3.18	2.41	2.04	1.82	1.60
	(0.04)	(0.03)	(0.03)	(0.02)	(0.02)
R^2	0.132	0.123	0.116	0.111	0.106
N	407,770	407,770	407,770	407,770	407,770

Table S6: Adding simulated i.i.d. measurement error. Regression coefficients and robust standard errors (in parentheses) from regressing average earnings at age 30–34 on test scores. All regressions also control for year of birth fixed-effects. The amount of measurement error is described as a percentage of the variance of the observed test scores. Test scores are scaled by the (observed) 1962 standard deviations. Earnings are in thousands of 2010 Euros. See section *Measurement error* for details.

	First-stage								IV estimates	OLS estimates	
	Leader-ship	Activity-energy	Achie-vement	Self-confidence	Deli-beration	Socia-bility	Duti-fulness	Mascu-linity	Brother sample	Brother sample	Full sample
A: Brothers											
Leadership motivation	0.26	0.10	0.14	0.10	0.03	0.13	0.10	-0.03	-0.54	1.32	1.15
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(1.50)	(0.12)	(0.05)
Activity-energy	-0.01	0.09	-0.01	-0.01	0.00	-0.01	-0.01	0.05	-4.78	-0.17	-0.25
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(1.34)	(0.10)	(0.04)
Achievement striving	0.05	0.04	0.13	0.03	0.02	0.03	0.05	0.00	7.23	1.94	1.99
	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(1.23)	(0.09)	(0.04)
Self-confidence	0.06	0.04	0.06	0.14	0.03	0.04	0.03	0.00	3.29	1.98	1.92
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(1.10)	(0.10)	(0.04)
Deliberation	-0.04	0.00	-0.02	-0.01	0.12	-0.03	0.03	0.01	2.87	0.78	0.78
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(1.11)	(0.08)	(0.03)
Sociability	-0.03	-0.03	-0.05	-0.02	-0.04	0.07	-0.03	0.00	3.75	-0.51	-0.31
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(1.34)	(0.11)	(0.04)
Dutifulness	0.04	0.02	0.05	0.02	0.05	0.01	0.13	-0.02	-1.17	0.02	0.08
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(1.30)	(0.09)	(0.04)
Masculinity	-0.04	0.01	-0.03	-0.02	-0.01	-0.02	-0.03	0.13	1.17	0.00	0.03
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.77)	(0.06)	(0.03)
N	71,489	71,489	71,489	71,489	71,489	71,489	71,489	71,489	71,489	71,489	413,203
R^2	0.12	0.07	0.09	0.10	0.04	0.09	0.07	0.02	0.006	0.114	0.111
F-stat for excl. instruments	957.1	477.6	734.3	526.8	328.3	542.6	546.6	180.7	.	.	.

Table S7: Anchoring personality test scores using brothers' test scores as instrumental variables. First-stage, IV and OLS coefficients and robust standard errors (in parentheses) from regressing later-life outcomes on personality test scores. All regressions also control for year of birth fixed-effects. Test scores are scaled by the 1962 standard deviations. See Appendix section *Measurement error* for details.

	First-stage								IV estimates	OLS estimates	
	Leader-ship	Activity-energy	Achie-vement	Self-confidence	Deli-beration	Socia-bility	Duti-fulness	Mascu-linity	Twins sample	Twins sample	Full sample
B: Twins											
Leadership motivation	0.32	0.12	0.13	0.09	0.00	0.17	0.07	0.04	-3.88	0.36	1.15
	(0.04)	(0.03)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)	(4.52)	(0.67)	(0.05)
Activity-energy	-0.01	0.13	0.01	0.00	-0.02	0.01	-0.02	0.06	2.23	0.48	-0.25
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(4.24)	(0.52)	(0.04)
Achievement striving	0.06	0.05	0.19	0.03	0.02	0.03	0.07	-0.02	6.88	1.65	1.99
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(4.32)	(0.49)	(0.04)
Self-confidence	0.02	0.04	-0.02	0.23	0.02	0.05	0.01	0.02	5.35	1.30	1.92
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(3.04)	(0.61)	(0.04)
Deliberation	-0.01	0.05	0.03	0.04	0.24	-0.03	0.07	0.06	1.72	0.22	0.78
	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(3.51)	(0.47)	(0.03)
Sociability	0.04	0.00	-0.01	-0.02	-0.03	0.17	-0.02	-0.05	1.02	0.16	-0.31
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(4.22)	(0.55)	(0.04)
Dutifulness	0.03	-0.02	0.07	-0.01	0.07	-0.02	0.20	-0.09	-2.82	0.60	0.08
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(5.21)	(0.47)	(0.04)
Masculinity	-0.04	0.01	0.01	-0.03	0.00	-0.04	-0.04	0.21	-1.99	0.12	0.03
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(2.46)	(0.36)	(0.03)
N	2,351	2,351	2,351	2,351	2,351	2,351	2,351	2,351	2,351	2,351	413,203
R^2	0.19	0.14	0.14	0.17	0.11	0.16	0.12	0.07	-0.004	0.115	0.111
F-stat for excl. instruments	54.0	35.3	43.7	34.5	30.3	42.1	35.5	11.9	.	.	.

Table S7: (cont') Anchoring personality test scores using twin brothers' test scores as instrumental variables. First-stage, IV and OLS coefficients and robust standard errors (in parentheses) from regressing later-life outcomes on personality test scores. All regressions also control for year of birth fixed-effects. Test scores are scaled by the 1962 standard deviations. See Appendix section *Measurement error* for details.

	First-stage			IV estimates	OLS estimates	
	Visuo-spatial	Verbal	Arith-metic	Brother/Twins sample	Brother/Twins sample	Full sample
A: Brothers						
Visuospatial	0.20	0.05	0.06	3.08	1.35	1.40
	(0.00)	(0.00)	(0.00)	(0.54)	(0.09)	(0.04)
Verbal	0.08	0.28	0.11	-1.46	0.95	0.97
	(0.01)	(0.00)	(0.01)	(0.49)	(0.10)	(0.04)
Arithmetic	0.12	0.12	0.31	5.21	3.18	3.17
	(0.01)	(0.01)	(0.01)	(0.53)	(0.10)	(0.04)
N	69,435	69,423	69,437	69,421	69,421	404,291
R^2	0.16	0.18	0.20	0.11	0.13	0.13
F-stat for excl. instruments	3376.0	4787.8	5169.7	.	.	.
B: Twins						
Visuospatial	0.31	0.05	0.03	3.53	1.69	1.40
	(0.03)	(0.03)	(0.03)	(1.61)	(0.45)	(0.04)
Verbal	0.14	0.43	0.14	-1.04	1.03	0.97
	(0.03)	(0.03)	(0.03)	(1.59)	(0.46)	(0.04)
Arithmetic	0.11	0.13	0.45	3.16	2.34	3.17
	(0.03)	(0.03)	(0.03)	(1.65)	(0.51)	(0.04)
N	2,245	2,244	2,245	2,351	2,351	413,203
R^2	0.28	0.34	0.35	0.00	0.11	0.11
F-stat for excl. instruments	221.1	318.4	346.2	.	.	.

Table S8: Anchoring cognitive ability test scores using brothers' (panel A) or twin brothers' (panel B) test scores as instrumental variables. First-stage, IV and OLS coefficients and robust standard errors (in parentheses) from regressing later-life outcomes on personality test scores. All regressions also control for year of birth fixed-effects. Test scores are scaled by the 1962 standard deviations. See Appendix section *Measurement error* for details.

	OLS				SEM			
	Anchored separately		Anchored jointly		Anchored separately		Anchored jointly	
	Personality	Cognitive	Personality	Cognitive	Personality	Cognitive	Personality	Cognitive
1962 (baseline)	0	0	0	0	0	0	0	0
1963	323	419	185	348	292	466	164	394
1964	539	401	337	327	506	440	289	366
1965	724	522	457	421	719	587	409	484
1966	874	831	652	663	870	911	578	740
1967	959	750	590	597	980	798	540	632
1968	1,104	974	702	793	1,157	1,023	657	813
1969	1,059	1,079	663	882	1,098	1,122	624	887
1970	1,232	1,305	786	1,058	1,326	1,365	754	1,077
1971	1,464	1,613	938	1,314	1,562	1,679	892	1,329
1972	1,679	1,749	1,081	1,427	1,779	1,827	1,015	1,449
1973	2,087	1,758	1,329	1,440	2,164	1,806	1,238	1,433
1974	2,255	2,039	1,416	1,659	2,294	2,119	1,313	1,684
1975	2,357	2,126	1,495	1,719	2,411	2,228	1,393	1,771
1976	2,474	2,219	1,586	1,793	2,546	2,328	1,481	1,856

Table S9: Mean differences of regression-anchored test scores (see Fig. 3) and of factor means by cohort in comparison to the average of the 1962 birth cohort. Anchoring variable is earnings at age 30–34 in 2010 euros.

	Baseline	IPW		Bounds	
		Spec. 1	Spec. 2	Lower	Upper
A: Changes in personality between 1962–1976					
Self-confidence	0.65	0.63	0.60	0.50	0.91
Sociability	0.58	0.57	0.55	0.39	0.87
Leadership motivation	0.55	0.53	0.51	0.27	0.86
Activity-energy	0.47	0.46	0.44	0.29	0.79
Achievement striving	0.38	0.36	0.33	0.08	0.67
Dutifulness	0.27	0.25	0.23	0.06	0.67
Deliberation	0.26	0.25	0.22	0.07	0.63
Masculinity	0.03	0.03	0.03	-0.29	0.38
Anchored test score	2,588	2,486	2,335	1,426	3,990
B: Changes in cognitive ability between 1962–1976					
Visuospatial	0.55	0.53	0.50	0.30	0.80
Arithmetic	0.40	0.37	0.33	0.15	0.66
Verbal	0.21	0.18	0.15	-0.04	0.48
Anchored test score	2,176	2,015	1,842	1,109	3,394
C: Changes in personality between 1964–1976					
Self-confidence	0.51	0.51	0.49	0.50	0.51
Sociability	0.47	0.47	0.46	0.46	0.47
Leadership motivation	0.45	0.45	0.44	0.43	0.47
Activity-energy	0.38	0.37	0.36	0.36	0.38
Achievement striving	0.30	0.29	0.28	0.28	0.30
Dutifulness	0.25	0.24	0.23	0.24	0.25
Deliberation	0.21	0.20	0.18	0.19	0.22
Masculinity	0.00	0.00	-0.01	-0.01	0.00
Anchored test score	2,025	2,002	1,923	1,895	2,232
D: Changes in cognitive ability between 1964–1976					
Visuospatial	0.48	0.47	0.45	0.38	0.57
Arithmetic	0.32	0.31	0.29	0.25	0.39
Verbal	0.15	0.14	0.12	0.07	0.23
Anchored test score	1,782	1,732	1,633	1,403	2,145

Table S10: Selectivity. The first column reports changes in average test scores between the 1962 and 1976 birth cohorts. Test scores are scaled by the 1962 standard deviations. Other columns report the corresponding changes adjusted for changes in selection into military service. See section *Selectivity in test score data* for details.

	Year of birth					
	1962–64	1965–67	1968–70	1971–73	1974–76	1977–79
A: Age at test						
18 or less	0.06	0.09	0.09	0.13	0.24	0.99
19	0.48	0.49	0.47	0.49	0.52	0.01
20	0.37	0.35	0.36	0.30	0.18	0.00
21	0.04	0.03	0.04	0.04	0.03	0.00
22 or more	0.05	0.04	0.05	0.04	0.03	0.00
B: Mother's level of education						
Secondary	0.23	0.28	0.33	0.38	0.41	0.44
Tertiary (lower)	0.09	0.11	0.14	0.17	0.21	0.23
Tertiary (higher)	0.01	0.02	0.03	0.04	0.05	0.06
Less / unknown	0.67	0.59	0.51	0.41	0.33	0.27
C: Father's level of education						
Secondary	0.19	0.23	0.27	0.31	0.35	0.38
Tertiary (lower)	0.10	0.12	0.14	0.17	0.18	0.20
Tertiary (higher)	0.04	0.05	0.06	0.07	0.08	0.09
Less / unknown	0.66	0.60	0.53	0.45	0.39	0.33
D: Municipality type						
Urban	0.59	0.61	0.63	0.64	0.65	0.64
Semi-urban	0.18	0.17	0.17	0.18	0.17	0.18
Rural	0.23	0.21	0.20	0.18	0.17	0.18
E: Sibship size						
1	0.08	0.10	0.12	0.12	0.11	0.10
2	0.31	0.36	0.40	0.44	0.44	0.43
3	0.27	0.27	0.27	0.27	0.27	0.29
4	0.16	0.14	0.12	0.10	0.10	0.11
5	0.08	0.06	0.05	0.03	0.03	0.03
6 or more	0.10	0.07	0.05	0.04	0.04	0.04
Annual earnings at age 30	19,220	20,730	24,000	26,360	28,410	29,190
Annual earnings at age 30–34	21,490	24,160	26,930	29,380	31,400	.
Parental income	32,956	38,086	41,006	44,912	47,688	49,989
Individuals	120,337	115,517	104,428	92,013	100,794	99,303
... with test scores	89,231	92,648	84,030	72,739	80,875	75,780
Share with test scores	0.74	0.80	0.80	0.79	0.80	0.76

Table S11: Means of background variables and later-life outcomes by three-year birth cohorts. Municipality type is based on the municipality of residence in the first census year after the year of birth. Sibship size is the number of children with the same biological mother. Earnings are measured as the sum of annual labor market income and entrepreneurial income. Parental income is measured as the sum of father's and mother's annual earnings, taxable transfers, and capital income, and averaged over the period when the child was 10–25 years old. Earnings and income are measured in year 2010 Euros.

		Conditional on					
	Baseline	Age at test (%)		Lie-score (%)		Cognitive ability (%)	
A: Personality							
Self-confidence	0.65	0.66	*-2*	0.60	*8*	0.53	*18*
Sociability	0.58	0.58	*0*	0.55	*5*	0.51	*12*
Leadership motivation	0.55	0.56	*-2*	0.51	*7*	0.43	*22*
Activity-energy	0.47	0.45	*4*	0.42	*11*	0.40	*15*
Achievement striving	0.38	0.38	*0*	0.36	*5*	0.24	*37*
Dutifulness	0.27	0.31	*-15*	0.19	*30*	0.17	*37*
Deliberation	0.26	0.28	*-8*	0.18	*31*	0.20	*23*
Masculinity	0.03	-0.02	*167*	0.03	*0*	0.02	*33*
Anchored test score	0.57	0.58	*-2*	0.52	9	0.43	*25*
B: Cognitive ability							
Visuospatial	0.55	0.58	*-4*	0.56	*-2*	.	.
Arithmetic	0.40	0.43	*-9*	0.41	*-4*	.	.
Verbal	0.21	0.25	*-19*	0.24	*-10*	.	.
Anchored test score	0.44	0.48	*-8*	0.46	*-4*	.	.

Table S12: Robustness checks. The first column reports changes in average test scores between the 1962 and 1976 birth cohorts. Other columns report changes adjusted for changes in age at test, Lie-scores, and cognitive ability test scores (see sections *Age at test, Validity of test responses,* and *Cognitive ability and personality test scores* for details). Each supercolumn reports the conditional change and the percentage of baseline change attributable to each adjustment (in *italics*). Test scores are scaled by the 1962 standard deviations.

		Predicted change in test score due to changes in:									
	Baseline	Age at test		Mother's education		Father's education		Municipality type		Sibship size	
A: Personality											
Self-confidence	0.65	-0.04	*-6*	0.12	*19*	0.10	*16*	0.02	*3*	0.05	*8*
Sociability	0.58	-0.02	*-3*	0.09	*16*	0.08	*14*	0.02	*3*	0.05	*8*
Leadership motivation	0.55	-0.04	*-8*	0.14	*25*	0.12	*22*	0.02	*4*	0.05	*9*
Activity-energy	0.47	-0.02	*-4*	0.07	*14*	0.05	*11*	0.01	*1*	0.02	*5*
Achievement striving	0.38	-0.03	*-8*	0.11	*30*	0.11	*28*	0.01	*4*	0.04	*11*
Dutifulness	0.27	-0.04	*-15*	0.10	*36*	0.09	*31*	0.00	*2*	0.03	*12*
Deliberation	0.26	-0.03	*-11*	0.04	*16*	0.04	*14*	0.00	*-2*	0.02	*8*
Masculinity	0.03	0.02	*68*	-0.04	*-134*	-0.02	*-68*	0.00	*-4*	0.01	*23*
Anchored test score	0.57	-0.04	*-8*	0.14	*24*	0.12	*21*	0.02	*3*	0.05	*9*
B: Cognitive ability											
Visuospatial	0.55	-0.04	*-6*	0.14	*30*	0.12	*26*	0.02	*4*	0.05	*12*
Arithmetic	0.40	-0.03	*-14*	0.16	*47*	0.14	*40*	0.02	*5*	0.07	*19*
Verbal	0.21	-0.06	*-26*	0.19	*88*	0.16	*72*	0.02	*9*	0.08	*37*
Anchored test score	0.44	-0.06	*-13*	0.19	*45*	0.15	*38*	0.02	*5*	0.08	*18*

Table S13: DFL decomposition of the changes in test scores one variable at a time. The first column reports the actual observed difference between the means of the 1976 and 1962 cohort distributions. Other columns report the difference between the mean of the counterfactual distribution that would have prevailed if the 1962 cohort had had the same distribution of that particular background characteristic as the 1976 cohort; this difference is reported both in standard deviation units and as a percentage share of the total observed change for that particular trait.

	1	2	3	4	5	6	7	8	9	10	11	12
1. Extraversion												
2. Neuroticism	**-0.55**											
3. Conscientiousness	**0.14**	**-0.15**										
4. Agreeableness	**0.38**	**-0.26**	0.12									
5. Openness to Experience	**0.28**	-0.09	-0.02	**0.22**								
6. Leadership motivation	**0.59**	**-0.35**	**0.15**	0.06	**0.22**							
7. Activity–Energy	**0.47**	**-0.38**	**0.34**	0.10	**0.08**	**0.32**						
8. Achievement striving	0.10	0.03	**0.40**	**-0.19**	**0.12**	**0.22**	**0.14**					
9. Self-confidence	**0.52**	**-0.64**	**0.28**	**0.31**	**0.07**	**0.36**	**0.34**	0.05				
10. Deliberation	**-0.43**	**0.34**	**0.47**	**-0.15**	**-0.22**	**-0.24**	-0.08	**0.24**	**-0.20**			
11. Sociability	**0.78**	**-0.43**	0.04	**0.26**	**0.24**	**0.70**	**0.36**	0.05	**0.48**	**-0.42**		
12. Dutifulness	0.08	-0.11	**0.30**	**0.20**	**0.21**	**0.15**	0.04	**0.27**	**0.21**	0.08	0.00	
13. Masculinity	-0.01	-0.11	0.10	-0.10	-0.15	0.02	**0.22**	0.00	0.11	0.09	-0.09	0.02

Table S14: Pearson correlation coefficients between Finnish Defence Forces (FDF) and Five Factor Model (FFM) traits in a convenience sample of 231 participants. Correlations printed in bold are statistically significant ($p<0.05$). See section *The Relation of the FDF test and the Five Factor Model* for details.

	Leadership motivation	Activity–Energy	Achievement striving	Self-confidence	Deliberation	Sociability	Dutifulness	Masculinity
Extraversion	0.606	0.394	0.186	0.220	-0.355	0.800	-0.108	-0.034
	(0.067)	(0.071)	(0.073)	(0.062)	(0.061)	(0.053)	(0.079)	(0.083)
Conscientiousness	0.085	0.282	0.438	0.168	0.545	-0.061	0.293	0.090
	(0.053)	(0.057)	(0.058)	(0.049)	(0.048)	(0.042)	(0.062)	(0.066)
Neuroticism	-0.049	-0.150	0.133	-0.467	0.217	-0.003	-0.073	-0.154
	(0.063)	(0.066)	(0.069)	(0.058)	(0.057)	(0.050)	(0.074)	(0.078)
Agreeableness	-0.216	-0.116	-0.310	0.101	0.004	-0.041	0.139	-0.108
	(0.057)	(0.060)	(0.062)	(0.052)	(0.052)	(0.045)	(0.067)	(0.071)
Openness to Experience	0.083	-0.015	0.158	-0.047	-0.088	0.022	0.205	-0.131
	(0.055)	(0.058)	(0.060)	(0.050)	(0.050)	(0.043)	(0.065)	(0.068)
N	231	231	231	231	231	231	231	231
R^2	0.389	0.325	0.276	0.487	0.498	0.62	0.159	0.06

Table S15: Multivariate regression models predicting the Finnish Defence Forces (FDF) personality traits with Five Factor Model (FFM) traits in a convenience sample of 231 participants. Values are standardized beta coefficients (and their standard errors) of 8 multivariate linear regression models predicting each of the FDF trait with all FFM traits. See section *The Relation of the FDF test and the Five Factor Model* for details.

	Extraversion	Conscienti-ousness	Neuroticism	Agreeableness	Openness to Experience
Leadership motivation	0.033	0.033	-0.083	-0.232	0.021
	(0.055)	(0.069)	(0.071)	(0.086)	(0.091)
Activity-Energy	0.190	0.265	-0.169	0.040	0.038
	(0.043)	(0.054)	(0.055)	(0.067)	(0.071)
Achievement striving	0.056	0.191	0.056	-0.241	0.096
	(0.041)	(0.052)	(0.054)	(0.065)	(0.069)
Self-confidence	0.144	0.238	-0.514	0.205	-0.100
	(0.045)	(0.057)	(0.058)	(0.071)	(0.074)
Deliberation	-0.157	0.489	0.201	0.020	-0.189
	(0.043)	(0.055)	(0.056)	(0.067)	(0.071)
Sociability	0.553	0.002	0.019	0.322	0.162
	(0.061)	(0.078)	(0.080)	(0.096)	(0.101)
Dutifulness	0.034	0.146	-0.013	0.249	0.214
	(0.040)	(0.051)	(0.052)	(0.063)	(0.067)
Masculinity	-0.017	-0.034	-0.030	-0.108	-0.126
	(0.039)	(0.050)	(0.051)	(0.062)	(0.065)
N	231	231	231	231	231
R^2	0.692	0.508	0.485	0.242	0.155

Table S16: Multivariate regression models predicting the Five Factor Model (FFM) personality traits with Finnish Defence Forces (FDF) personality traits in a convenience sample of 231 participants. Values are standardized beta coefficients (and their standard errors) of 5 multivariate linear regression models predicting each of the FFM traits with all the FDF traits. See section *The Relation of the FDF test and the Five Factor Model* for details.

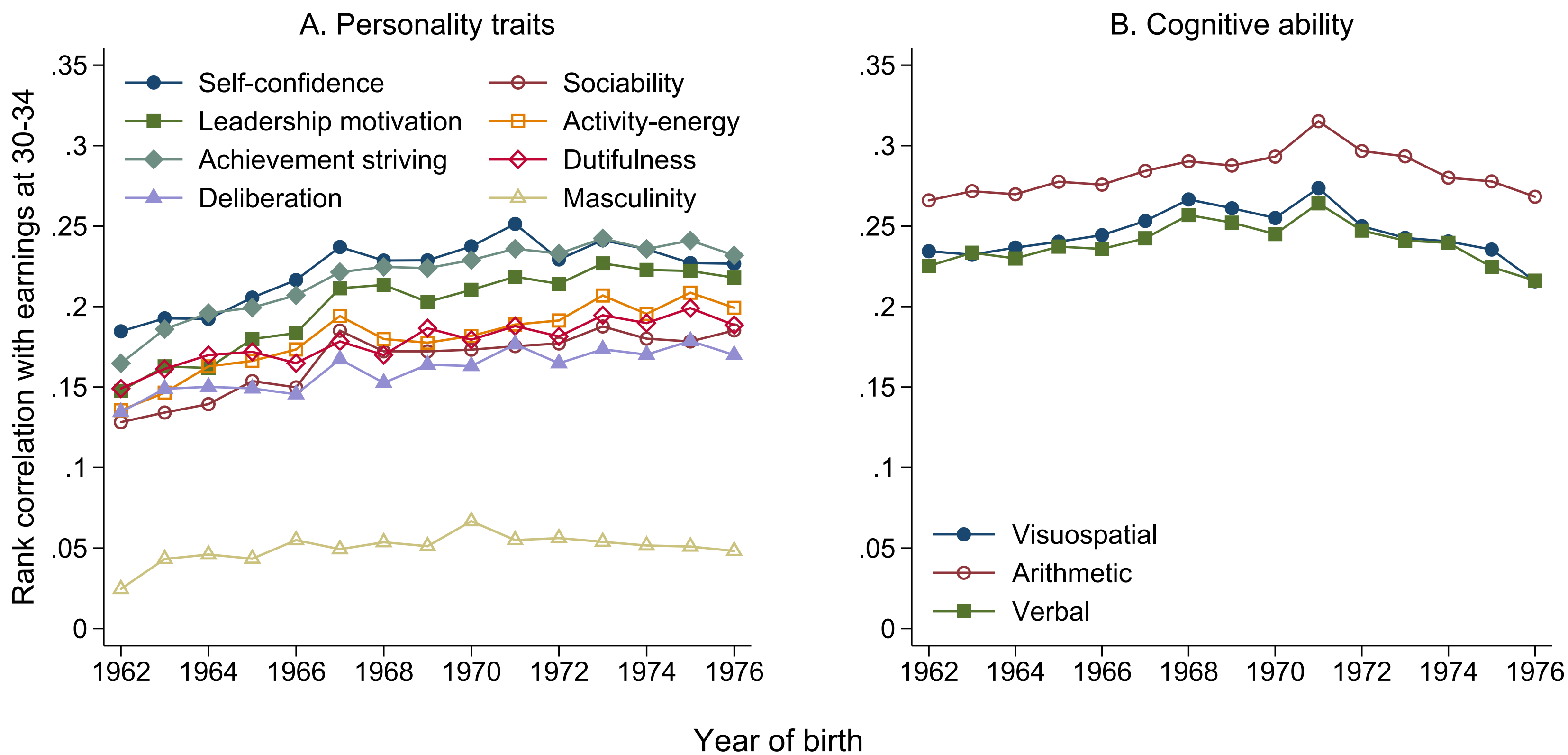

Figure S1: The relation of earnings and (**A**) personality traits and (**B**) cognitive ability by birth cohort. The relations are measured as the within-cohort rank correlation between the test score and earnings at age 30–34.

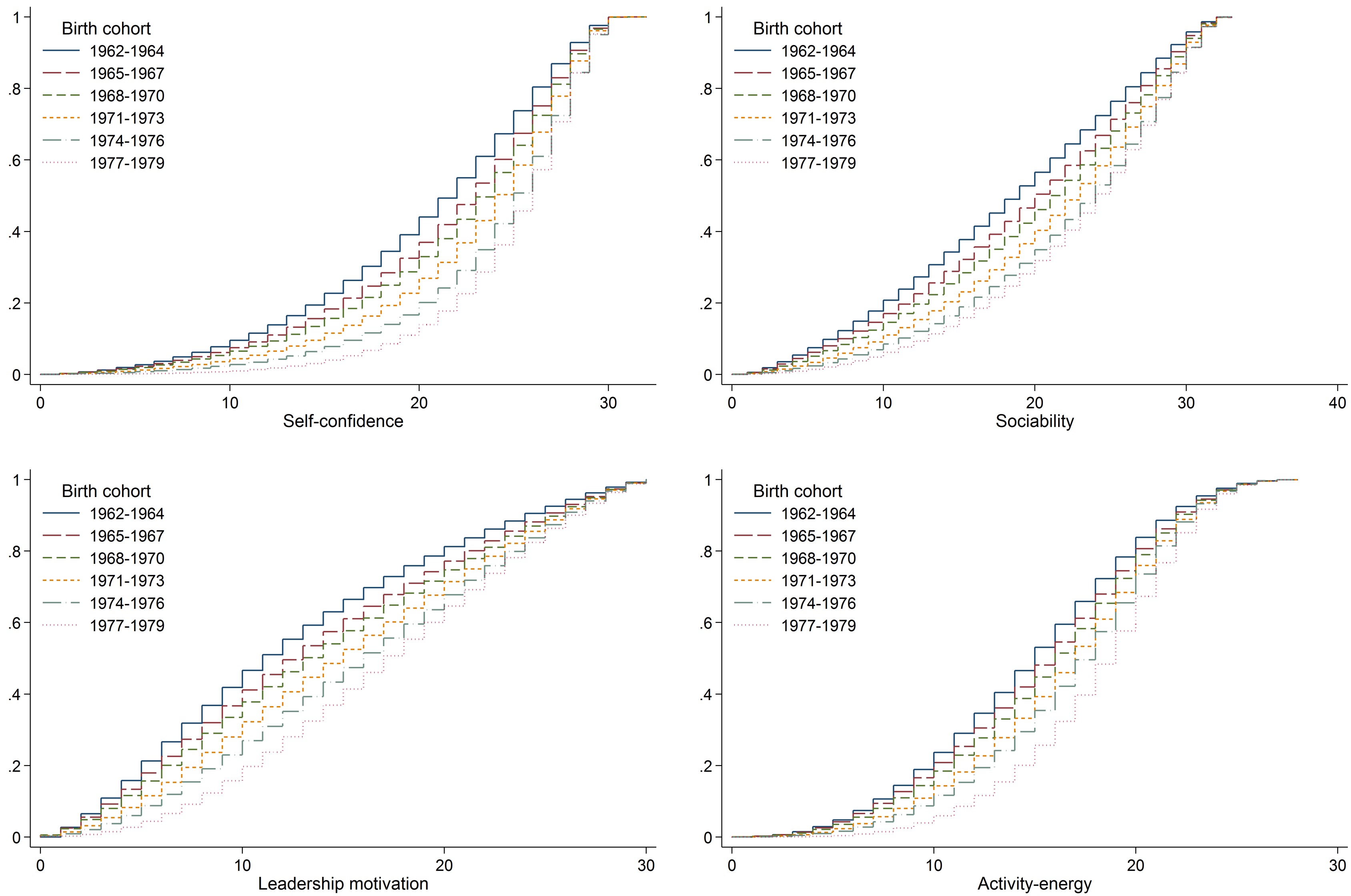

Figure S2: (1) CDFs of the raw personality test scores by birth cohort

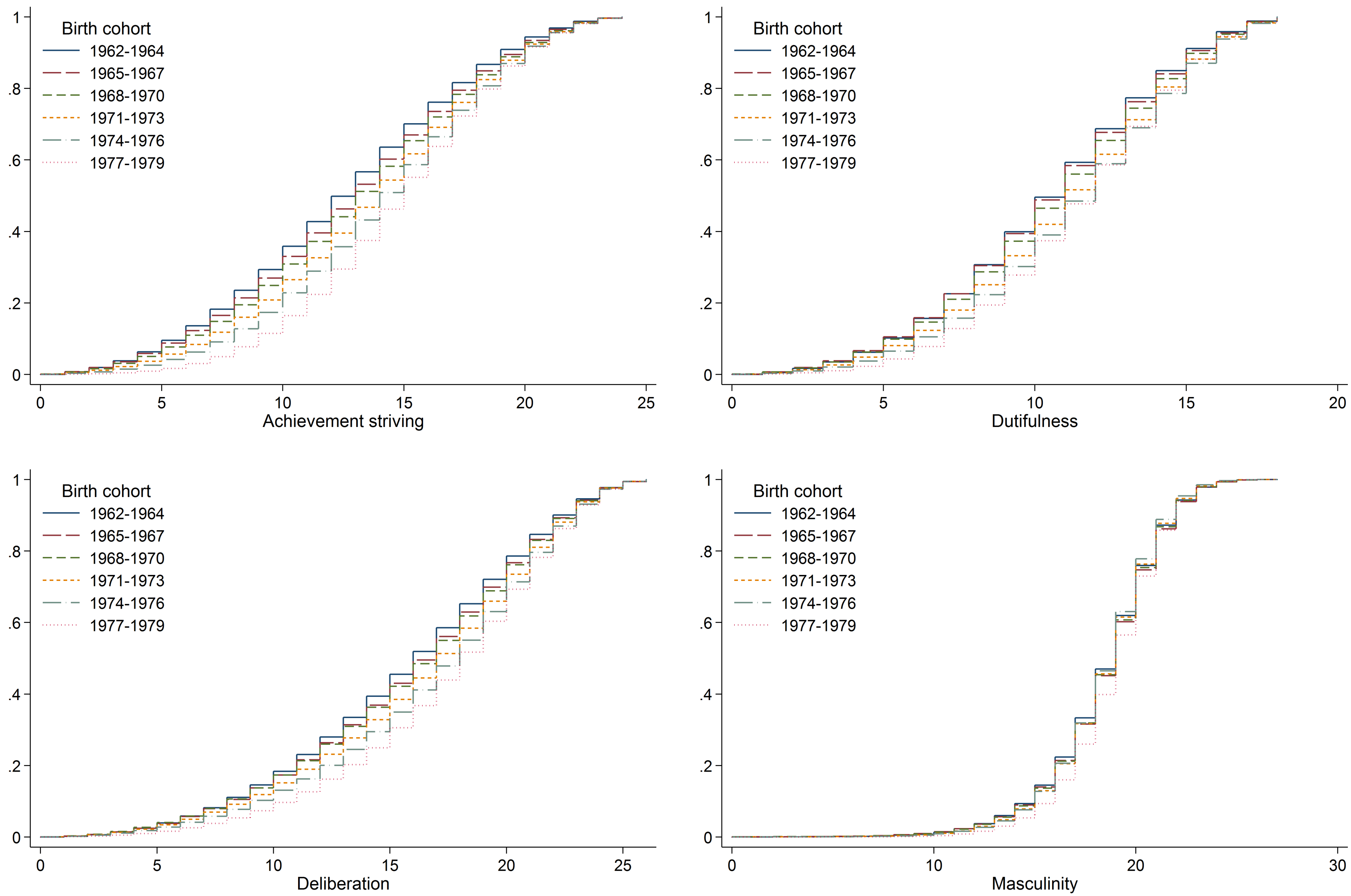

Figure S2: (2) CDFs of the raw personality test scores by birth cohort

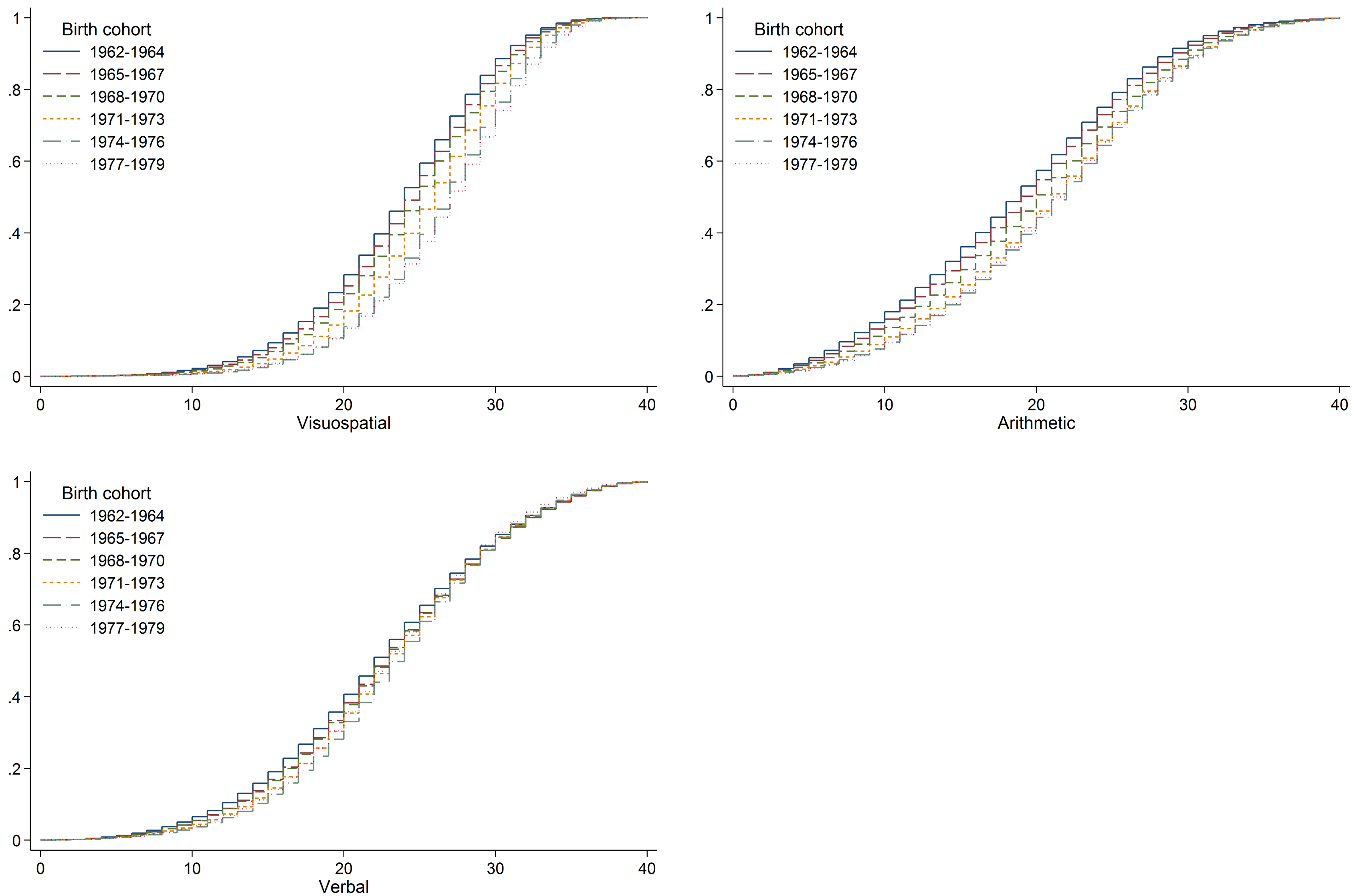

Figure S3: CDFs of the raw cognitive ability test scores by birth cohort

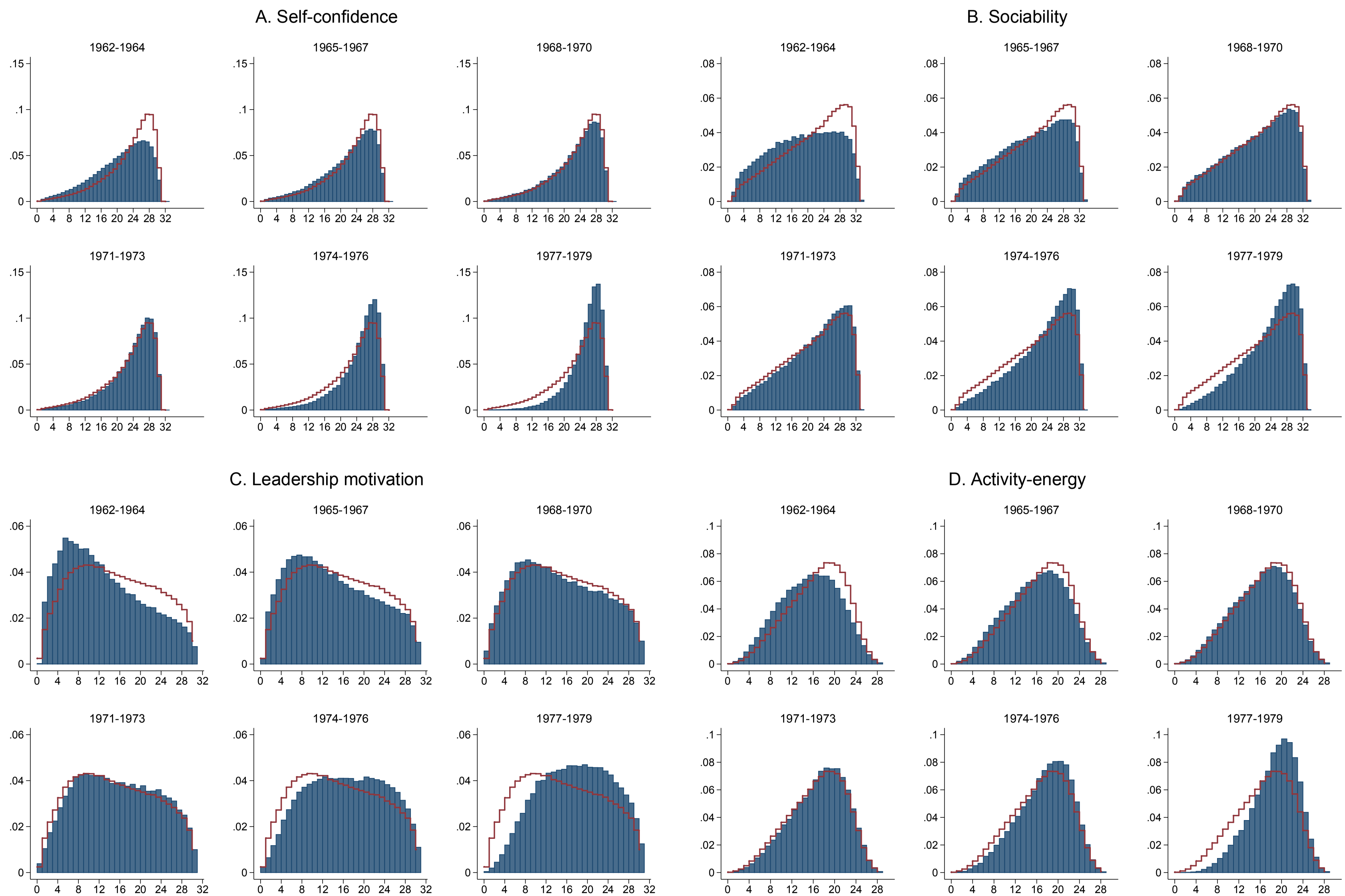

Figure S4: (1) Distributions of the raw personality test scores by birth cohort. The red curves depict the distributions for the 1962–1976 cohorts.

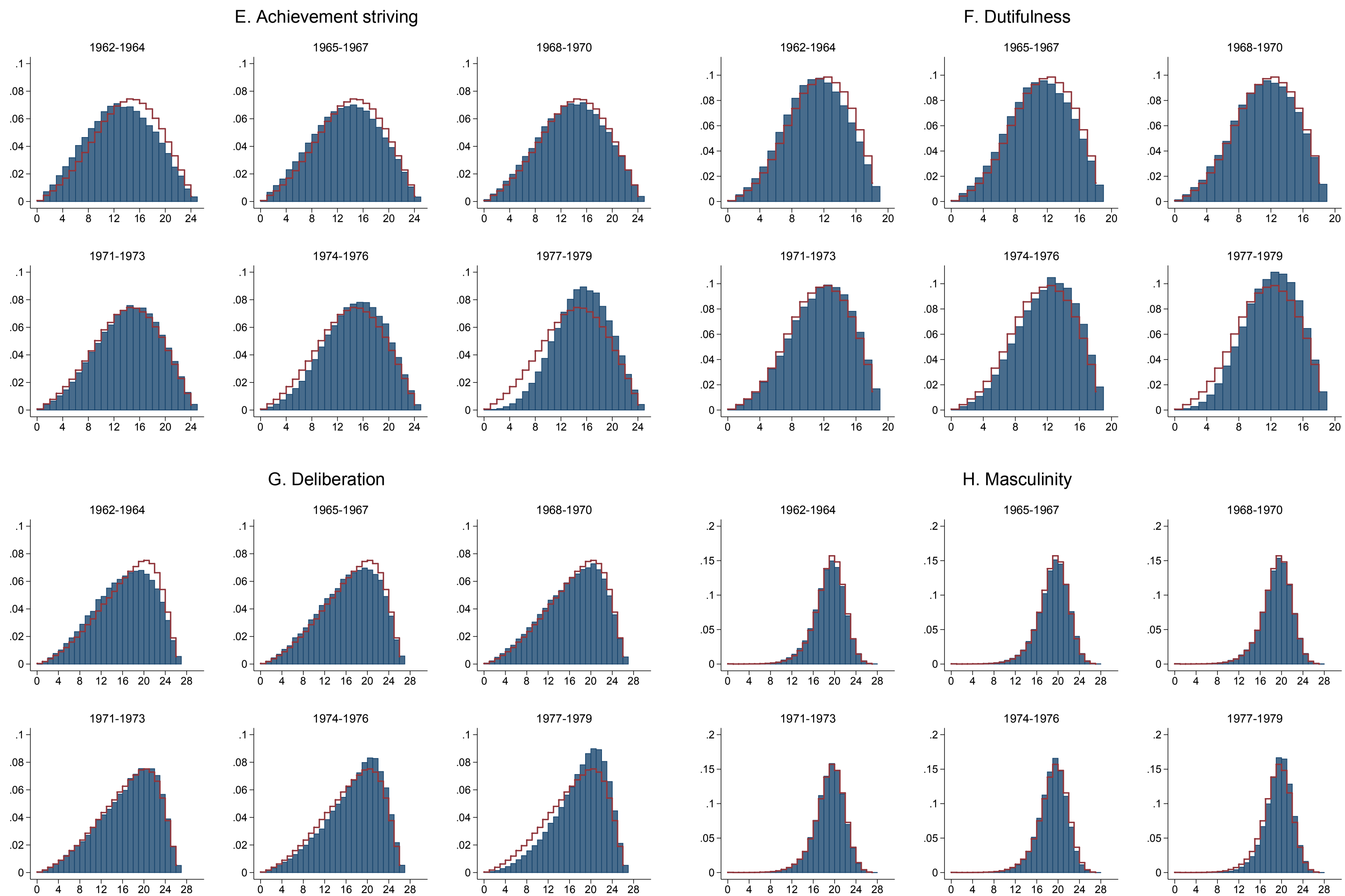

Figure S4: (2) Distributions of the personality test raw scores by birth cohort. The red curves depict the distributions for the 1962–1976 cohorts.

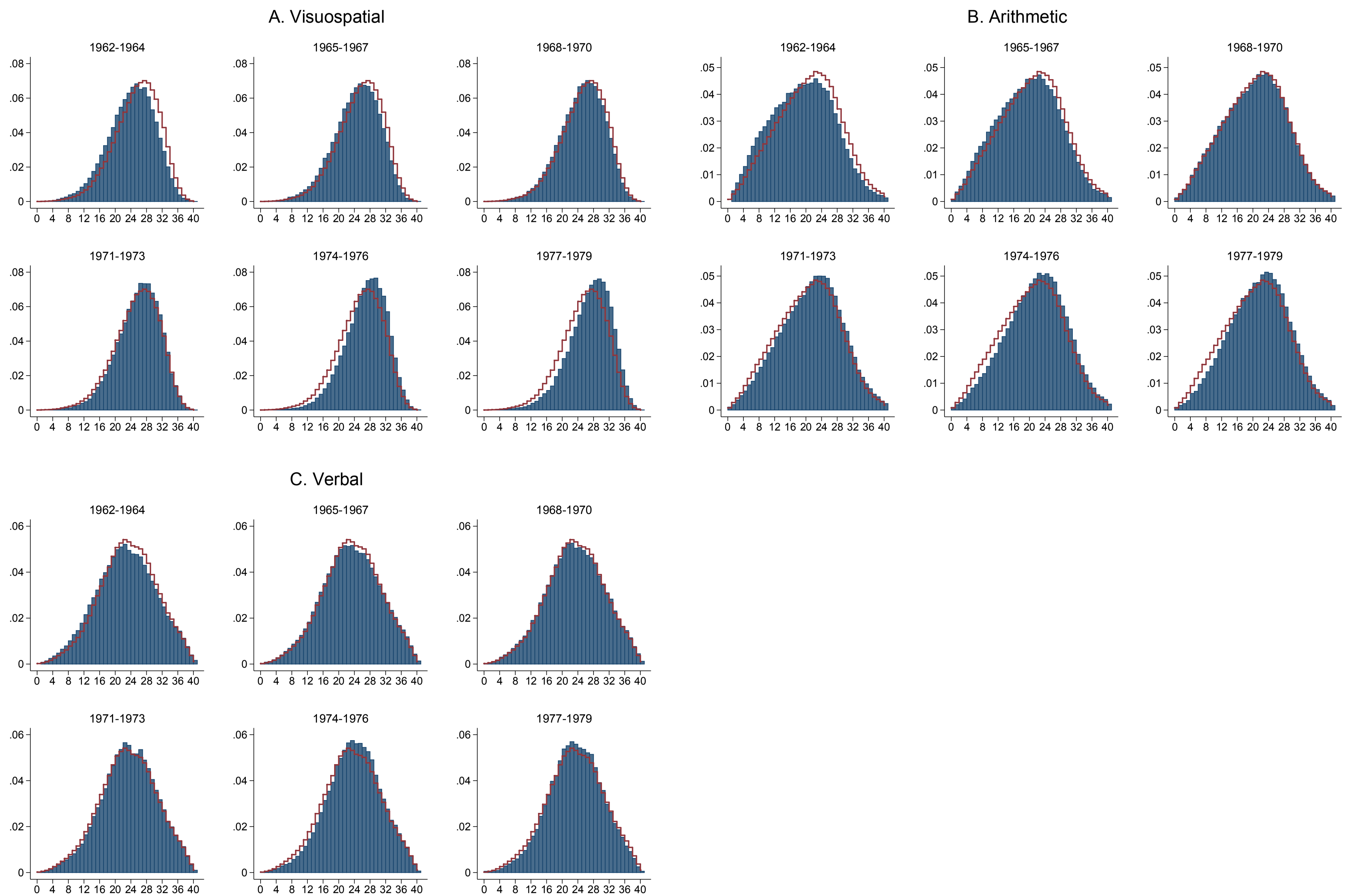

Figure S5: Distributions of the raw cognitive ability test scores by birth cohort. The red curves depict the distributions for the 1962–1976 cohorts.

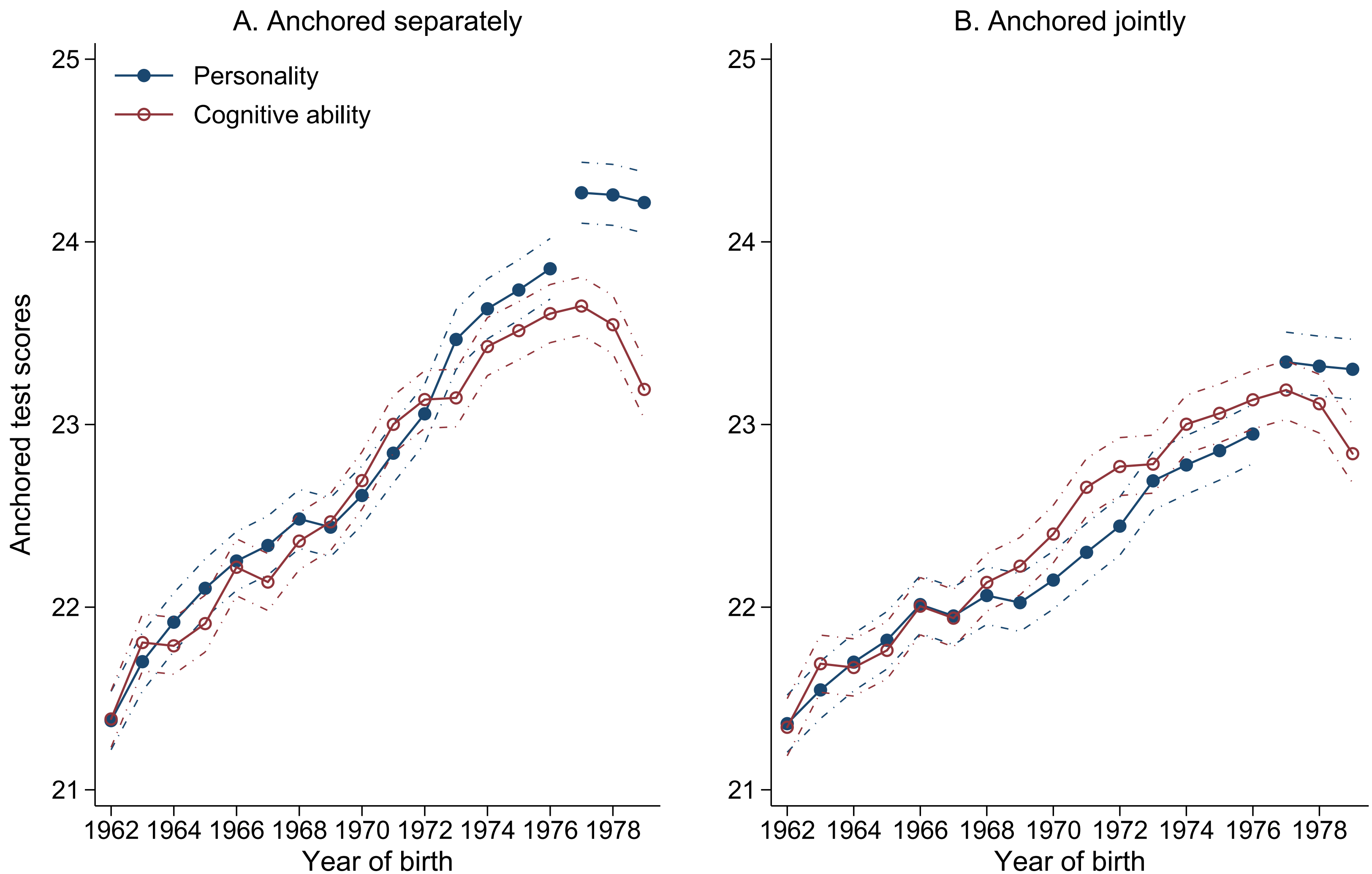

Figure S6: Anchored test scores based on (A) separate regressions and (B) joint regressions of the test scores. See section *Anchored test scores* for details. Panel A reproduces Fig 3 of the main text for reference.

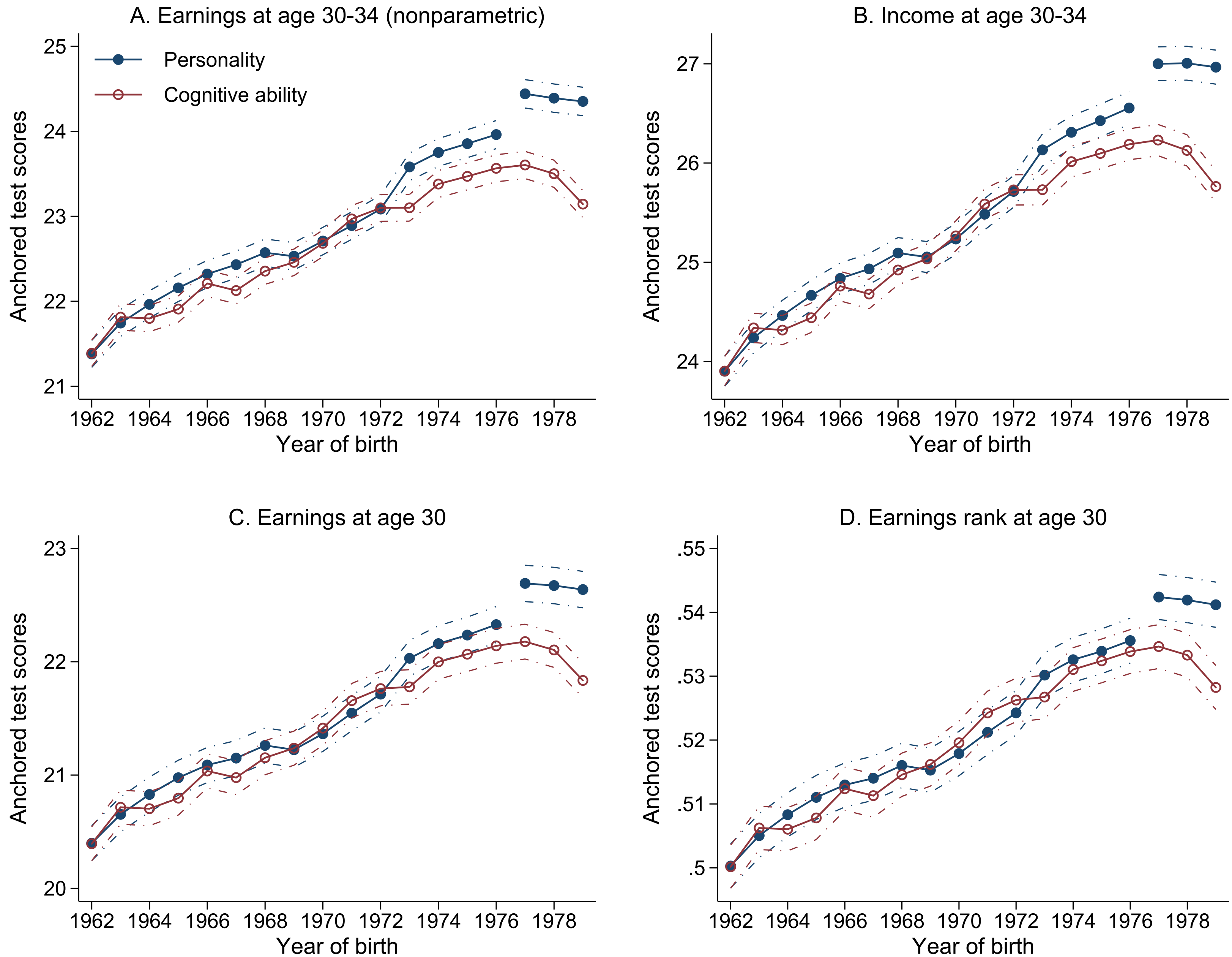

Figure S7: Anchored test scores using alternative anchoring variables. See section *Alternative anchoring variables.*

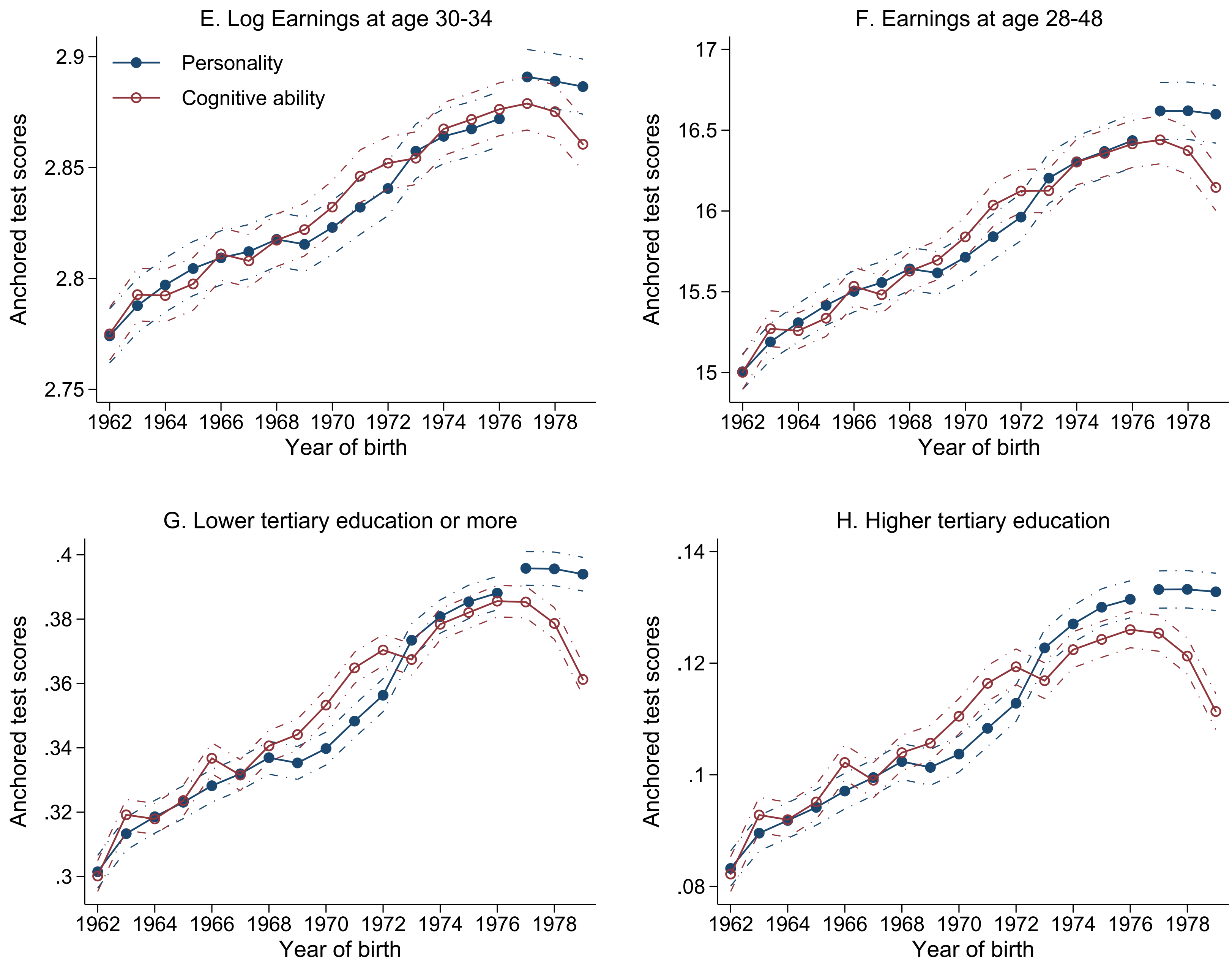

Figure S7: (cont') Anchored test scores using alternative anchoring variables. See section *Alternative anchoring variables.*

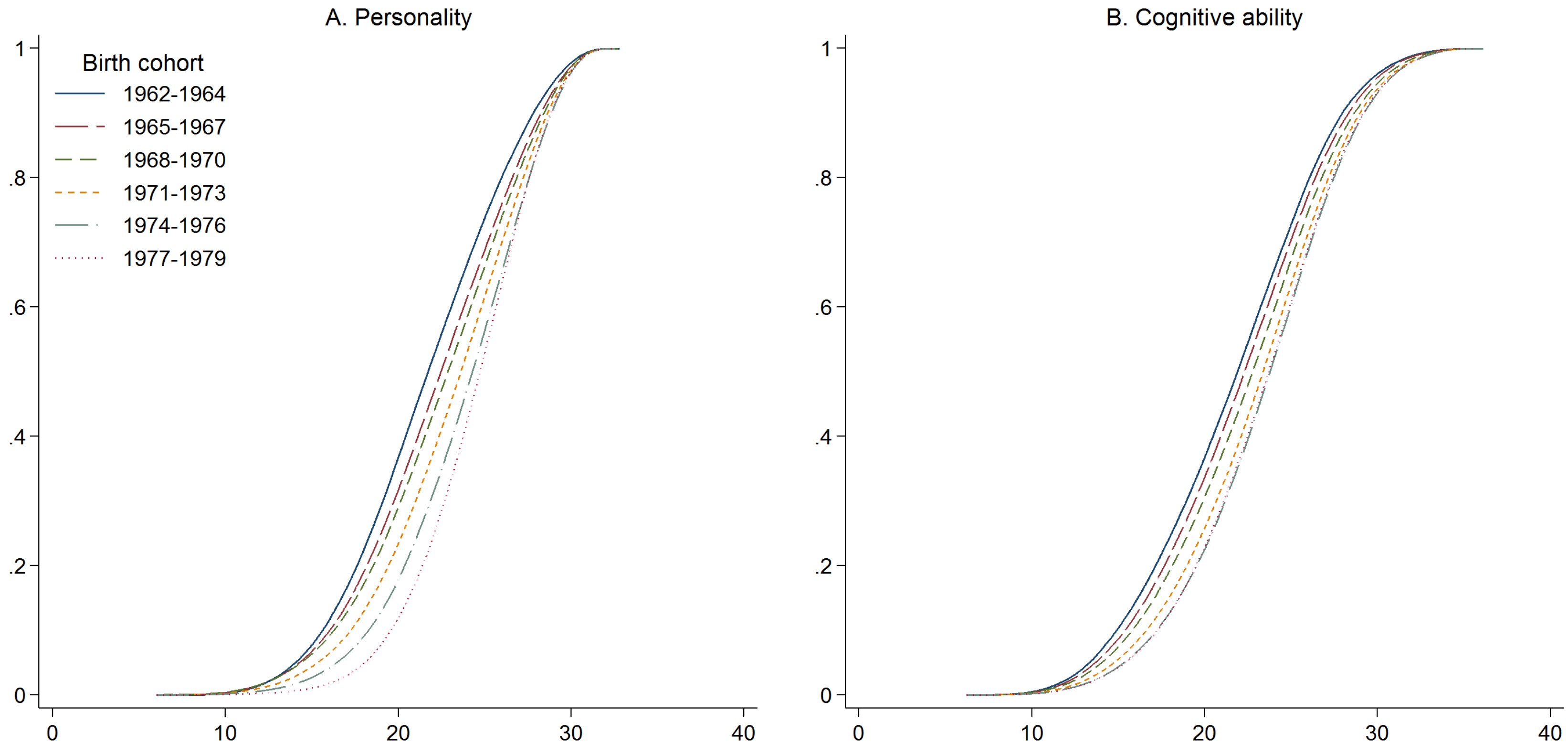

Figure S8: CDFs of anchored test scores. The cohort means of these scores are reported in Fig 3 of the main text.

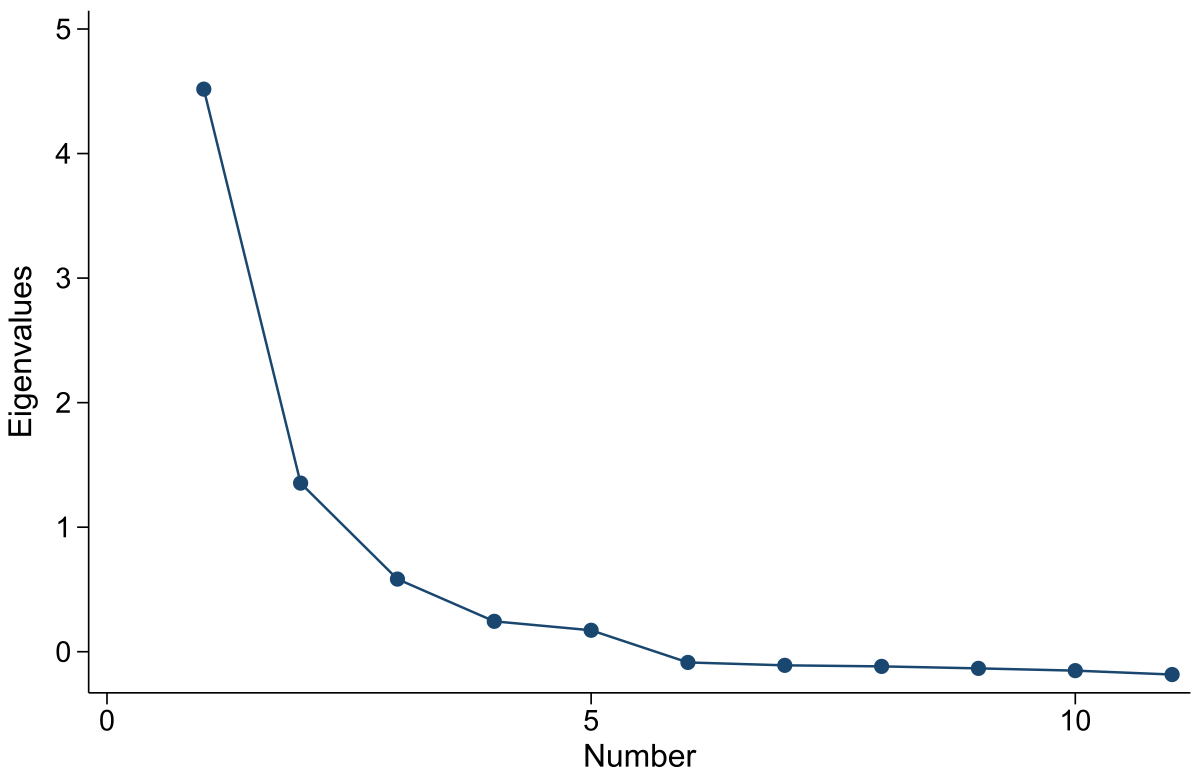

Figure S9: Eigenvalue plot of results from exploratory factor analysis of the test score data. See section *Exploratory factor analysis* for details.

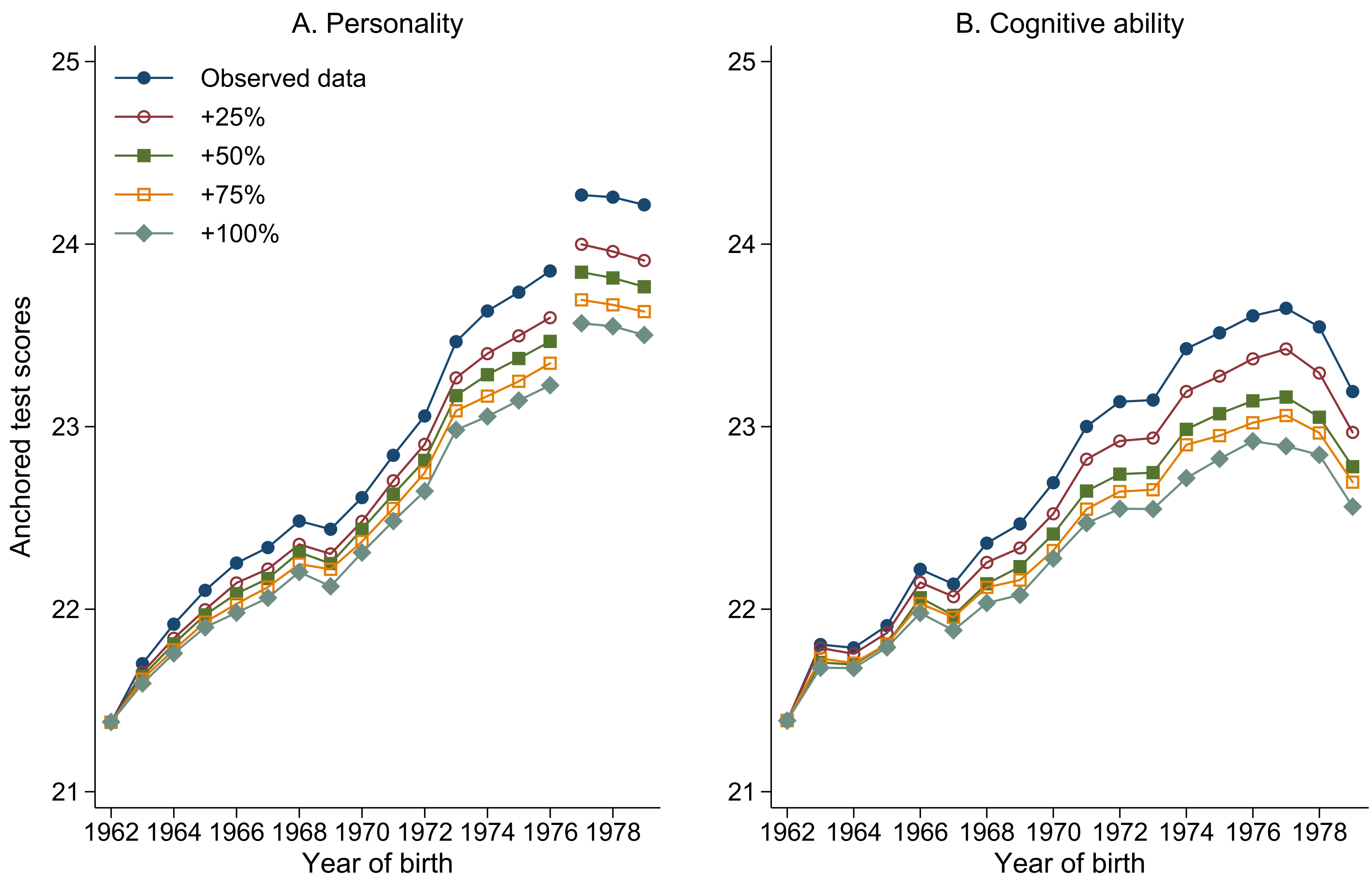

Figure S10: Trends in anchored test scores after adding measurement error to the observed test scores. The measurement error is i.i.d, normally distributed, and measured as a percentage of the variance of the observed test scores. See section *Measurement error* for details.

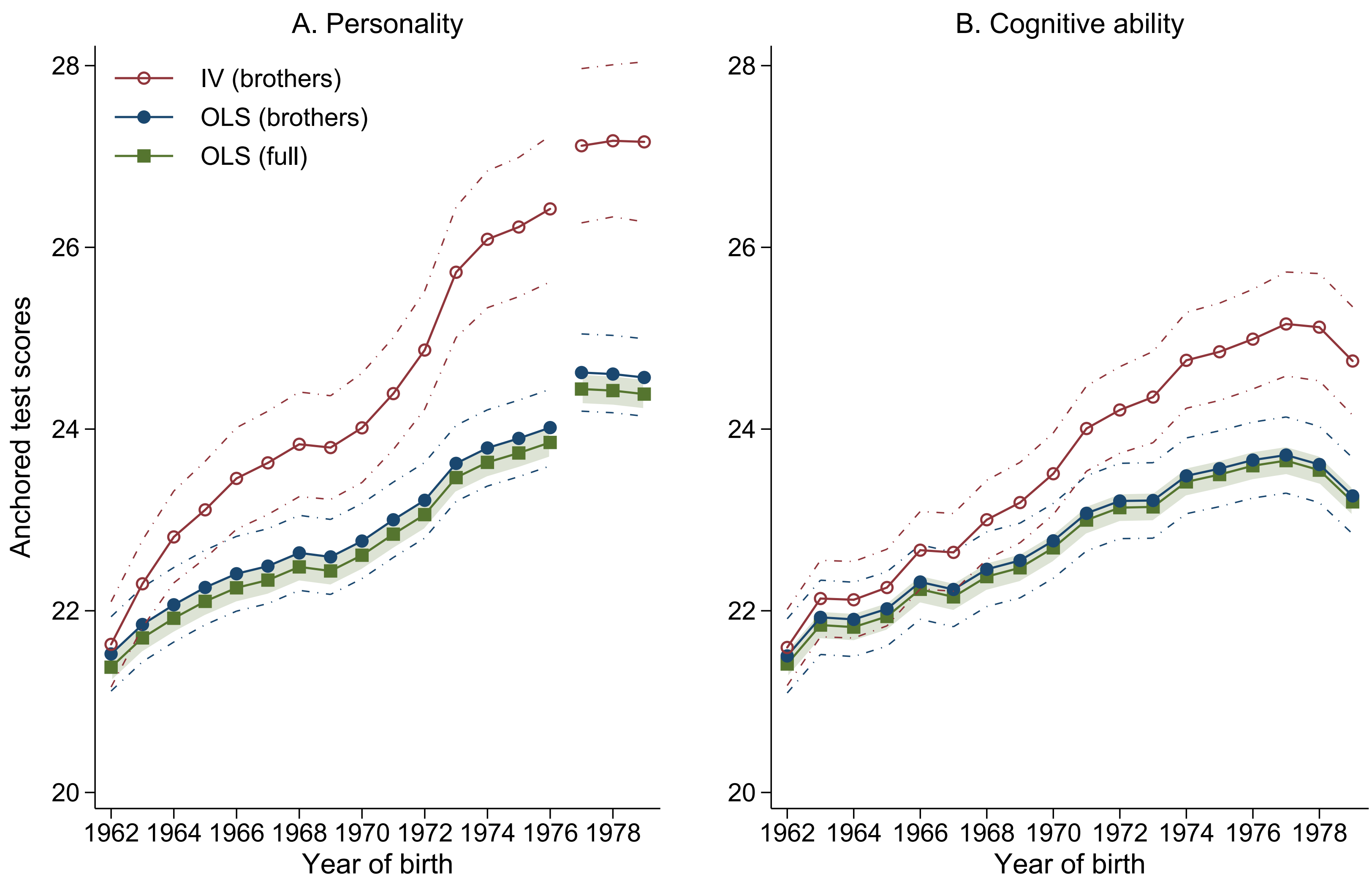

Figure S11: Trends in achored test scores based on IV estimates (brothers). See section *Instrumental variables estimates* for details.

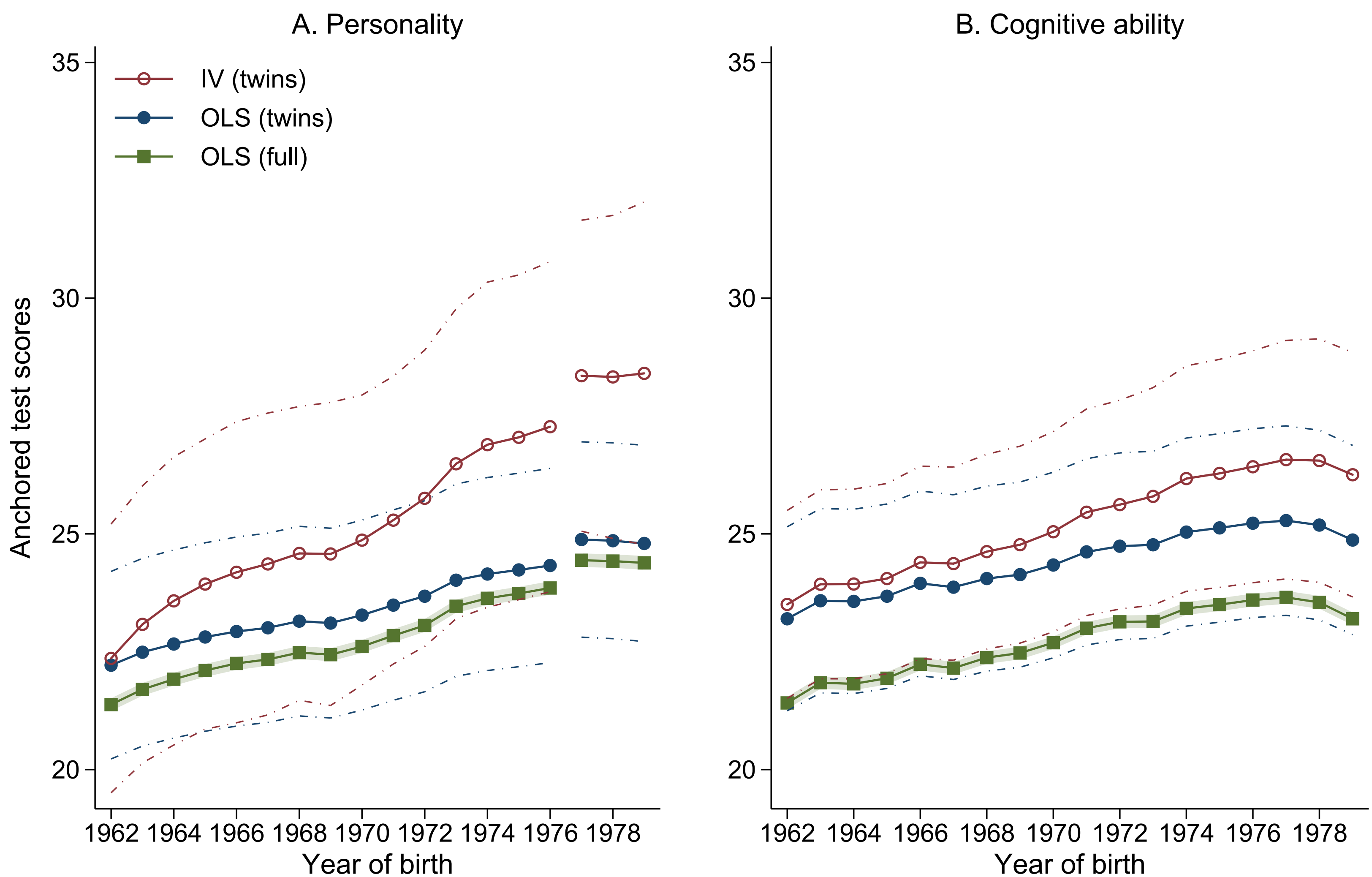

Figure S12: Trends in achored test scores based on IV estimates (twins). See section *Instrumental variables estimates* for details.

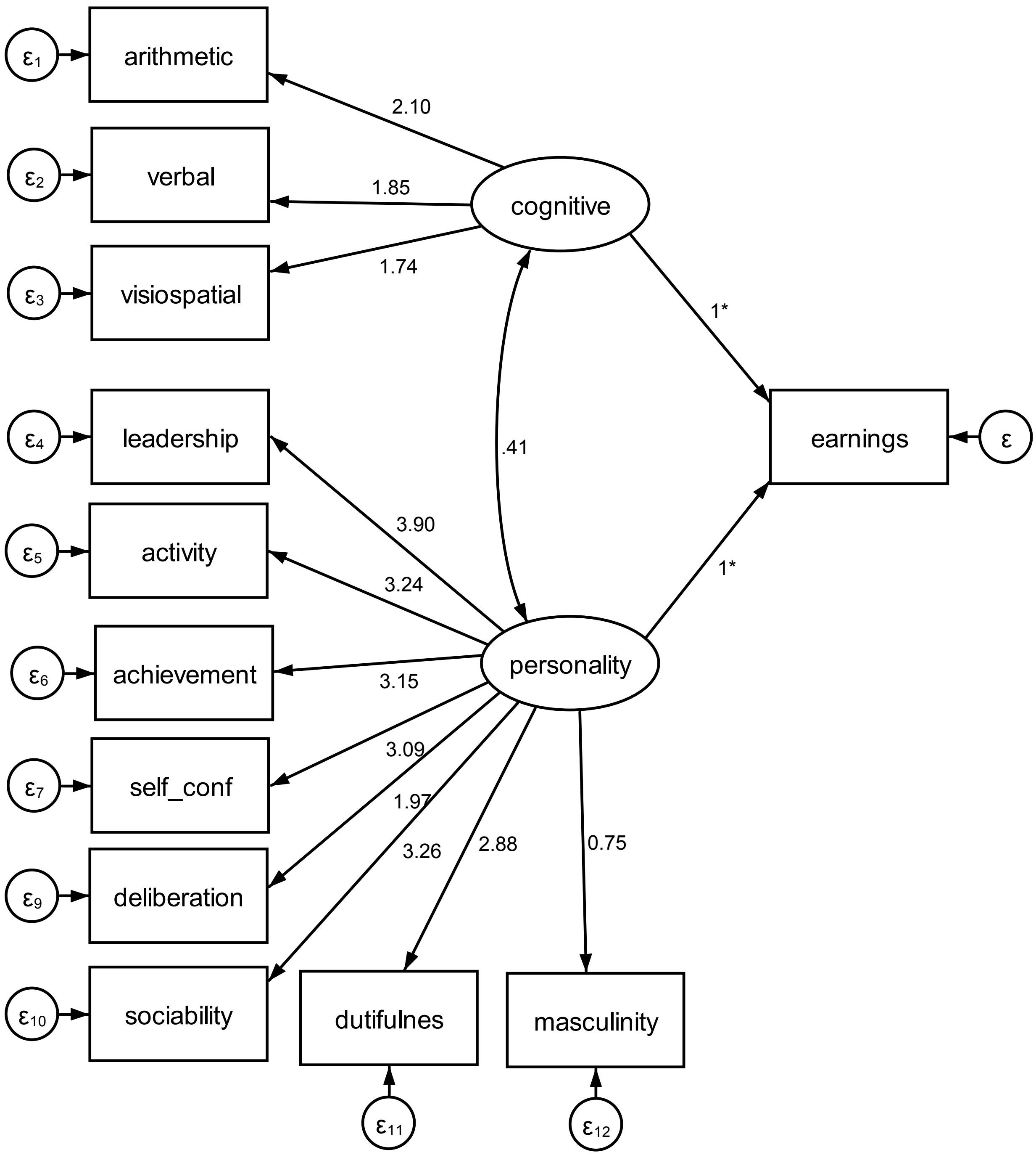

Figure S13: Path diagram of factor structure; see section *Structural equations model* for details.

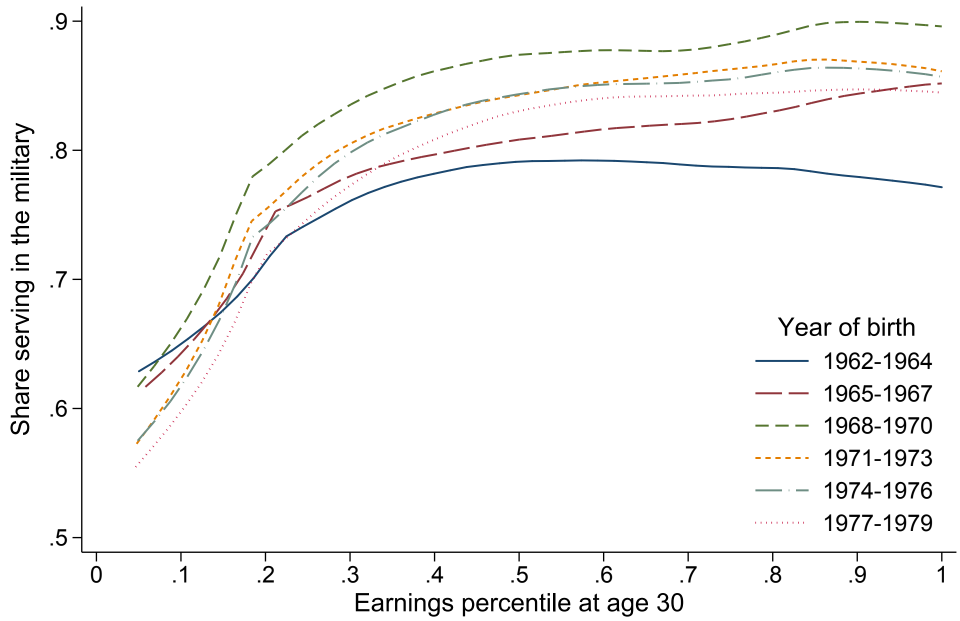

Figure S14: Selection of men to the FDF test data by later-life income. See section *Selectivity in test score data* for details.

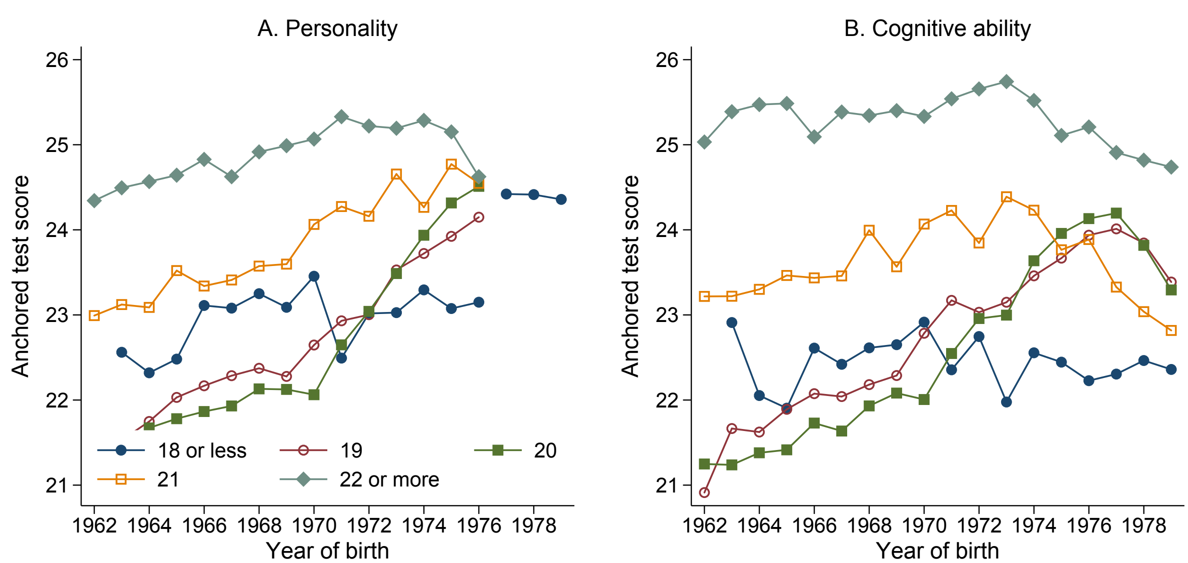

Figure S15: Trends in anchored test scores by age at test.

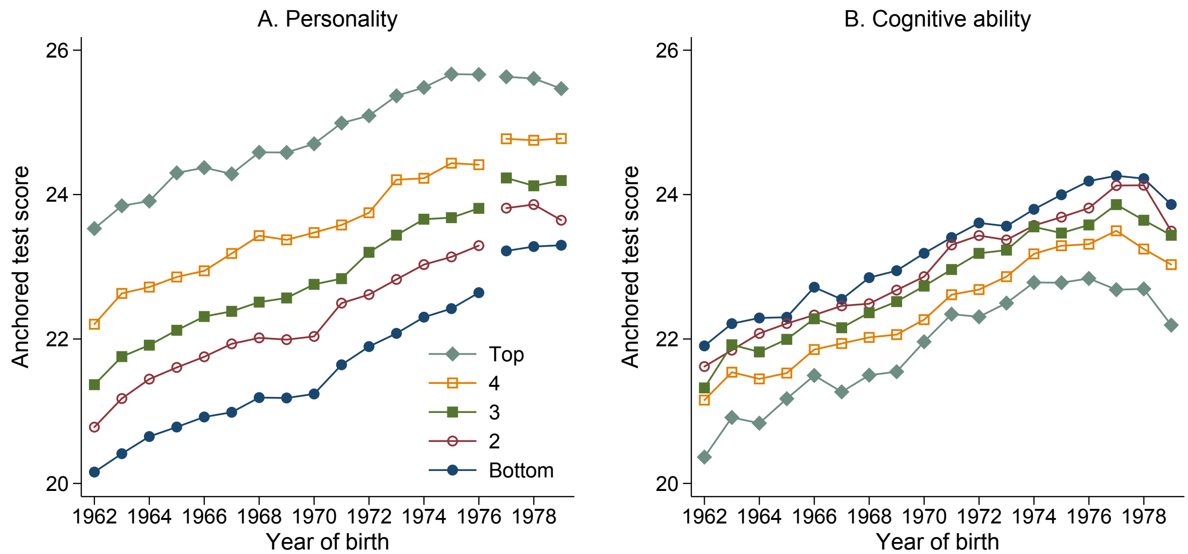

Figure S16: Trends in anchored test scores by quintiles of the Lie-score

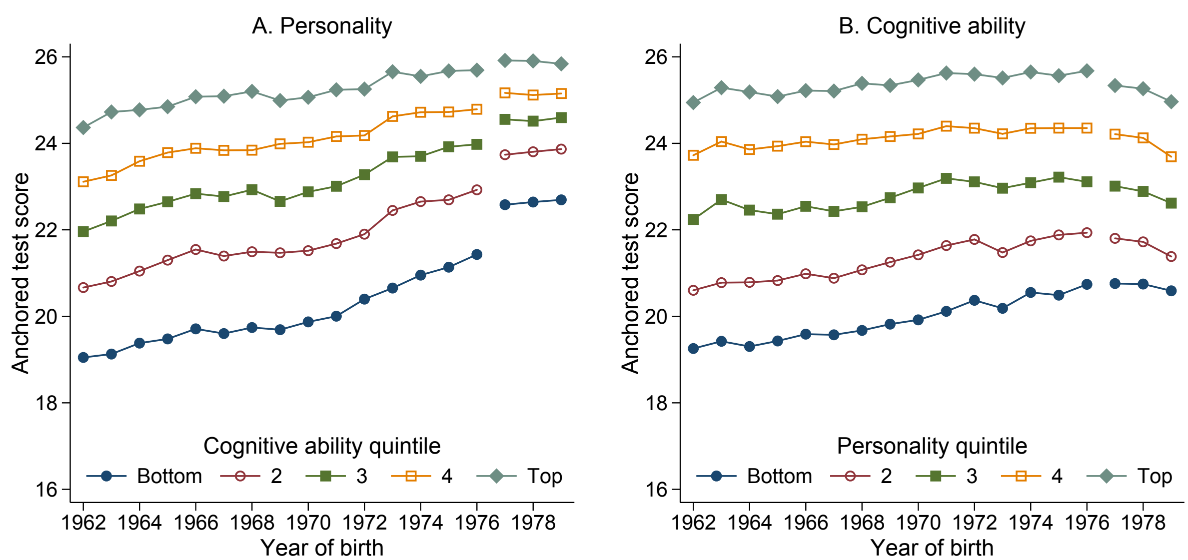

Figure S17: Trends in (A) anchored personality test scores by quintiles of anchored cognitive ability test scores and (B) anchored cognitive ability test scores by quintiles of anchored personality test scores .

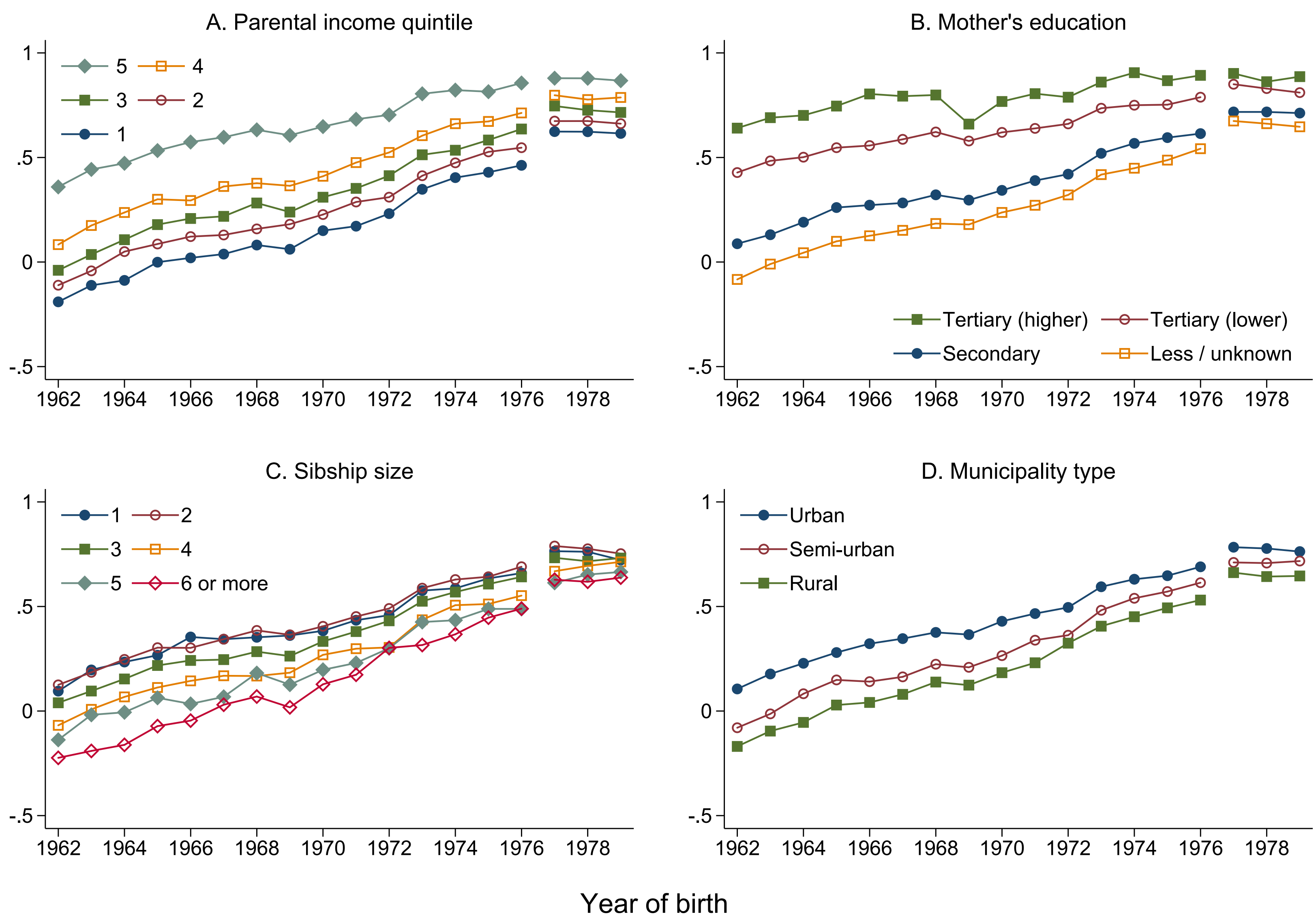

Figure S18: (1) Evolution of *Self-confidence* across birth cohorts by (**A**) parental income quintile, (**B**) mother's education, (**C**) sibship size, and (**D**) urbanization of birth place.

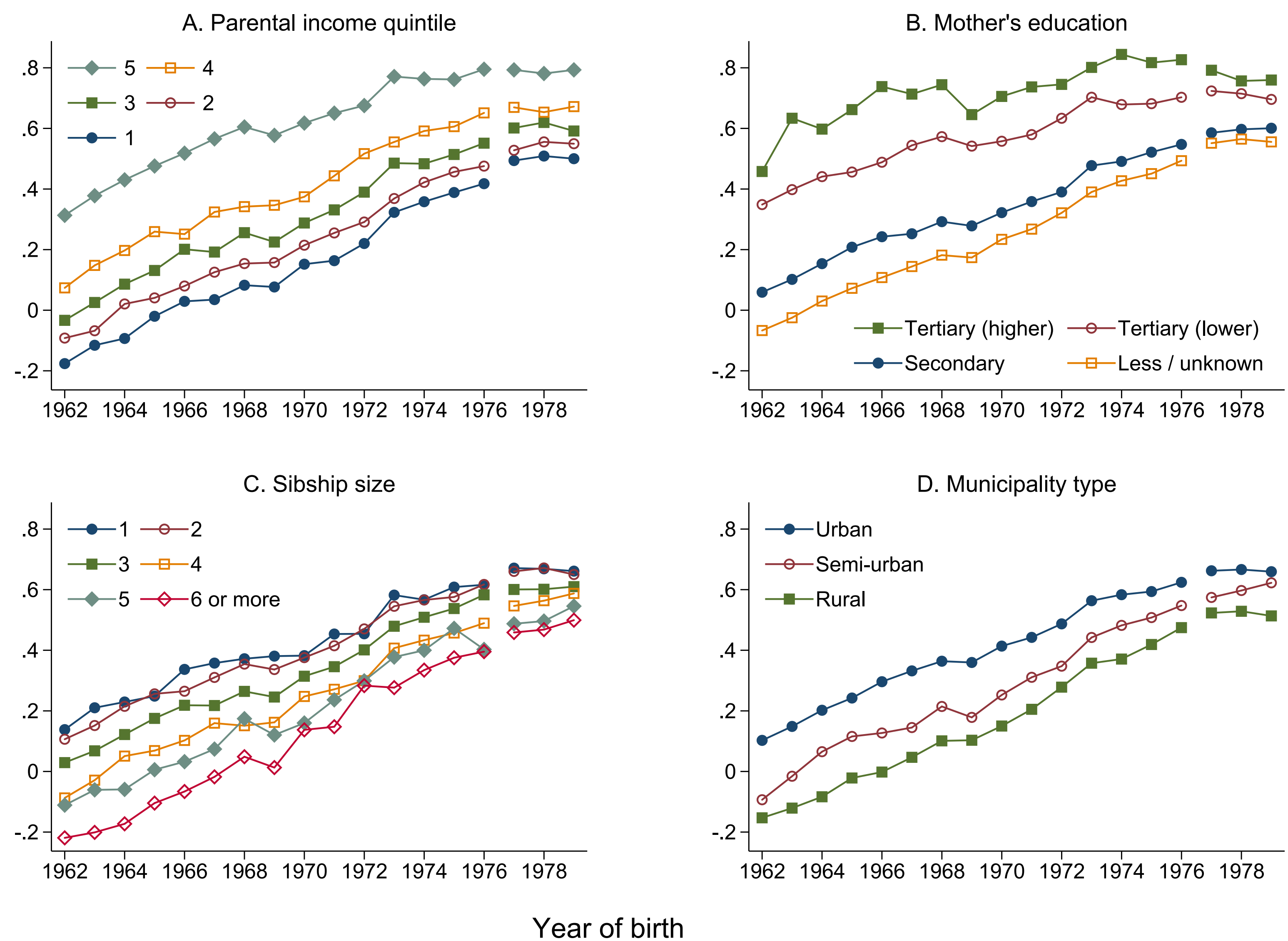

Figure S18: (2) Evolution of *Sociability* across birth cohorts by (**A**) parental income quintile, (**B**) mother's education, (**C**) sibship size, and (**D**) urbanization of birth place.

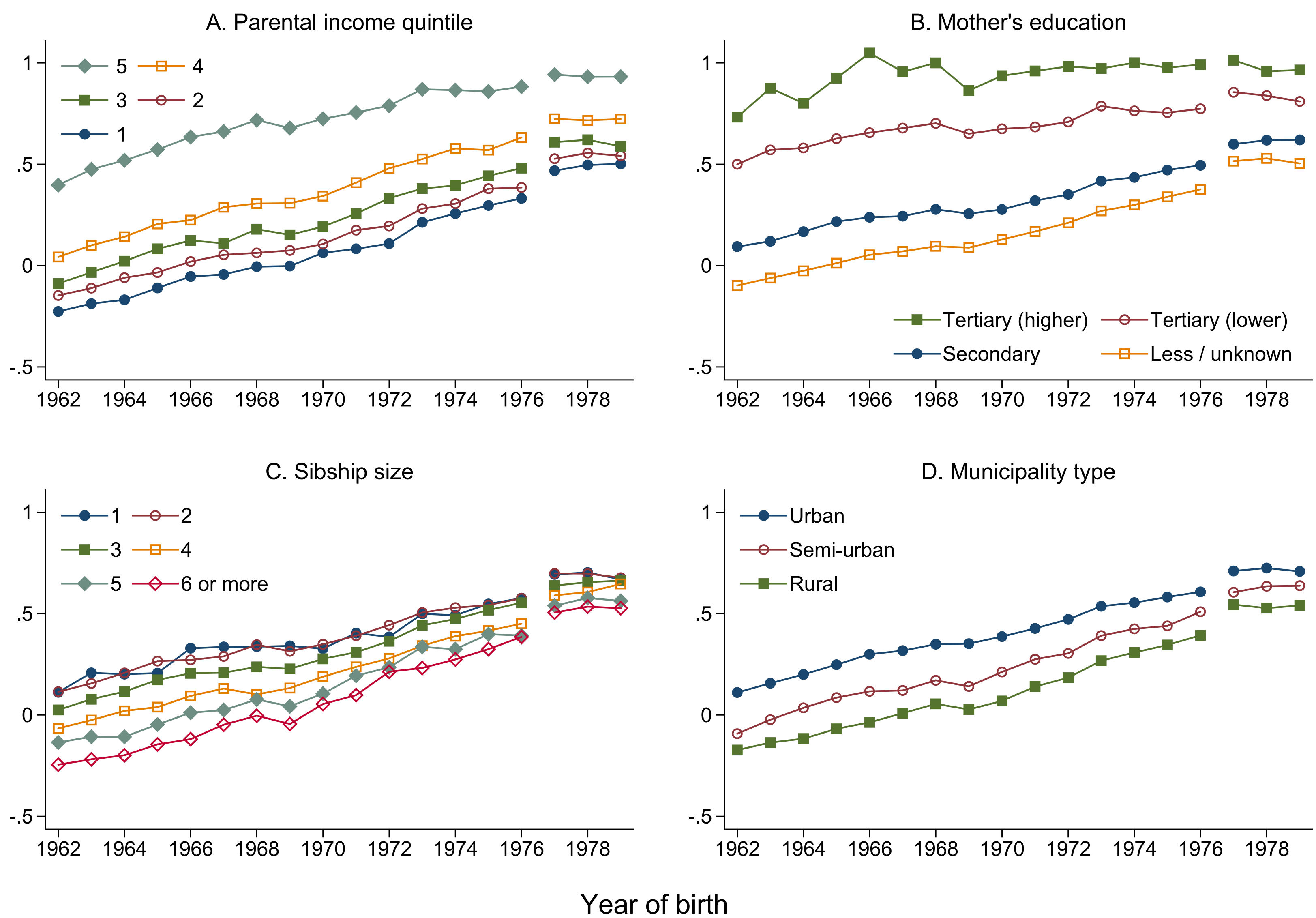

Figure S18: (3) Evolution of *Leaderhsip motivation* across birth cohorts by (**A**) parental income quintile, (**B**) mother's education, (**C**) sibship size, and (**D**) urbanization of birth place.

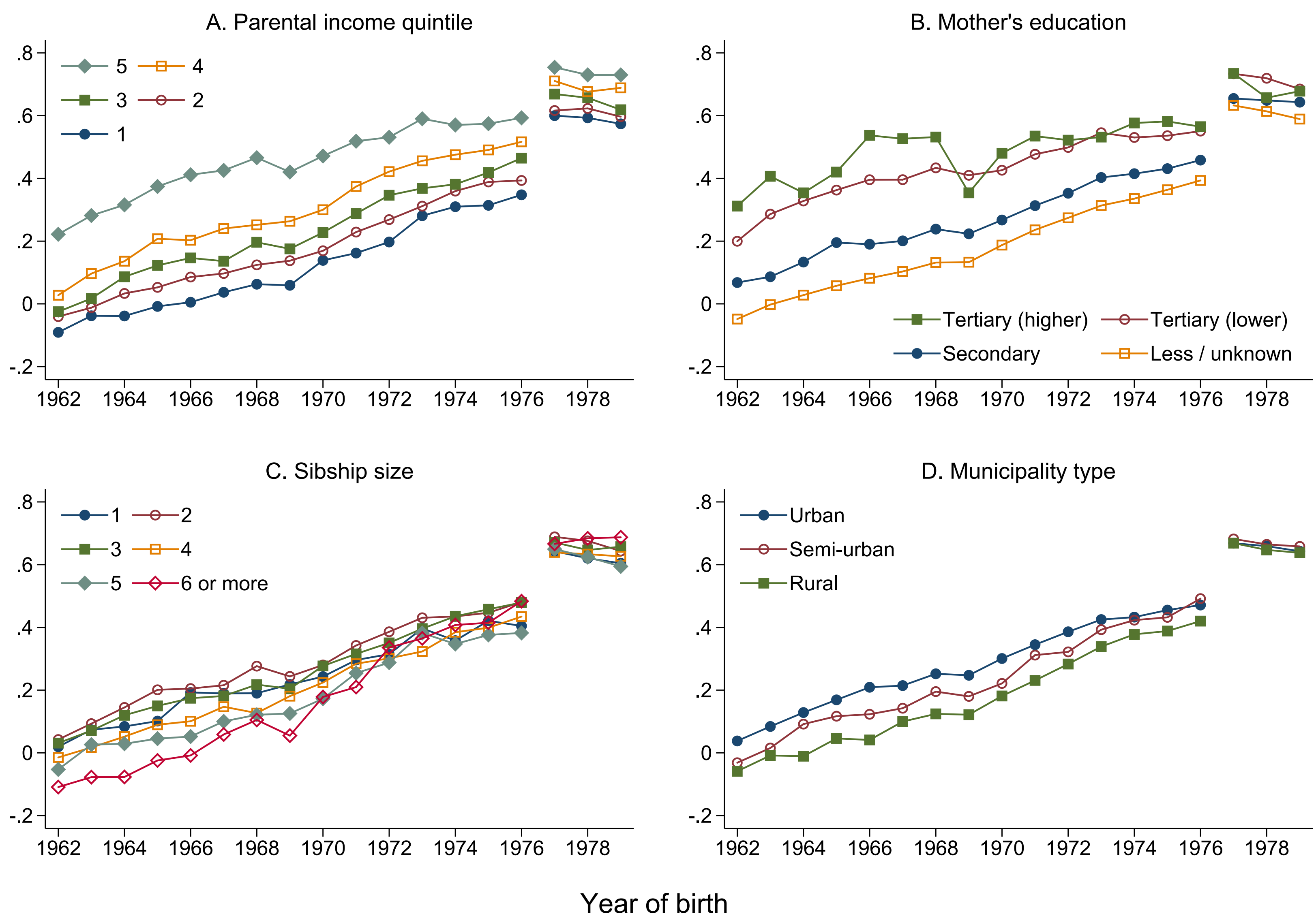

Figure S18: (4) Evolution of *Activity-energy* across birth cohorts by (**A**) parental income quintile, (**B**) mother's education, (**C**) sibship size, and (**D**) urbanization of birth place.

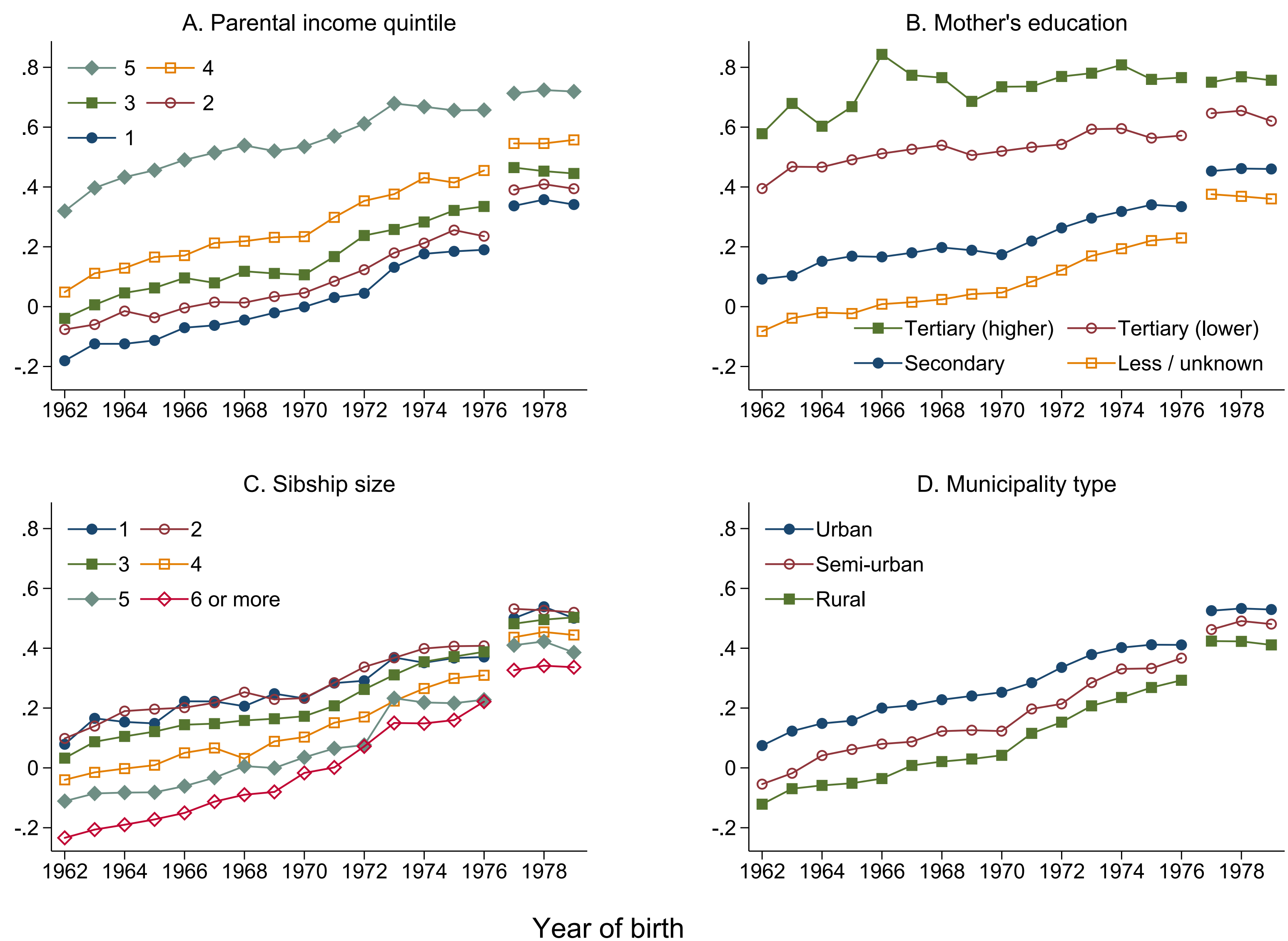

Figure S18: (5) Evolution of *Achievement striving* across birth cohorts by (**A**) parental income quintile, (**B**) mother's education, (**C**) sibship size, and (**D**) urbanization of birth place.

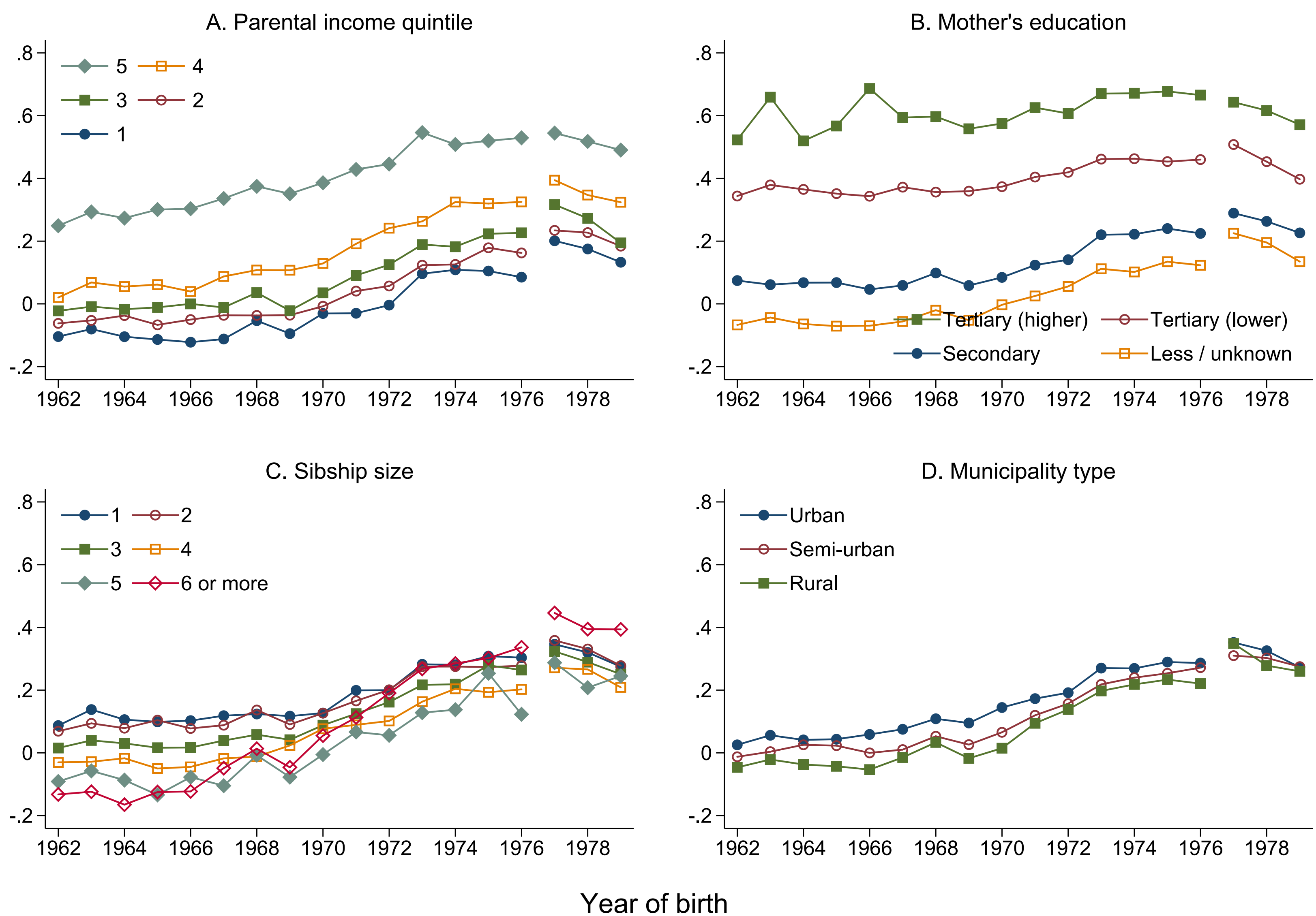

Figure S18: (6) Evolution of *Dutifulness* across birth cohorts by (**A**) parental income quintile, (**B**) mother's education, (**C**) sibship size and (**D**) urbanization of birth place.

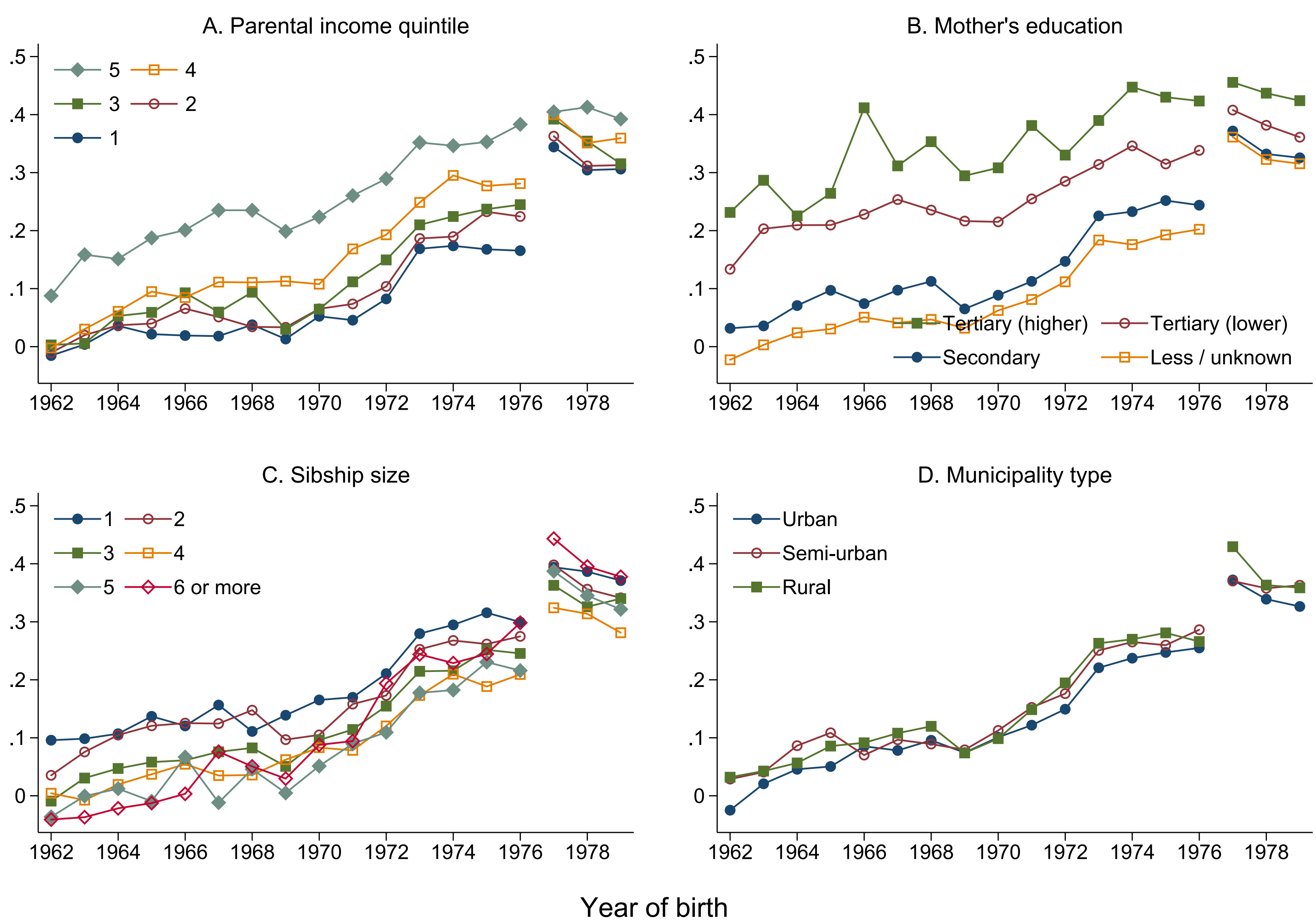

Figure S18: (7) Evolution of *Deliberation* across birth cohorts by (**A**) parental income quintile, (**B**) mother's education, (**C**) sibship size, and (**D**) urbanization of birth place.

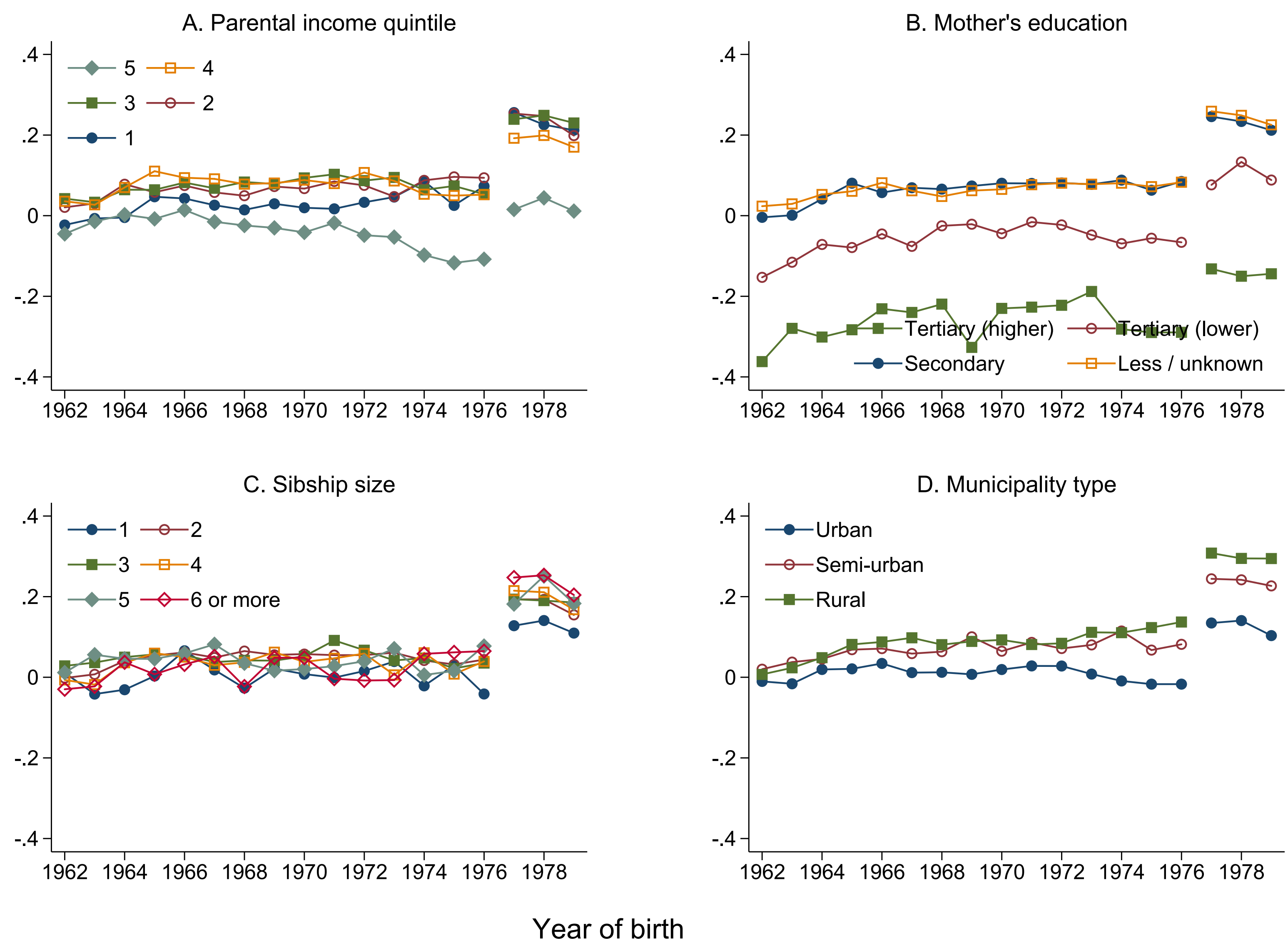

Figure S18: (8) Evolution of *Masculinity* across birth cohorts by (**A**) parental income quintile, (**B**) mother's education, (**C**) sibship size, and (**D**) urbanization of birth place.

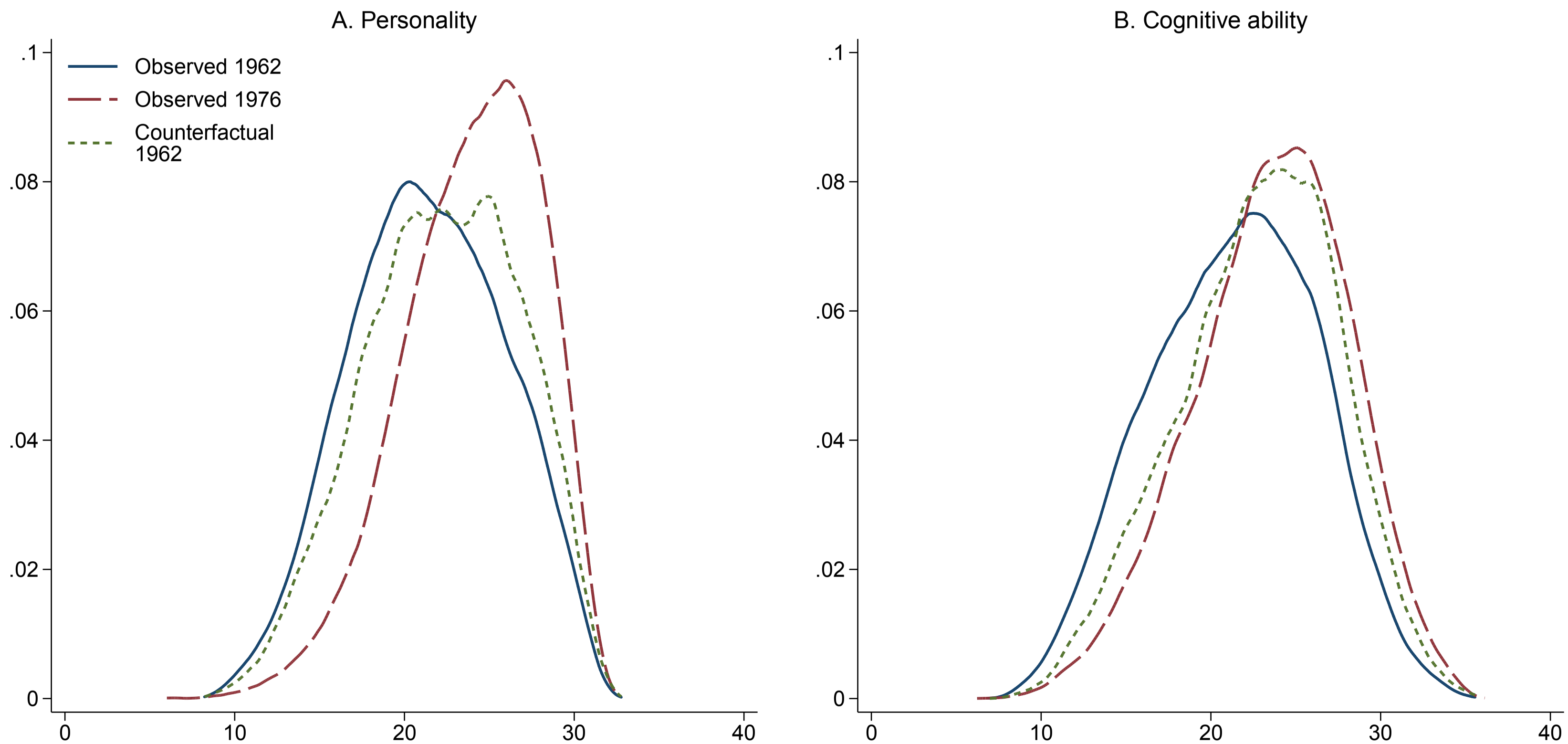

Figure S19: DFL decomposition of anchored test scores. Each figure shows the test score distribution of the 1962 birth cohort (solid line), the observed test score distribution of the 1976 birth cohort (dashed line), and the counterfactual distribution where we reweight the 1962 test score distribution to correspond to the 1976 distribution of background characteristics (dotted line). See section *Trends in background variables* for details.