I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17687

# Revisiting the Dunning-Kruger Effect: Composite Measures and Heterogeneity by Gender

Anna Adamecz
Radina Ilieva
Nikki Shure

FEBRUARY 2025

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17687

# Revisiting the Dunning-Kruger Effect: Composite Measures and Heterogeneity by Gender

**Anna Adamecz**
*University College London, Institute of Economics, Centre for Economic and Regional Studies (KRTK KTI) and IZA*

**Radina Ilieva**
*Informa Connect*

**Nikki Shure**
*University College London and IZA*

FEBRUARY 2025



ISSN: 2365-9793

**IZA – Institute of Labor Economics**

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

# ABSTRACT

## Revisiting the Dunning-Kruger Effect: Composite Measures and Heterogeneity by Gender*

The Dunning-Kruger effect (DKE) states that people with lower levels of the ability tend to self-assess their ability less accurately than people with relatively higher levels of the ability. Thus, the correlation between one's objective cognitive abilities and self-assessed abilities is higher at higher levels of objective cognitive abilities. There has been much debate as to whether this effect actually exists or is a statistical artefact. This paper replicates and extends Gignac and Zajenkowski (2020) and Dunkel, Nedelec, and van der Linden (2023) to test whether the DKE exists using several measures of ability and nationally representative data from a British birth cohort study. To do this, we construct a measure of objective cognitive abilities using 18 tests conducted at ages 5, 10, and 16, and a measure of subjective self-assessed abilities using estimates of school performance and being clever at ages 10 and 16. We replicate their models and show that the DKE exists in our secondary data. Importantly, we are the first to look at whether this relationship is heterogeneous by gender and find that while the self-assessment bias is gender specific, the DKE is not. The DKE comes from men relatively overestimating and women relatively underestimating their abilities.

**JEL Classification:** J16, J24, D90

**Keywords:** Dunning-Kruger effect, overconfidence, underconfidence, gender differences

**Corresponding author:**
Nikki Shure
UCL Social Research Institute
University College London
55-59 Gordon Square
London WC1H 0NU
Great Britain
E-mail: nikki.shure@ucl.ac.uk

* This work was supported by the Economic and Social Research Council [grant number ES/T013850/1]. This project received ethics approval from the UCL IOE Research Ethics Committee [REC 1374]. The data used in this paper is available via the UK Data Service.

## 1. Introduction

The Dunning-Kruger effect (DKE) refers to the observation that people who are less competent in certain intellectual and social domains are also worse at estimating their performance than those who are more competent (Kruger and Dunning 1999). Furthermore, Kruger and Dunning (1999) argue that this correlation between an individual's self-assessed and objective abilities can be accounted for by the fact that the skills and knowledge necessary for one to perform well on the cognitive task are often the same as those they need to be able to accurately evaluate one's performance in that domain, the *meta-cognitive task*. The result of this phenomenon is a higher degree of overconfidence among the low performers, as they overestimate their performance compared to their actual, poor performance. Studies focussed on a range of domains have found that high performers are better at the meta-cognitive task of assessing their own performance, but that in general most people overestimate their abilities (Dunning 2011). Dunning (2011) also highlights that the DKE is pervasive across a range of contexts from chess players to gun owners to medical residents.

Since their seminal paper, many attempts have been made to test – and criticize – the DKE. This has ranged from a critique of the representativeness of their sample (Krajc and Ortmann, 2008), that the effect is driven by regression to the mean (Krueger and Mueller 2002), and that the effect is the result of noise plus bias (McIntosh and Della Sala 2022). More recent work has proposed that as a result of these critiques, the classic test for the DKE is confounded and instead researchers should use alternative methods, for example, introducing a quadratic term into their linear regressions (Gignac and Zajenkowski 2020). Another frequently used method to test the DKE is the Glejser test for heteroscedastic residuals; however, its validity in this context has also been questioned (Gignac 2022).

Krajc and Ortmann (2008) critique the lack representativeness among the participants in the studies conducted by Dunning, Kruger, and their collaborators, as most of them are Cornell University undergraduate students. They conclude that "little can be said about (the lack of) metacognitive ability if one does not control for the distribution of real abilities" (p. 736). Additionally, they propose that the inability of low-performing students to accurately estimate their performance may be explained by a greater difficulty to extract information from the feedback provided to them, identified as a "signal-extraction problem", rather than by a lack of metacognitive skills, as suggested by Dunning and Kruger.

Schlösser et al. (2013) assess the Krajc-Ortmann theoretical model and "signal-extraction" explanation across three studies. Their findings confirm the greater overestimation of self-

performance by poor performers, anticipated by the Dunning-Kruger framework. In contrast, the Krajc-Ortmann model fails to account for this overestimation, although it does well in accounting for the underestimation of abilities among high performers.

Krueger and Mueller (2002) offer an additional alternative explanation of the asymmetry in the ability to evaluate own performance of top and bottom performers. They suggest that the errors in an individual's estimation of self-performance could be predicted by statistical regression towards the mean, taking also into account the "Better-Than-Average" effect. That is, they argue that there cannot be a perfect correlation between two variables, such as objective and estimated performance, because of the errors that inevitably occur when measuring those variables in the first place.

Further empirical evidence for the argument that the DKE is simply a statistical artefact is the study conducted by Gignac and Zajenkowski (2020). Similar to previous studies, participants were asked to self-assess their intelligence on a scale from 1 to 25, and then to complete an intelligence test. Gignac and Zajenkowski introduce a quadratic term into the DKE regression and use a Glejser test to test for heteroscedasticity of the errors. The results from these statistical analyses suggest that there is no significant difference in the self-assessed abilities of individuals across the spectrum of objective abilities (Gignac and Zajenkowski 2020).

In a recent paper responding to and building on Gignac and Zajenkowski (2020), Dunkel, Nedelec, and van der Linden (2023) use nationally representative survey data from Ad Health in the US as well as the methods proposed by Gignac and Zajenkowski (2020) and find support for a modest DKE. They note, however, that their measure of self-assessed ability is based on two questions with categorical answers and urge future research to consider a continuous measure of self-assessed ability. They also highlight the limited nature of the objective ability measure used.

Despite the substantial amount of academic research on and debate around the DKE, very few papers try to address all these concerns in a unified framework. One way to address them would be to test the existence of the DKE incorporating new methodological considerations while also using secondary data, specifically birth cohort data. Birth cohort studies are designed to be representative of a generation of the population, so sample selection issues will not be a concern. We contribute to the literature by testing the existence of the DKE, replicating Gignac

and Zajenkowski (2020), to examine whether the DKE exists in nationally representative, secondary data while testing for its existence using traditional and newer methods.

Beyond replicating and confirming the existence of the DKE using secondary data, we provide the first estimates of heterogeneity by gender in its existence using nationally representative data. Surprisingly, there is no empirical examination of its heterogeneity across genders. This is despite a finding in the literature that men are more likely to overclaim knowledge or overestimate their ability than women (Ehrlinger and Dunning 2003; Möbius et al. 2011; Niederle and Vesterlund 2007; Adamecz-Völgyi and Shure 2022). There is also evidence that overconfidence is more beneficial to men, both from an evolutionary (von Hippel and Trivers 2011; Ronay, Maddux, and von Hippel 2022) and an economic standpoint (Adamecz-Völgyi and Shure 2022). Since overconfidence, a type of miscalibration (Moore and Healy 2008), is closely associated with the DKE, this makes exploring the prevalence of the DKE by gender interesting. This builds on a recent paper by Dunkel, Nedelec, and van der Linden (2023), which uses nationally representative survey data from the US as well as the methods proposed by Gignac and Zajenkowski (2020) to test for the DKE, but do not explore heterogeneity by gender.

We make two contributions to the literature. First, as opposed to a small sample of university students used in most papers, we test the DKE using a nationally representative sample of a British birth cohort born in 1970 with multiple measures of self-assessed and objective ability. This again builds on Dunkel, Nedelec, and van der Linden (2023), who also use secondary data to test for the DKE, but we extend their work by constructing more robust measures of objective and self-assessed ability. Our measures are not one-off intelligence tests or self-assessments of intelligence, but composite measures of several tests taken at ages five, 10 and 16. This should assuage concerns about measurement errors (Bollen 1989). In addition to the classic tests proposed by Kruger and Dunning (1999), we also use the methods proposed by Gignac and Zajenkowski (2020) to test for the existence of the DKE.

Second, we are the first to investigate whether the DKE is heterogeneous by gender. This is important given previous findings in the literature that show gender differences in overclaiming and overestimation and an association of overconfidence with men in popular culture. This extends the work of Dunkel, Nedelec, and van der Linden (2023) and allows us to answer a new and important research question.

We hypothesize that the DKE will be present in our nationally representative sample, but not differ between men and women. Men and women will be equally good or bad at assessing their own ability; however, we also hypothesize that women will underestimate their ability and men will overestimate it. These two biases taken together will contribute to the DKE. This highlights the importance of distinguishing between the DKE and self-assessment biases more broadly.

The rest of this paper is structured as follows. In section 2, we present the data and some descriptive statistics. In section 3, we discuss the methodology and in section 4, we present the results and discuss additional robustness checks. Finally, in section 5, we conclude.

## 2. Data and descriptive statistics

We use data from the British Cohort Study 1970 (BCS70, CLS n.d.) to test the DKE using nationally representative secondary data and take advantage of the multiple timepoints at which the young people were asked to assess their ability and take cognitive tests.[§] The BCS70 is a birth cohort study that follows the lives of approximately 17,000 individuals born in the UK in a specific week in 1970. As such, it is representative of a generation of individuals in the UK, not just those individuals who attend university. The BCS70 collects rich data on family background, childhood and adolescent cognitive and non-cognitive skills, preferences, and labor market and other life outcomes.

Our main analytical sample includes 4,429 individuals, who participated in the age five, 10, and 16 waves of the survey and have non-missing data on at least five objective cognitive measures and at least three measures of self-assessed ability.[**] As out of the 17,000 individuals we only include 4,429 in our analytical sample, we show that sample selection (attrition and non-response) does not bias our results, which we discuss in further detail with our other robustness checks.

We measure objective cognitive abilities via 18 tests taken at ages five, 10 and 16 (see the detailed explanation of measures in Table A1).[††] This includes cognitive ability measures

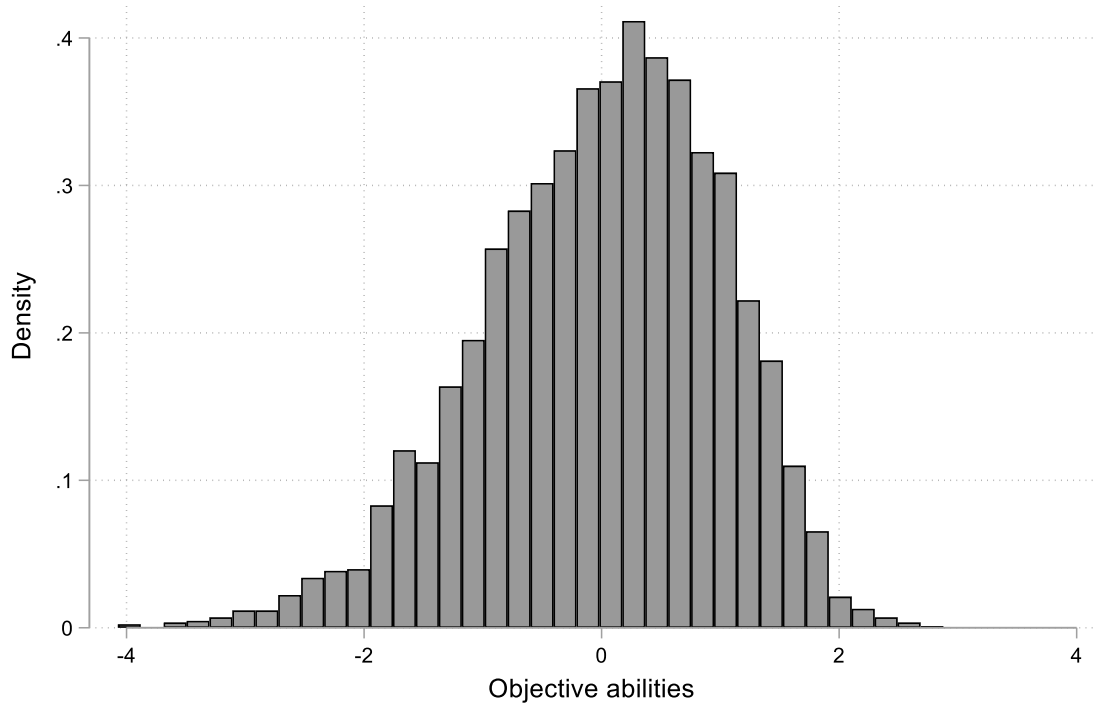
[§] We use safeguarded data (accessed through the UK Data Service) from the birth sweep (SN: 2666), the age five sweep (SN: 2699), the age 10 sweep (SN: 3732), and three data collections of the age 16 sweep (SN: 3535, 6095 and 8288).

[**] We also provide robustness checks using the complete case sample, i.e. those who provided data on all 18 objective and all seven self-assessed ability questions (N=1,337) and find very similar results.

[††]As age five is an early age that might make the assessment of objective cognitive skills challenging, we also provide a robustness check where we exclude the age five measures and use only the remaining 13 measures of objective cognitive ability from ages 10 and 16. This strategy leads to very similar results.

across a range of domains including mathematics, English, and IQ. As in previous studies exploring the importance of cognitive ability, we combine existing survey measures into an index (Bütikofer and Peri 2021; Lindqvist and Vestman 2011). The advantage of using longitudinal data is that we have many measures from several points in time, which we combine to create a composite measure of cognitive ability, less prone to measurement error (Bollen 1989). Cognitive ability is a latent construct that psychometricians try to measure with instruments (tests) and having only a one-shot measure increases the likelihood that the latent construct is measured with error. By using 18 tests from a variety of domains, we should get closer to the underlying latent construct of cognitive ability. Importantly, this is not a measure of IQ, but rather a construct that captures different facets of cognitive ability, including IQ. Duckworth & Seligman (2005) have shown that IQ tests also capture motivation and personality, which makes the case stronger for including multiple domains in our measure of ability. We create a standardized index of the resulting continuous scores of these 18 tests using Confirmatory Factor Analysis (CFA) (Thompson and Daniel 1996). This gives us an index that can be reasonably treated as continuous, which we then standardize. As shown in Figure 1, the measure exhibits a roughly normal distribution.

**Figure 1: The distribution of objective abilities**



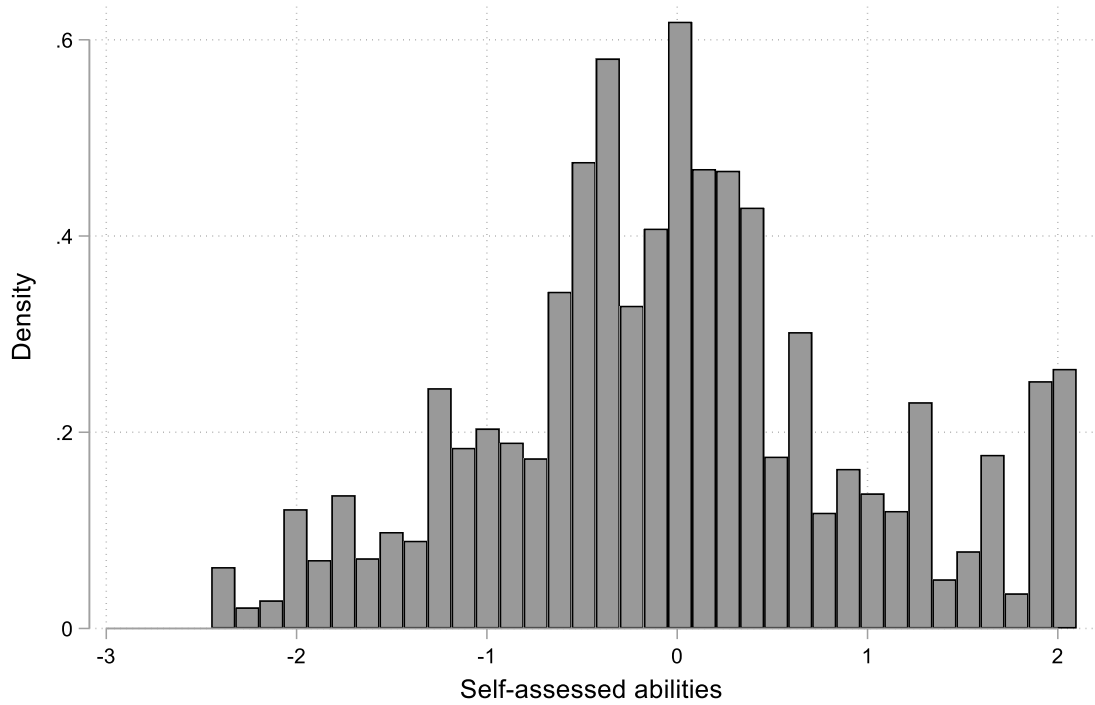

Source: BCS70 (CLS, n.d.). Number of observations = 4,429.

We construct a measure of self-assessed abilities via seven questions taken at age 10 and 16 (see the detailed explanation of measures in Table A2). This includes questions about how good the individual thinks they are at mathematics and whether they believe themselves to be clever. There are two types of questions. The "Are you good at mathematics?"-type questions have three potential answers: "Yes/No/I don't know". In these cases, it is not clear what the "I don't know" answer means: is it that they are somewhere in between "Yes" and "No", or that they are not able to tell. In our main specification, we create binary variables from these question that are equal to one if the answer is "Yes" and zero otherwise.[‡‡] In the case of the other types of questions, the answers are on three-category Likert scales and thus we use them as they are. For example, "Please say whether the following applies to you: I am clever. Applies very much/Applies somewhat/Does not apply" (see Table A2 in the Appendix for further details on the questions asked). These type of Likert scored self-assessment questions are standard in the DKE literature. For example, Dunkel et al. (2023) create their self-assessed ability measure using AdHealth data and two questions on self-assessed intelligence scored using a four-point Likert scale.

We create an index of these variables using Item Response Theory (IRT) (Edelen and Reeve 2007). Note that while on their own all questions are categorical, as we construct a summary index of these seven questions, we end up with a continuous measure of self-assessed abilities. Again, we standardize this measure. See Table A3 in the Appendix for an overview of their scale reliability. Figure 2 shows the distribution of this measure.

[‡‡] We also provide robustness checks using these answers as three-category variables and both strategies lead to similar conclusions.

**Figure 2: The distribution of self-assessed abilities**




Source: BCS70 (CLS, n.d.). Number of observations: 4,429.

Table 1 presents the descriptive statistics for these summary measures. As they have been standardized within our sample, the mean of the objective abilities and the self-assessed abilities measures is zero with a standard deviation of one.

**Table 1: Descriptive Statistics**

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Objective abilities | 4,429 | 0.000 | 1.000 | -4.074 | 2.878 |
| Objective abilities, squared | 4,429 | 1.000 | 1.432 | 0.000 | 16.597 |
| Objective abilities, cubic | 4,429 | -0.442 | 4.113 | -67.617 | 23.849 |
| Self-assessed abilities | 4,429 | 0.000 | 1.000 | -2.449 | 2.099 |
| Female | 4,429 | 0.570 | 0.495 | 0 | 1 |

Source: BCS70 (CLS, n.d.).

The most challenging task when testing the DKE is how to map one's self-assessed abilities onto the objective ability distribution (Hiller 2023). For example, what does it mean when people say *"I am clever"* in terms of their self-assessed place in the objective ability distribution? As we construct a continuous self-assessed ability measure using several questions answered via Likert scales, we are able to create a more continuous distribution of self-assessed ability with variation. Thus, we can compare everyone's place in the self-assessed and objective ability distribution.

## 3. Empirical methods

We test the statistical relationship between objective and self-assessed abilities in four ways. First, we present the classic Kruger and Dunning (1999) graphs, which plot self-assessed abilities and objective abilities by the four quartiles of objective abilities. This allows us to see whether the distance between self-assessed and objective abilities is constant along the objective abilities' distribution or whether the DKE exists.

Second, replicating Gignac and Zajenkowski (2020) and Dunkel, Nedelec, and van der Linden (2023), we look at whether a linear, a quadratic, or a cubic model fits the data the best. We assume that if the linear model fits better, the relationship between the two is constant over the distribution of objective cognitive skills, while if the quadratic or the cubic model fits better, it is not (i.e., indicating the presence of the DKE).

First, we regress estimated abilities on objective abilities in a linear model as:

$$SAA_i = \propto + \beta OA_i + \varepsilon_i \qquad (1)$$

where $SAA_i$ represents the self-assessed abilities of individual $i$, $OA_i$ represents the objective abilities of individual $i$, $\varepsilon_i$ is a heteroscedasticity-robust error term, and $\beta$ captures the correlation between $SAA_i$ and $OA_i$. Then, we extend the model with the second-order term of $OA_i$ as:

$$SAA_i = \propto + \beta_1 OA_i + \beta_2 OA_i^2 + \epsilon_i \qquad (2)$$

where $\beta_2$ captures if the relationship between $SAA_i$ and $OA_i$ is quadratic. Lastly, we also introduce the third-order term of objective abilities as:

$$SAA_i = \propto + \beta_1 OA_i + \beta_2 OA_i^2 + \beta_3 OA_i^3 + \epsilon_i \qquad (3)$$

Besides looking at whether the estimated coefficients are significant and positive, we also compare the adjusted R-squareds of the models as measures of model fit by estimating these models hierarchically using the *hireg* package in Stata.

As we will show, the quadratic model fits the data the best, so we repeat Equation (2) by gender. Then, we repeat the estimation on the pooled sample of men and women by adding the interaction terms of female and objective abilities to the model as:

$$SAA_i = \propto + \beta_1 * OA_i + \beta_2 OA_i^2 +$$

$$\beta_3 female_i + \beta_4 OA_i * female_i + \beta_5 OA_i^2 * female_i + e_i. \qquad (4)$$

Third, we also investigate the distribution of the residuals after estimating the linear model (Equation 1) by gender. Gignac (2022) and others have expressed concerns about the validity of the Glejser test to test for the DKE, thus we do not rely on the Glejser test in this paper. Instead, we directly investigate the distribution of residuals (after estimating Equation 1) by gender.

Lastly, we re-estimate the linear model (Equation 1) by the quintiles of objective abilities and show that the statistical relationship between self-assessed and objective abilities is indeed stronger as we move up the distribution of objective skills.

We provide the following robustness tests. First, we re-estimate our main results using the percentiles of objective and self-assessed abilities instead of their actual values (Model R1). Second, Gignac and Zajenkowski (2023) raised the problem of unequal test score reliability at the tails of objective ability distribution. Thus, we re-estimate our models by dropping the top and bottom 5% (Model R2) and 10% (Model R3) of the objective ability distribution to show this phenomenon is not driving our results (as well as to exclude outlier values, people with very low and very high objective cognitive skills). Third, we restrict the analytical sample to those who provided data on all 18 objective and seven self-assessed ability questions (Model R4). Fourth, we re-estimate our results using an alternative measure of objective cognitive abilities that does not rely on age 5 questions. Age 5 is a very early age that might makes the measurement of objective skills challenging. Thus, we construct an alternative measure using the 13 questions from age 10 and 16 and find very similar results (Model R5). Fifth, as mentioned above, we create an alternative measure of self-assessed abilities that handles the questions on math and reading performance as three-category variables (Model R6). Sixth, we use both the alternative objective ability measure from Model R5 and the alternative self-assessed ability measure from Model R6 in Model R7. Seventh, we exclude 441 individuals from the analytical sample who answered "Yes" to all four questions about math and spelling abilities. For these four questions, only Yes/I don't know/No answers were available; and we are worried that some people might just answer "Yes" to anything. We show that excluding these individuals would not change our results (Model R8).

Lastly, we show that selection to the analytical sample does not bias our results in Model R8 and Model R9. We do this by examining the relationship between the individual characteristics of those in the main sample and the probability of attrition and estimate a probit selection model to capture the selection mechanism. Using this model, we estimate the predicted probability of being in the analytical sample and use the inverse of these probabilities as weights to re-estimate our main results. We also make sure that selection to the analytical sample would not bias our

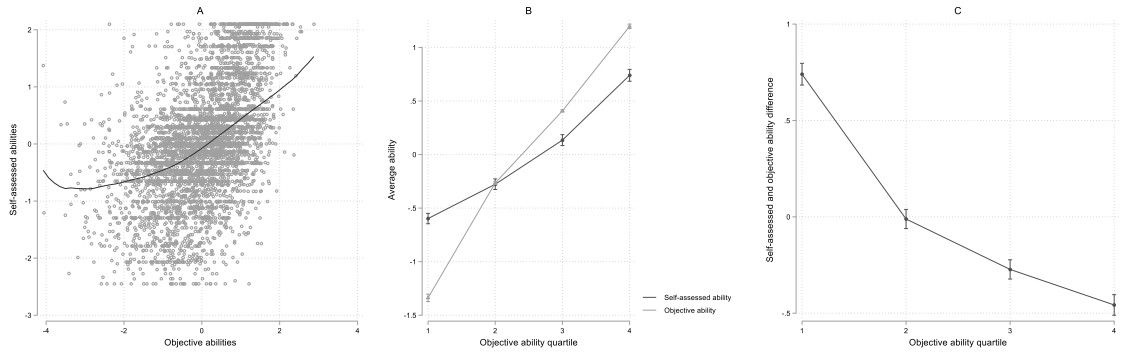
results by applying a balancing technique, entropy balancing (Hainmueller 2011), to construct individual-level weights to equate the first moments of these variables across the two groups. Using these entropy-balanced weights, we weight individuals in the main sample in such a way that their individual characteristics have the same distribution as the individual characteristics of those who were excluded from the sample.[§§]

## 4. Results

Figure 3 shows the classic DKE figures. Panel A of Figure 3 is a replication of Figure 1 in Gignac and Zajenkowski (2020), which plots self-assessed abilities vis-à-vis objective abilities and fits a nonparametric polynomial function on the data. The fitted line already suggests that the functional relationship is non-linear between the two measures. Panel B plots average self-assessed and objective abilities along the four quartiles of the objective ability distribution. Similarly, in Panel C, we plot the difference between average self-assessed and objective skills along the four quartiles to show that it decreases. Note however that the decrease of this difference along the distribution of objective skills is somewhat mechanical. While there are always going to be individuals who underestimate their abilities even with low levels of objective abilities or overestimate their abilities even with high levels of objective abilities, on average, there is more "room" for overestimation with low objective abilities and for underestimation for high objective abilities by design. This is the most important critique of using the simple difference of self-assessed and objective ability scores to describe over- and underconfidence, often referred to as floor and ceiling effects in the literature (Belmi et al. 2019).

[§§] Further detail on the creation of the weights is provided in the Appendix.

**Figure 3: Tests for the DKE using standardized measures of self-assessed and objective ability**




Notes: Replication of Figure 1 in Gignac and Zajenkowski (2020) using data from BCS70 (CLS, n.d.). No. of observations: 4,429. Panel A: scatter plot depicting the association between self-assessed and objective abilities; the line of best fit was estimated via LOESS estimation (kernel: Epanechnikov; bandwidth=0.5. Panel B: plot of self-assessed and objective ability means across the distribution of objective ability (quartiles); Panel C: plot of self-assessed and objective ability difference score means across the distribution of objective ability (quartiles).

The results shown in Figure 3 appear similar to the classic DKE figures presented in Gignac and Zajenkowski (2020). Panel A shows a positive correlation between self-assessed and objective abilities; Panel B shows support for the DKE as the magnitude of the difference between self-assessed and objective abilities is larger at the bottom of the objective ability distribution; and Panel C demonstrates a negative relationship between objective abilities and self-assessments. Unlike their simulated data, however, we observe the classic crossing of the lines in Panel B, indicating that individuals at the top of the objective ability distribution underestimate their ability while individuals at the bottom of the objective ability distribution overestimate it. Panel C highlights the same phenomenon since the difference score is negative for individuals at the top of the objective ability distribution.

**Table 2: The relationship between self-assessed and objective abilities**

| | (1) Model 1 | (2) Model 2 | (3) Model 3 |
|---|---|---|---|
| | Linear model | Quadratic model | Cubic model |
| Objective abilities | 0.500*** | 0.553*** | 0.560*** |
| | (0.014) | (0.013) | (0.021) |
| Objective abilities, squared | | 0.119*** | 0.115*** |
| | | (0.010) | (0.011) |
| Objective abilities, cubic | | | -0.003 |
| | | | (0.006) |
| Constant | 0.000 | -0.119*** | -0.117*** |
| | (0.013) | (0.016) | (0.016) |
| Observations | 4429 | 4429 | 4429 |
| $R^2$ | 0.250 | 0.276 | 0.276 |
| Adjusted $R^2$ | 0.250 | 0.276 | 0.276 |
| Change of model fit after extensions from *hireg* in Stata | | | |
| $R^2$ change | | 0.026 | 0.000 |
| F(df) change | | 160.718(1,4426) | 0.278(1,4425) |
| p-values | | 0.000 | 0.598 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 2 shows our main results more formally. The linear relationship (correlation) between self-assessed and objective abilities is 0.5, significant at the 1% significance level (Model 1). The estimated magnitude is similar, albeit somewhat larger, to those estimated in other studies, including the estimate of 0.33 obtained in a meta-analysis (Freund and Kasten 2012). This could be because both our measures are based on several questions and times of observations, thus they are most likely less biased by measurement error (i.e., they are less "noisy"). In Model 2, the estimated coefficient of the second-order term is again significant and positive, 0.119. This indicates that the relationship between self-assessed and objective abilities gets stronger as we move up the distribution, lending support to the DKE. Adding the second-order term increases the adjusted R-squared of the model by 10 percent (change divided by baseline, 0.026/0.25), suggesting that the quadratic model is a better representation of the data generating process than the linear model.

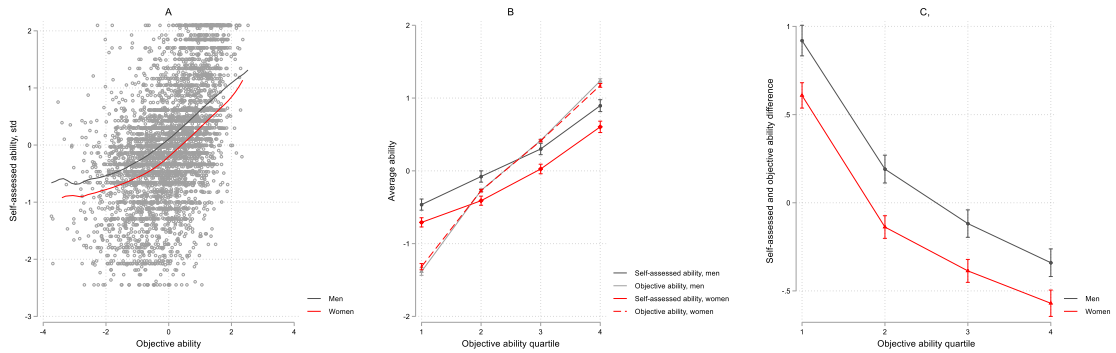
To probe this further, we follow Dunkel, Nedelec, and van der Linden (2023) and introduce a cubic term in Model 3. As opposed to the findings of Dunkel, Nedelec, and van der Linden (2023), the estimated coefficient of the cubic term is neither significant, nor does it increase the adjusted R-squared of the model. Thus, we conclude that the quadratic model fits the data best.

The results of our robustness checks are shown in Tables A4-A15 in the Appendix. Using the percentiles of the measures instead of their standardized values (Table A4), dropping

the top and bottom 5 or 10% of the sample i.e. excluding floor and ceiling effects (Tables A5 and A6), using the complete case sample (Table A7), using alternative measures of objective or self-assessed abilities (Tables A8–A10), dropping individuals who answered "Yes" to all four questions on math and spelling abilities (Table A11), as well as reweighting the sample using probit-based inverse probability weights (Table A14) or entropy-balanced weights (Table A15) all lead to the same conclusion: the relationship between self-assessed and objective skills is not linear.

We extend previous work in this area by investigating the presence of the DKE differentially by gender. The same graphs in Figure 3 are now produced separately by gender in Figure 4. Panel A shows that the curvature of the fitted line is similar for women than it is for men. Panel B shows that the crossing of the objective and self-assessed abilities lines occurs much earlier in the objective ability distribution for women than for men. This highlights how women are more underconfident in assessing their abilities. The same may be seen in Panel C, where the difference between self-assessed and objective ability is already below zero by the second quartile of the objective ability distribution for women.

**Figure 4: Tests for the DKE using standardized measures of self-assessed and objective ability by gender**




Notes: Replication of Figure 1 in Gignac and Zajenkowski (2020) using data from BCS70 (CLS, n.d.). No. of observations: 4,429. Panel A: scatter plot depicting a linear association between self-assessed and objective abilities; the line of best fit was estimated via LOESS estimation (kernel: Epanechnikov; bandwidth=0.5). Panel B: plot of self-assessed and objective ability means across the distribution of objective ability (quartiles); Panel C: plot of self-assessed and objective ability difference score means across the distribution of objective ability (quartiles).

When we compare men and women more formally in the regression setup (Table 3), the results show that women on average have about 0.3 SD lower self-assessment than men in all models. Still, the interaction terms of female and objective skills are not significantly different from zero. These results imply that women are similarly affected by the DKE as men; however,

they do not specify whether there is a gender difference in the direction of the self-assessment bias. Thus, we examine the gender differences in the DKE further.

**Table 3: The heterogeneity of the relationship between self-assessed and objective abilities by gender**

| | (1) Main model | (2) Model R1 | (3) Model R2 | (4) Model R3 | (5) Model R4 | (6) Model R5 | (7) Model R6 | (8) Model R7 |
|---|---|---|---|---|---|---|---|---|
| Female | -0.282*** | -6.724*** | -0.283*** | -0.314*** | -0.332*** | -0.284*** | -0.278*** | -0.282*** |
| | (0.032) | (2.172) | (0.037) | (0.040) | (0.057) | (0.032) | (0.031) | (0.032) |
| Objective abilities | 0.541*** | 0.311*** | 0.562*** | 0.561*** | 0.557*** | 0.544*** | 0.543*** | 0.546*** |
| | (0.019) | (0.075) | (0.026) | (0.033) | (0.037) | (0.020) | (0.019) | (0.020) |
| Objective abilities, squared | 0.108*** | 0.002*** | 0.136*** | 0.075 | 0.100*** | 0.108*** | 0.105*** | 0.105*** |
| | (0.013) | (0.001) | (0.030) | (0.050) | (0.027) | (0.014) | (0.013) | (0.014) |
| Female*objective abilities | 0.015 | -0.083 | -0.004 | 0.005 | 0.004 | 0.017 | 0.010 | 0.013 |
| | (0.027) | (0.100) | (0.035) | (0.043) | (0.050) | (0.027) | (0.027) | (0.027) |
| Female*objective abilities, squared | 0.010 | 0.001 | 0.014 | 0.114* | 0.036 | 0.010 | 0.012 | 0.012 |
| | (0.019) | (0.001) | (0.039) | (0.065) | (0.037) | (0.020) | (0.019) | (0.020) |
| Constant | 0.047* | 32.338*** | 0.033 | 0.053* | 0.030 | 0.049** | 0.047* | 0.049** |
| | (0.024) | (1.658) | (0.028) | (0.031) | (0.044) | (0.024) | (0.024) | (0.024) |
| Observations | 4429 | 4429 | 3986 | 3544 | 1337 | 4429 | 4429 | 4429 |
| $R^2$ | 0.294 | 0.284 | 0.237 | 0.191 | 0.305 | 0.294 | 0.294 | 0.293 |
| Adjusted $R^2$ | 0.294 | 0.284 | 0.236 | 0.190 | 0.303 | 0.293 | 0.293 | 0.292 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. The specifications and the samples of Models R1–R7 are explained in Section 4 in more details.

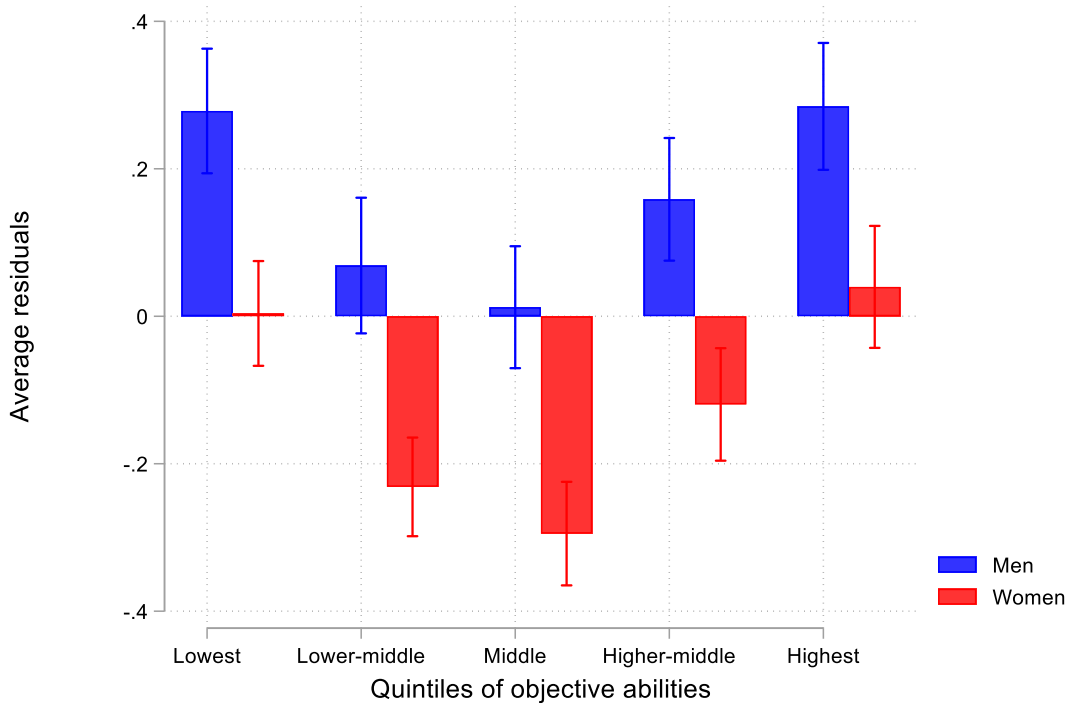
We do this by examining the sign and magnitude of the residuals obtained from the OLS estimation of Equation 1 by gender.[***] Note that the type of over- and underestimation captured by the OLS residuals is of a different nature than what is captured by the simple difference of self-assessed and objective abilities as plotted in Panels C of Figures 3 and 4 (also called a 'difference score' in the literature (Belmi et al. 2019)). The residuals of the OLS model are set to mean zero by construction, i.e. zero captures the average extent of overconfidence in the sample (Adamecz-Völgyi and Shure 2022). This will result in two things: there will be people with both negative and positive residuals along the entire distribution of objective skills and, as men are more overconfident than women (see Adamecz-Völgyi and Shure 2022), the share of those with positive and negative residuals will differ by gender.

Figure 5 plots the mean residuals by quintiles of objective ability and gender, while Figure A2 plots the share of individuals with positive residuals (i.e., those who are overconfident according to the OLS model) in the same categories. Figure A2 shows that the

[***] While we prefer not to report the results of the Glejser test in this paper due to the criticism it received, we run the test using the lmhgl package of Stata. The test rejected the null hypothesis of the homoscedasticity of the error terms (p=0.0000).

share of overconfident individuals among men is larger than 50% in four quintiles, while the share of overconfident women is only larger than 50% in the lowest quintile. Figure 5 confirms that on average, men have an overly positive evaluation of their abilities in the lowest quintile; however, women do not (at least compared to the average level of overconfidence in the sample, as explained above). In the middle three quintiles, women have on average lower self-assessment what they should have based on their objective skills, i.e., women are underconfident while men are not. In the highest quintile, men tend to be overly positive about their skills while women are not.

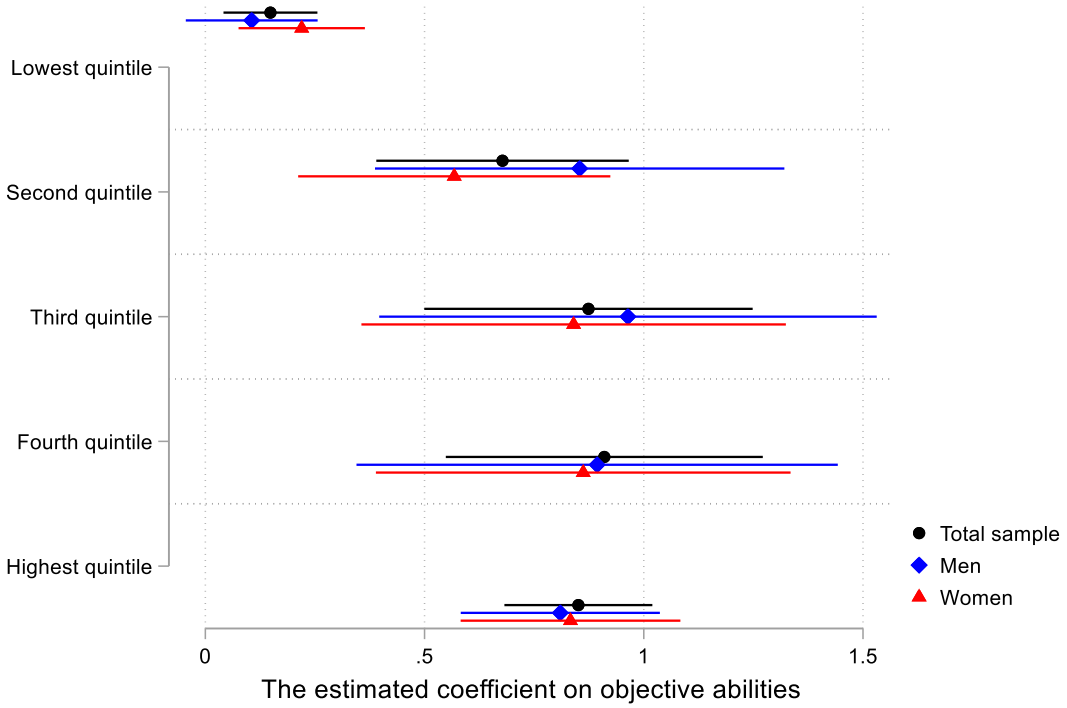
**Figure 5: Mean residuals from the linear model by gender and objective ability quintiles**




Source: BCS70 (CLS, n.d.) OLS residuals from Equation 1. N=4,429.

Lastly, Figure 6 provides further support for the DKE. It plots the estimated coefficients after re-estimating the linear model (Equation 1) by the quintiles of objective ability. These results confirm our earlier findings that indeed, the correlation between objective and subjective cognitive abilities is larger with higher levels of cognitive skills (i.e., the DKE holds in the data). As shown in Table 3, these within-quintile correlations are similar for women and men, suggesting that while the level of self-assessment is lower among women than among men, the slopes of the linear functions are similar.

**Figure 6: The estimated coefficients on objective abilities in linear models by the quintiles of objective abilities**




Source: BCS70 (CLS, n.d.). All coefficients are estimated in linear models according to Equation 1 and plotted with 95% confidence intervals based on heteroscedasticity-robust standard errors. N=4,429.

## 5. Discussion

This paper replicates and extends Gignac and Zajenkowski (2020) and Dunkel, Nedelec, and van der Linden (2023) to better understand if the DKE is indeed a statistical artefact or an actual phenomenon. We use birth cohort data instead of data collected from university students or participants in a laboratory setting. These data are nationally representative, limiting concerns around sample selection driving results. We also use multiple and repeated measures of self-assessed ability and objective ability collected during childhood and adolescence to construct our measures of objective ability and self-assessed ability. These measures are constructed to capture latent constructs, which should reduce concerns around measurement error. We combine the usage of nationally representative data with robust methods to test for the DKE that draw on the traditional tests in the literature as well as more current advancements. We are also the first paper to explore heterogeneity by gender in the DKE.

Unlike Gignac and Zajenkowski (2020), we find support for the DKE. This is in line with Dunkel, Nedelec, and van der Linden (2023), the only other study to our knowledge that uses secondary, nationally representative data to test for the DKE. This highlights the importance of sample selection issues. Unlike Dunkel, Nedelec, and van der Linden (2023), however, we find that the quadratic function better fits the data than the cubic function.

We extend previous work on the DKE by testing its heterogeneity by gender. We find that although women and men are equally bad on average at estimating their own abilities, the direction of the bias works in opposite directions: women tend to be more underconfident while men tend to be more overconfident. We find that along the entire distribution of objective skills, women tend to assess their abilities about 0.3 standard deviation lower than men. The DKE itself is similar among women and men, although we show that men are relatively more overconfident in their abilities, especially at the bottom and the top of the ability distribution, while women are relatively underconfident in their abilities, especially in the middle of the ability distribution. This is in line with popular sentiment and a range of literature that men are more overconfident than women. It also highlights the importance of assessing the direction or type of self-assessment biases. This will have implications for developing school or work-based interventions to reduce miscalibration of beliefs given the difference in the direction for men and women. While women should be encouraged to have higher self-belief and assess themselves more positively, the same is not true for men.

Our results highlight the need for better, more representative data to test for the DKE as well as better measures of objective and self-assessed ability. We improve on existing work in the field by constructing latent indices of objective and self-assessed ability using multiple measures, which should more accurately capture these constructs. It would be interesting to extend our work to include later measures of self-assessed and objective abilities to understand if the DKE holds within individuals over time.

Social scientists should care about the DKE because it is a form of miscalibration, which has always been of interest to economists in particular (e.g. Ben-David et al., 2013). Individuals hold beliefs about their abilities and these beliefs shape their actions and behavior. When these beliefs are miscalibrated to actual abilities, then individuals can make mistakes that are costly or take risks that come with a high reward. In short, it changes the distribution of outcomes,

which can have implications for markets and welfare. In the case of gender-specific self-assessments, this can lead to inequality in outcomes, e.g. wages.

The DKE has become ubiquitous in popular culture and our results lend support for its existence. One takeaway from our paper is that perhaps its magnitude is somewhat larger than previously thought. The DKE's magnitude is generally captured by how the R-squared in the regression model changes when introducing the quadratic term. We find an improvement of 2% (e.g., Table A4) to 13% (e.g. Table A14) depending on the specification. Our main specification in Table 2 estimates a 10% change in the R2 with the introduction of the quadratic term. Taken together, this range of estimates is larger than the magnitude observed in Dunkel et al. (2023) of 1.1% and includes the range of the "minimally substantively significant" size of 2-4% proposed by Gignac & Zajenkowski (2020). As Dunkel et al. (2023) also use a nationally representative sample, our improvement in explanatory power could be due to the improved measures of objective and self-assessed ability, highlighting the importance of efforts to capture latent constructs using multiple measures in representative samples.

**References**


Adamecz-Völgyi, Anna, and Nikki Shure. 2022a. 'The Gender Gap in Top Jobs – the Role of Overconfidence'. *Labour Economics*, October, 102283. https://doi.org/10.1016/j.labeco.2022.102283.

———. 2022b. 'The Gender Gap in Top Jobs – the Role of Overconfidence'. *Labour Economics*, October, 102283. https://doi.org/10.1016/j.labeco.2022.102283.

Belmi, Peter, Margaret A. Neale, David Reiff, and Rosemary Ulfe. 2019. 'The Social Advantage of Miscalibrated Individuals: The Relationship between Social Class and Overconfidence and Its Implications for Class-Based Inequality.' *Journal of Personality and Social Psychology*, May. https://doi.org/10.1037/pspi0000187.

Ben-David, Itzhak, John R. Graham, and Campbell R. Harvey. 2013. 'Managerial Miscalibration*'. *The Quarterly Journal of Economics* 128 (4): 1547–84. https://doi.org/10.1093/qje/qjt023.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. 1st ed. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118619179.

Bütikofer, Aline, and Giovanni Peri. 2021. 'How Cognitive Ability and Personality Traits Affect Geographic Mobility'. *Journal of Labor Economics* 39 (2): 559–95. https://doi.org/10.1086/710189.

CSL. n.d. '1970 British Cohort Study [Data Collection]. UK Data Service. SN: 2000001'. University College London, Institute of Education, Centre for Longitudinal Studies. https://beta.ukdataservice.ac.uk/datacatalogue/series/series?id=200001.

Duckworth, Angela L., and Martin E.P. Seligman. 2005. 'Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents'. *Psychological Science* 16 (12): 939–44. https://doi.org/10.1111/j.1467-9280.2005.01641.x.

Dunkel, Curtis S., Joseph Nedelec, and Dimitri van der Linden. 2023. 'Reevaluating the Dunning-Kruger Effect: A Response to and Replication Of'. *Intelligence* 96 (January):101717. https://doi.org/10.1016/j.intell.2022.101717.

Dunning, David. 2011. 'Chapter Five - The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance'. In *Advances in Experimental Social Psychology*, edited by James M. Olson and Mark P. Zanna, 44:247–96. Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00005-6.

Edelen, Maria Orlando, and Bryce B. Reeve. 2007. 'Applying Item Response Theory (IRT) Modeling to Questionnaire Development, Evaluation, and Refinement'. *Quality of Life Research* 16:5–18.

Ehrlinger, Joyce, and David Dunning. 2003. 'How Chronic Self-Views Influence (and Potentially Mislead) Estimates of Performance.' *Journal of Personality and Social Psychology* 84 (1): 5–17. https://doi.org/10.1037/0022-3514.84.1.5.

Freund, Philipp Alexander, and Nadine Kasten. 2012. 'How Smart Do You Think You Are? A Meta-Analysis on the Validity of Self-Estimates of Cognitive Ability.' *Psychological Bulletin* 138 (2): 296–321. https://doi.org/10.1037/a0026556.

George, Darren, and Paul Mallery. 2003. *SPSS for Windows Step by Step: A Simple Guide and Reference, 11.0 Update*. Allyn and Bacon.

Gignac, Gilles E. 2022. 'The Association between Objective and Subjective Financial Literacy: Failure to Observe the Dunning-Kruger Effect'. *Personality and Individual Differences* 184 (January):111224. https://doi.org/10.1016/j.paid.2021.111224.

Gignac, Gilles E., and Marcin Zajenkowski. 2020. 'The Dunning-Kruger Effect Is (Mostly) a Statistical Artefact: Valid Approaches to Testing the Hypothesis with Individual Differences Data'. *Intelligence* 80 (May):101449. https://doi.org/10.1016/j.intell.2020.101449.

———. 2023. ‘Still No Dunning-Kruger Effect: A Reply to Hiller’. *Intelligence* 97 (March):101733. https://doi.org/10.1016/j.intell.2023.101733.

Hainmueller, Jens. 2011. ‘Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies’. SSRN Scholarly Paper ID 1904869. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=1904869.

Hiller, Avram. 2023. ‘Comment on Gignac and Zajenkowski, “The Dunning-Kruger Effect Is (Mostly) a Statistical Artefact: Valid Approaches to Testing the Hypothesis with Individual Differences Data”’. *Intelligence* 97 (March):101732. https://doi.org/10.1016/j.intell.2023.101732.

Hippel, William von, and Robert Trivers. 2011. ‘The Evolution and Psychology of Self-Deception’. *The Behavioral and Brain Sciences* 34 (1): 1–16; discussion 16-56. https://doi.org/10.1017/S0140525X10001354.

Krajc, Marian, and Andreas Ortmann. 2008. ‘Are the Unskilled Really That Unaware? An Alternative Explanation’. *Journal of Economic Psychology* 29 (5): 724–38. https://doi.org/10.1016/j.joep.2007.12.006.

Krueger, Joachim, and Ross A. Mueller. 2002. ‘Unskilled, Unaware, or Both? The Better-than-Average Heuristic and Statistical Regression Predict Errors in Estimates of Own Performance.’ *Journal of Personality and Social Psychology* 82 (2): 180–88. https://doi.org/10.1037/0022-3514.82.2.180.

Kruger, Justin, and David Dunning. 1999. ‘Unskilled and Unaware of It: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self-Assessments.’ *Journal of Personality and Social Psychology* 77 (6): 1121–34. https://doi.org/10.1037/0022-3514.77.6.1121.

Lindqvist, Erik, and Roine Vestman. 2011. ‘The Labor Market Returns to Cognitive and Noncognitive Ability: Evidence from the Swedish Enlistment’. *American Economic Journal: Applied Economics* 3 (1): 101–28.

McIntosh, Robert D., and Sergio Della Sala. 2022. *The Persistent Irony of the Dunning-Kruger Effect*. *PSYCHOLOGIST*. Vol. 35. BRITISH PSYCHOLOGICAL SOC ST ANDREWS HOUSE, 48 PRINCESS RD EAST, LEICESTER ….

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. 2011. ‘Managing Self-Confidence: Theory and Experimental Evidence’. Working Paper 17014. National Bureau of Economic Research. https://doi.org/10.3386/w17014.

Moore, Don A., and Paul J. Healy. 2008. ‘The Trouble with Overconfidence’. *Psychological Review* 115 (2): 502–17. https://doi.org/10.1037/0033-295X.115.2.502.

Moulton, V., E. McElroy, E. Fitzsimons, Richards, M., K. Northstone, G. Conti, G.B. Ploubidis, A. Sullivan, and D. O’Neill. 2020. ‘A Guide to the Cognitive Measures in Five British Birth Cohort Studies’. London, UK: CLOSER. https://www.closer.ac.uk/cross-study-data-guides/cognitive-measures-guide/.

Niederle, Muriel, and Lise Vesterlund. 2007. ‘Do Women Shy Away From Competition? Do Men Compete Too Much?’ *The Quarterly Journal of Economics* 122 (3): 1067–1101. https://doi.org/10.1162/qjec.122.3.1067.

Ronay, Richard, William W. Maddux, and William von Hippel. 2022. ‘The Cocksure Conundrum: How Evolution Created a Gendered Currency of Corporate Overconfidence’. *Adaptive Human Behavior and Physiology* 8 (4): 557–78. https://doi.org/10.1007/s40750-022-00197-5.

Schlösser, Thomas, David Dunning, Kerri L. Johnson, and Justin Kruger. 2013. ‘How Unaware Are the Unskilled? Empirical Tests of the “Signal Extraction” Counterexplanation for the Dunning–Kruger Effect in Self-Evaluation of Performance’. *Journal of Economic Psychology* 39:85–100. https://doi.org/10.1016/j.joep.2013.07.004.

*Structural Equation Modeling Reference Manual, Release 16*. 2017. Stata Press. https://www.stata.com/bookstore/structural-equation-modeling-reference-manual/.

Tavakol, Mohsen, and Reg Dennick. 2011. 'Making Sense of Cronbach's Alpha'. *International Journal of Medical Education* 2 (June):53–55. https://doi.org/10.5116/ijme.4dfb.8dfd.

Thompson, Bruce, and Larry G. Daniel. 1996. 'Factor Analytic Evidence for the Construct Validity of Scores: A Historical Overview and Some Guidelines'. *Educational and Psychological Measurement* 56 (2): 197–208. https://doi.org/10.1177/0013164496056002001.

## Appendix

**Table A1: Measures on cognitive abilities in BCS70, age 5, 10 and 16**

| | |
|---|---|
| Age 5 | |
| English Picture Vocabulary Test | 56 sets of four different pictures with a particular word associated with each set of four pictures, increasing in difficulty. The child was asked to indicate the one picture that corresponded to the given word until the child made five mistakes in a run of eight consecutive items. The first two words were drum and time, the last two are reel and coast. |
| Copying Designs Test | The child was given a booklet, and asked to copy 8 drawings, one at a time twice on two consecutive pages of booklet. |
| Human Figure Drawing | The child was asked to 'make a picture of a man or a lady'. (Terms such as 'daddy', 'mummy', 'boy', 'girl', etc., could be used if the child responded better to those). They were asked to make the best picture they could and to draw a whole person, not just a face or head. When the child had finished, if anything was not clear, the child was asked what the various parts of the drawings were and these were labelled. |
| Complete a Profile Test | The child was asked to complete an outline picture of a human face in profile by filling in features (eyes, ears, nostrils, mouth, hair etc.). |
| Schonell Reading Test | Children's reading age (of children between age 5 and 14+ years). Reading age is calculated from the number of words read correctly and compared to the child's chronological age. Before the test was administered, the child's mother was asked if she thought the child had begun to read at all. If the mother said the child could read some words or some sentences the child was given a card with 50 words on it, which were read from left to right. When a child struggled with a word, they were asked to sound it out. If the child still couldn't say what the word was, they were asked to try the next one. The test was stopped when the child made five consecutive mistakes. |
| Age 10 | |
| Edinburgh Reading Test | A test of word recognition, which examined vocabulary, syntax, sequencing, comprehension and retention. Items were carefully selected to cover a wide age range of ability from seven to thirteen years in a form suitable to straddle the ten-year cohort. Particular attention was paid to the lower limit to allow a score to be allocated for very poor readers. |
| Friendly Maths Test | Mathematical competence, ranging from early awareness of number operations to expected mathematics ability at 13 years old, including arithmetic, number skills, fractions, measures, algebra, geometry and statistics. |
| Spelling Dictation Task | A paragraph was dictated to the child including both real and made up words. A sentence could be repeated once and an imaginary word in the middle of the passage could be repeated twice. |
| British Ability Scales (BAS) Word Definitions | For each item on the scale, a word was orally presented to the child who was asked what the word meant. Items were scored as correct or incorrect according to whether or not the child expressed key concepts of the word's meaning. The assessment was stopped after four successive incorrect or partially incorrect words. |
| BAS Word Similarities | The test consisted of 21 items made up of 3 words e.g. orange, banana, strawberry. The teacher read the three words and asked the child to name another word consistent with the group i.e. another type of fruit. The child then had to say what the words had in common i.e. they are all fruits. When the child was unable to name a group example and name on four successive attempts the test was stopped. |
| BAS Recall of Digits | For each item the teacher read out digits and asked the child to repeat them. The exercise increased in difficulty from remembering and repeating two digits to three digits and then up to eight digits. If the child asked for a repeat of the numbers, this was scored as incorrect. The test was stopped after four consecutive incorrect responses. |
| BAS Matrices | Each matrix was a square consisting of four or nine cells, with a blank cell in the lower right corner of each matrix. The teacher asked the child to complete each item by drawing the appropriate shape in the empty square. There were seven example items, three at the start of the exercise, then four examples when the level of difficulty increased. The task was stopped when four successive items were drawn incorrectly or when it was apparent that the level of difficulty was too great. |
| Pictorial Language Comprehension Test | The test consisted of 100 sets of four different pictures with a particular word associated with each set of four pictures, increasing in difficulty. The child was asked to indicate |

| | |
|---|---|
| | the one picture that corresponded to the given word. For the vocabulary Items only, the test continued until the child had five successive failures. |
| Age 16 | |
| Applied Psychology Unit (APU) Arithmetic Test | Measures general arithmetic attainment (and not aptitude). Designed to test arithmetic concepts through calculation. Covers evaluation of arithmetic expressions, knowledge of proportion, percentage, estimation of area and simple probability. It tests the ability to reproduce and therefore the aptitude to learning arithmetic processes. |
| APU Vocabulary | 75 words in the test. Each word was followed by a multiple-choice list of 5 words from which the respondent picked the one with the same meaning as the first word. The test got progressively harder. |
| BAS Matrices | Same procedure as at age 10. |
| Edinburgh Reading Test | Measures reading skills, and includes five sub-scales examining vocabulary, syntax, sequencing, comprehension and retention. |
| Spelling Test | Spelling was assessed by two tests (A and B). 100 words for each test - some spelt correctly and some incorrectly, CM identifies whether correct or incorrect. The words get harder as the test progresses. Order of test rotated by odd and even days. |

Source: Moulton et al. (2020) reproduced in Adamecz-Völgyi and Shure (2022). We construct a summary index from these 18 measures the following way. First, we standardize all these continuous measures to mean 0 and SD 1. Then, we use Confirmatory Factor Analysis (CFA) to estimate the underlying objective cognitive skills variable via Full Information Maximum Likelihood (*Structural Equation Modeling Reference Manual,* 2017). Thus, if at least one of these measures is available for a person, we will estimate the index for them.

## Table A2: Measures on subjective estimated abilities in BCS70, age 10 and 16

| | |
|---|---|
| Age 10 | |
| Good at math | Question: Are you good at mathematics? Yes/No/I don't know. |
| Good at spelling | Question: Are you good at spelling? Yes/No/I don't know. |
| Age 16 | |
| Good at math | Question: Are you good at mathematics? *Yes/No/I don't know* |
| Good at spelling | Question: Are you good at spelling? *Yes/No/I don't know* |
| Clever | Please say whether the following applies to you. *Applies very much/Applies somewhat/Does not apply*<br>I am clever. |
| Good at exams | Please say whether the following applies to you. *Applies very much/Applies somewhat/Does not apply*<br>I am good at exams. |
| Not good at school (inverted) | Please say whether the following applies to you. *Applies very much/Applies somewhat/Does not apply*<br>I am not very good at school. |

Source: Adamecz-Völgyi and Shure (2022). We construct a summary index from these seven categorical (ordinal) measures using Item Response Theory (IRT). We fit graded response models to these measures, and we allow them to vary in their difficulty and discrimination. Again, we exploit all information: if at least one of these measures is available for a person, we will estimate the latent index for them.

## Table A3: Scale reliability measures

| Index | Number of items | Average inter-item covariance | Cronbach's alpha |
|---|---|---|---|
| Cognitive index (CFA) | 18 | 0.326 | 0.897 |
| Self-assessment index (IRT) | 7 | 0.068 | 0.678 |

Source: BCS70 (CLS, n.d.). The Cronbach's alpha for the cognitive index is well above the rule of thumb threshold of 0.7 (George and Mallery 2003); however, the Cronbach's alpha for the self-assessment index is just below 0.7. This could be partially explained by the lower number of items used to create this scale (Tavakol and Dennick 2011). Given its proximity to 0.7, this is still acceptable.

**Table A4: The relationship between self-assessed and objective abilities – Robustness test No. 1: using the percentiles of the objective and self-assessed ability measures instead of their standardized values (Model R1)**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| | Linear model | Quadratic model | Cubic model |
| | Outcome variable: Self-assessed ability percentiles | | |
| Objective ability percentiles | 0.510*** | 0.239*** | 0.443*** |
| | (0.012) | (0.050) | (0.130) |
| | | | |
| Objective abilities percentiles, squared | | 0.003*** | -0.002 |
| | | (0.000) | (0.003) |
| | | | |
| Objective abilities percentiles, cubic | | | 0.000* |
| | | | (0.000) |
| | | | |
| Constant | 24.240*** | 28.832*** | 27.078*** |
| | (0.726) | (1.088) | (1.494) |
| Observations | 4429 | 4429 | 4429 |
| $R^2$ | 0.261 | 0.266 | 0.266 |
| Adjusted $R^2$ | 0.261 | 0.266 | 0.266 |
| Change of model fit after extensions from *hireg* in Stata | | | |
| $R^2$ change | | 0.005 | 0.000 |
| F(df) change | | 28.939(1,4426) | 2.844(1,4425) |
| p-values | | 0.000 | 0.092 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table A5: The relationship between self-assessed and objective abilities – Robustness test No. 2: dropping the top and bottom 5% of the objective ability distribution (Model R2)**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| | Linear model | Quadratic model | Cubic model |
| | Outcome variable: Self-assessed abilities | | |
| Objective abilities | 0.544*** | 0.563*** | 0.584*** |
| | (0.017) | (0.017) | (0.036) |
| | | | |
| Objective abilities, squared | | 0.153*** | 0.149*** |
| | | (0.020) | (0.021) |
| | | | |
| Objective abilities, cubic | | | -0.016 |
| | | | (0.024) |
| | | | |
| Constant | -0.038*** | -0.137*** | -0.135*** |
| | (0.014) | (0.018) | (0.018) |
| Observations | 3986 | 3986 | 3986 |
| $R^2$ | 0.205 | 0.217 | 0.217 |
| Adjusted $R^2$ | 0.205 | 0.217 | 0.217 |
| Change of model fit after extensions from *hireg* in Stata | | | |
| $R^2$ change | | 0.012 | 0.000 |
| F(df) change | | 59.381(1,3983) | 0.469(1,3982) |
| p-values | | 0.000 | 0.494 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table A6: The relationship between self-assessed and objective abilities – Robustness test No. 3: dropping the top and bottom 10% of the objective ability distribution (Model R3)**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| | Linear model | Quadratic model | Cubic model |
| | Outcome variable: Self-assessed abilities | | |
| Objective abilities | 0.566*** | 0.569*** | 0.577*** |
| | (0.021) | (0.021) | (0.049) |
| | | | |
| Objective abilities, squared | | 0.147*** | 0.147*** |
| | | (0.032) | (0.032) |
| | | | |
| Objective abilities, cubic | | | -0.009 |
| | | | (0.049) |
| | | | |
| Constant | -0.067*** | -0.135*** | -0.135*** |
| | (0.014) | (0.020) | (0.020) |
| Observations | 3544 | 3544 | 3544 |
| $R^2$ | 0.167 | 0.172 | 0.172 |
| Adjusted $R^2$ | 0.167 | 0.172 | 0.172 |
| Change of model fit after extensions from *hireg* in Stata | | | |
| $R^2$ change | | 0.005 | 0.000 |
| F(df) change | | 21.560(1,3541) | 0.037(1,3540) |
| p-values | | 0.000 | 0.848 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table A7: The relationship between self-assessed and objective abilities – Robustness test No. 4: complete case sample of those who provided data on all 18 objective and seven self-assessed ability measure (Model R4)**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| | Linear model | Quadratic model | Cubic model |
| | Outcome variable: Self-assessed abilities | | |
| Objective abilities | 0.522*** | 0.561*** | 0.605*** |
| | (0.025) | (0.025) | (0.039) |
| | | | |
| Objective abilities, squared | | 0.122*** | 0.106*** |
| | | (0.019) | (0.024) |
| | | | |
| Objective abilities, cubic | | | -0.019 |
| | | | (0.013) |
| | | | |
| Constant | -0.045* | -0.159*** | -0.151*** |
| | (0.023) | (0.029) | (0.030) |
| Observations | 1337 | 1337 | 1337 |
| $R^2$ | 0.258 | 0.280 | 0.281 |
| Adjusted $R^2$ | 0.257 | 0.279 | 0.280 |
| Change of model fit after extensions from *hireg* in Stata | | | |
| $R^2$ change | | 0.023 | 0.001 |
| F(df) change | | 41.742(1,1334) | 1.977(1,1333) |
| p-values | | 0.000 | 0.160 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table A8: The relationship between self-assessed and objective abilities – Robustness test No. 5: using an alternative measure of cognitive skills relying only on data from ages 10 and 16 (Model R5)**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| | Linear model | Quadratic model | Cubic model |
| | Outcome variable: Self-assessed abilities | | |
| Objective abilities | 0.500*** | 0.557*** | 0.559*** |
| | (0.014) | (0.014) | (0.021) |
| | | | |
| Objective abilities, squared | | 0.120*** | 0.119*** |
| | | (0.010) | (0.012) |
| | | | |
| Objective abilities, cubic | | | -0.001 |
| | | | (0.007) |
| | | | |
| Constant | 0.000 | -0.120*** | -0.119*** |
| | (0.013) | (0.016) | (0.017) |
| Observations | 4429 | 4429 | 4429 |
| $R^2$ | 0.250 | 0.276 | 0.276 |
| Adjusted $R^2$ | 0.250 | 0.275 | 0.275 |
| Change of model fit after extensions from *hireg* in Stata | | | |
| $R^2$ change | | 0.026 | 0.000 |
| F(df) change | | 158.931(1,4426) | 0.008(1,4425) |
| p-values | | 0.000 | 0.929 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table A9: The relationship between self-assessed and objective abilities – Robustness test No. 6: using an alternative measure of self-assessed abilities (Model R6)**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| | Linear model | Quadratic model | Cubic model |
| | Outcome variable: Self-assessed abilities | | |
| Objective abilities | 0.501*** | 0.553*** | 0.559*** |
| | (0.014) | (0.014) | (0.021) |
| | | | |
| Objective abilities, squared | | 0.117*** | 0.114*** |
| | | (0.010) | (0.011) |
| | | | |
| Objective abilities, cubic | | | -0.002 |
| | | | (0.006) |
| | | | |
| Constant | 0.000 | -0.117*** | -0.115*** |
| | (0.013) | (0.016) | (0.016) |
| Observations | 4429 | 4429 | 4429 |
| $R^2$ | 0.251 | 0.276 | 0.276 |
| Adjusted $R^2$ | 0.251 | 0.276 | 0.276 |
| Change of model fit after extensions from *hireg* in Stata | | | |
| $R^2$ change | | 0.025 | 0.000 |
| F(df) change | | 155.812(1,4426) | 0.174(1,4425) |
| p-values | | 0.000 | 0.677 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table A10: The relationship between self-assessed and objective abilities – Robustness test No. 7: using an alternative measure of cognitive skills relying only on data from ages 10 and 16 and an alternative measure of self-assessed abilities (Model R7)**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| | Linear model | Quadratic model | Cubic model |
| | Outcome variable: Self-assessed abilities | | |
| Objective abilities | 0.500*** | 0.557*** | 0.556*** |
| | (0.014) | (0.014) | (0.021) |
| | | | |
| Objective abilities, squared | | 0.118*** | 0.118*** |
| | | (0.010) | (0.012) |
| | | | |
| Objective abilities, cubic | | | 0.000 |
| | | | (0.007) |
| | | | |
| Constant | 0.000 | -0.118*** | -0.118*** |
| | (0.013) | (0.016) | (0.017) |
| Observations | 4429 | 4429 | 4429 |
| $R^2$ | 0.250 | 0.275 | 0.275 |
| Adjusted $R^2$ | 0.250 | 0.275 | 0.274 |
| Change of model fit after extensions from *hireg* in Stata | | | |
| $R^2$ change | | 0.025 | 0.000 |
| F(df) change | | 152.934(1,4426) | 0.000(1,4425) |
| p-values | | 0.000 | 0.994 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Table A11: The relationship between self-assessed and objective abilities – Robustness test No. 8: excluding the 441 people from the sample who answered "Yes" to all four self-assessment questions (Model R8)**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| | Linear model | Quadratic model | Cubic model |
| | Outcome variable: Self-assessed abilities | | |
| Objective abilities | 0.435*** | 0.495*** | 0.498*** |
| | (0.014) | (0.015) | (0.021) |
| | | | |
| Objective abilities, squared | | 0.103*** | 0.101*** |
| | | (0.010) | (0.013) |
| | | | |
| Objective abilities, cubic | | | -0.001 |
| | | | (0.007) |
| | | | |
| Constant | -0.092*** | -0.188*** | -0.187*** |
| | (0.013) | (0.016) | (0.017) |
| Observations | 3988 | 3988 | 3988 |
| $R^2$ | 0.207 | 0.228 | 0.228 |
| Adjusted $R^2$ | 0.207 | 0.227 | 0.227 |
| Change of model fit after extensions from *hireg* in Stata | | | |
| $R^2$ change | | 0.021 | 0.000 |
| F(df) change | | 107.237(1,3985) | 0.041(1,3984) |
| p-values | | 0.000 | 0.840 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. In this robustness check we exclude those people (n=441) who answered "Yes" to all of the following four

questions: Are you good at math, age 10; Are you good at math, age 16; Are you good at spelling, age 10; Are you good at spelling, age 16.

**Table A12: The balance table of the analytical sample compared to the rest of the Wave 1 sample of BCS70**

| | Mean | Mean | Diff. | SE | t-test |
|---|---|---|---|---|---|
| | Rest of the sample | Analytical sample | | | p-value |
| Female | 0.45 | 0.57 | 0.12 | 0.01 | 0.000 |
| Parental high SES | 0.30 | 0.38 | 0.08 | 0.01 | 0.000 |
| Mother has qualification | 0.37 | 0.61 | 0.24 | 0.01 | 0.000 |
| Father has qualification | 0.67 | 0.94 | 0.27 | 0.01 | 0.000 |
| Low birth weight | 0.08 | 0.06 | -0.03 | 0.00 | 0.000 |
| Mother's year of birth | 1944.21 | 1943.84 | -0.37 | 0.09 | 0.000 |
| Ethnicity: English | 0.64 | 0.93 | 0.29 | 0.01 | 0.000 |
| Ethnicity: Irish | 0.00 | 0.00 | 0.00 | 0.00 | 0.970 |
| Ethnicity: Other European | 0.00 | 0.00 | 0.00 | 0.00 | 0.553 |
| Ethnicity: West Indian | 0.01 | 0.01 | 0.00 | 0.00 | 0.028 |
| Ethnicity: Indian | 0.01 | 0.01 | 0.00 | 0.00 | 0.395 |
| Ethnicity: Pakistani | 0.00 | 0.00 | 0.00 | 0.00 | 0.500 |
| Ethnicity: Bangladeshi | 0.00 | 0.00 | 0.00 | 0.00 | 0.025 |
| Ethnicity: Other | 0.00 | 0.00 | 0.00 | 0.00 | 0.115 |
| Ethnicity: Missing | 0.33 | 0.04 | -0.29 | 0.01 | 0.000 |
| Region: North | 0.05 | 0.07 | 0.01 | 0.00 | 0.001 |
| Region: Yorks and Humberside | 0.09 | 0.08 | 0.00 | 0.00 | 0.516 |
| Region: East Midland | 0.06 | 0.07 | 0.01 | 0.00 | 0.013 |
| Region: East | 0.03 | 0.04 | 0.02 | 0.00 | 0.000 |
| Region: South East | 0.29 | 0.26 | -0.03 | 0.01 | 0.000 |
| Region: South West | 0.06 | 0.07 | 0.01 | 0.00 | 0.007 |
| Region: West Midlands | 0.09 | 0.11 | 0.02 | 0.01 | 0.002 |
| Region: North West | 0.12 | 0.13 | 0.01 | 0.01 | 0.022 |
| Region: Wales | 0.04 | 0.07 | 0.02 | 0.00 | 0.000 |
| Region: Scotland | 0.09 | 0.09 | 0.00 | 0.01 | 0.563 |
| Region: Northern Ireland | 0.05 | 0.00 | -0.05 | 0.00 | 0.000 |
| Observations | 13,154 | 4,429 | | | |

Source: BCS70 (CLS, n.d.). Total sample of 17,583 individuals.

In Table A12, we look at how the individual characteristics of those in the main sample relate to the characteristics of those who dropped out or did not report data. As we cannot construct cognitive ability and subjective self-assessment measures for the whole sample, we use characteristics that are available for everybody from the first two waves: gender, region of birth, socio-economic background of parents, whether their mother and father had any qualifications,

mother's year of birth, ethnicity and low (<2500 g) birthweight. As we find that there are some differences between the two groups, we estimate a probit selection model to capture the selection mechanism (Table A13). Using this model, we estimate the predicted probability of being in the analytical sample and use the inverse of these probabilities as weights to re-estimate our main results (Table A14).

We show in Figure A1 that using these weights eliminates statistical differences between those in the analytical sample and those who were excluded. Re-estimating our (unweighted) main results using these entropy-balanced weights leads to similar results (Table A15); thus, we are confident that (observed) sample selection is not driving our results.

**Table A13: Selection to the analytical sample – probit model based on the observable characteristics of cohort members at ages 0 and 5**

| | (1) | |
|---|---|---|
| | Model 1 | |
| | | |
| Female | 0.308*** | (0.022) |
| Parental high SES | 0.130*** | (0.025) |
| Mother has qualification | 0.341*** | (0.024) |
| Father has qualification | 0.399*** | (0.042) |
| Low birth weight | -0.106** | (0.046) |
| Mother's year of birth | -0.006*** | (0.002) |
| Ethnicity baseline: English | | |
| Ethnicity: Irish | -0.045 | (0.179) |
| Ethnicity: Other European | 0.026 | (0.174) |
| Ethnicity: West Indian | -0.313*** | (0.121) |
| Ethnicity: Indian | 0.022 | (0.113) |
| Ethnicity: Pakistani | -0.097 | (0.171) |
| Ethnicity: Bangladeshi | 0.000 | (.) |
| Ethnicity: Other | -0.418 | (0.298) |
| Ethnicity: Missing | -0.921*** | (0.045) |
| Region baseline: North | | |
| Region: Yorks and Humberside | -0.149*** | (0.057) |
| Region: East Midland | 0.053 | (0.062) |
| Region: East | 0.188*** | (0.073) |
| Region: South East | -0.204*** | (0.049) |
| Region: South West | -0.055 | (0.061) |
| Region: West Midlands | -0.003 | (0.055) |
| Region: North West | -0.033 | (0.053) |
| Region: Wales | 0.081 | (0.063) |
| Region: Scotland | -0.137** | (0.056) |
| Region: Northern Ireland | -1.009*** | (0.229) |
| Constant | 11.020*** | (4.024) |
| Observations | 17152 | |

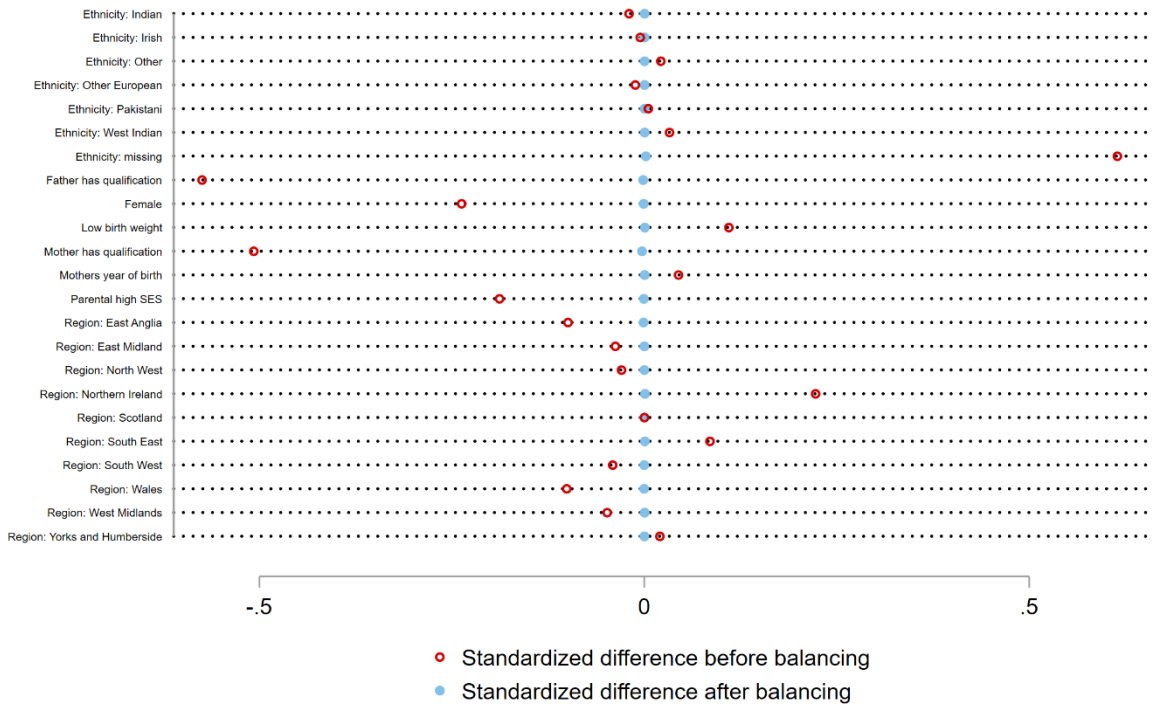
Source: BCS70 (CLS, n.d.). Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. ROC measure: 0.753. No standard errors are reported for the "Ethnicity: Bangladeshi" category, because this category perfectly predicts failure and thus was automatically omitted from the estimation sample. Five individuals reported Bangladeshi ethnicity in the sample.

**Table A14: The relationship between self-assessed and objective abilities – Robustness test No. 9: regressions with inverse probability weighting to handle selection to the analytical sample (Model R9)**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| | Linear model | Quadratic model | Cubic model |
| | Outcome variable: Self-assessed abilities | | |
| Objective abilities | 0.457*** | 0.545*** | 0.554*** |
| | (0.025) | (0.022) | (0.037) |
| | | | |
| Objective abilities, squared | | 0.118*** | 0.112*** |
| | | (0.014) | (0.016) |
| | | | |
| Objective abilities, cubic | | | -0.004 |
| | | | (0.009) |
| | | | |
| Constant | 0.023 | -0.096*** | -0.092*** |
| | (0.022) | (0.027) | (0.027) |
| Observations | 4429 | 4429 | 4429 |
| $R^2$ | 0.234 | 0.263 | 0.264 |
| Adjusted $R^2$ | 0.233 | 0.263 | 0.263 |
| Change of model fit after extensions from *hireg* in Stata | | | |
| $R^2$ change | | 0.030 | 0.000 |
| F(df) change | | 179.876(1,4426) | 0.597(1,4425) |
| p-values | | 0.000 | 0.440 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Figure A1: The balance of the analytical sample before and after entropy balancing**



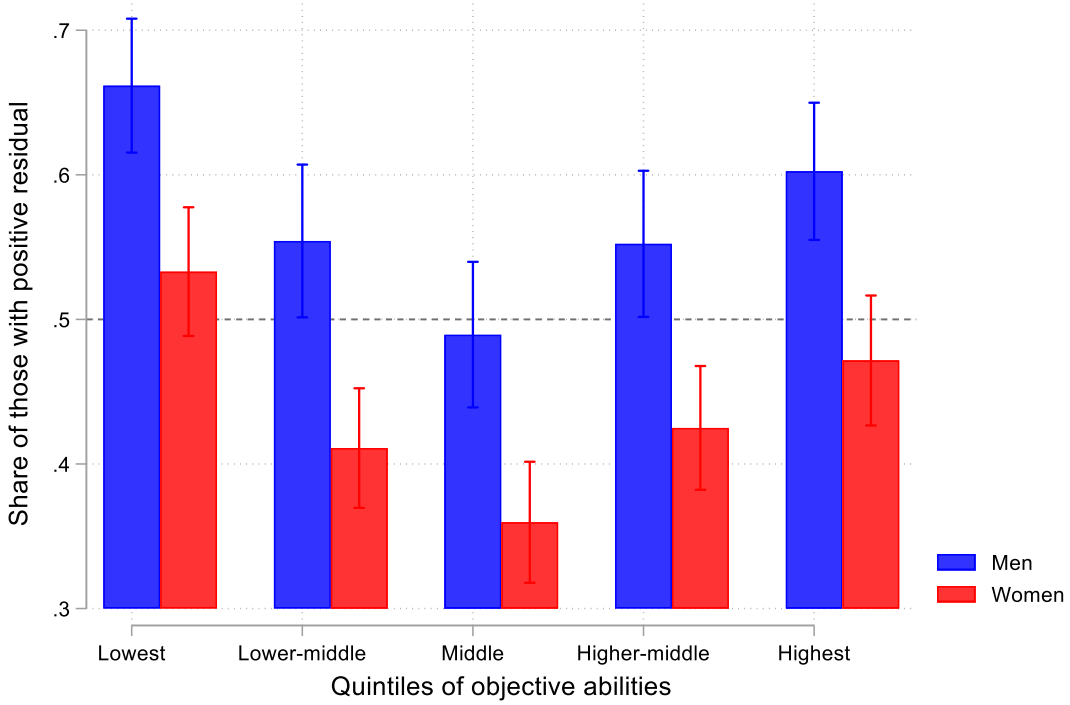
Source: BCS70 (CLS, n.d.).Baseline categories of categorical variables are not plotted (Ethnicity: English; Region: North). Entropy balancing is a reweighting procedure to achieve covariate balance with binary treatments based on the moments of the covariates (Hainmueller 2011).

**Table A15: The relationship between self-assessed and objective abilities – Robustness test No. 10: regressions with entropy-balanced weights to handle selection to the analytical sample (Model R10)**

| | (1) | (2) | (3) |
|---|---|---|---|
| | Model 1 | Model 2 | Model 3 |
| | Linear model | Quadratic model | Cubic model |
| | Outcome variable: Self-assessed abilities | | |
| Objective abilities | 0.423*** | 0.527*** | 0.523*** |
| | (0.036) | (0.034) | (0.053) |
| | | | |
| Objective abilities, squared | | 0.117*** | 0.120*** |
| | | (0.021) | (0.023) |
| | | | |
| Objective abilities, cubic | | | 0.002 |
| | | | (0.012) |
| | | | |
| Constant | 0.023 | -0.091** | -0.094** |
| | (0.032) | (0.040) | (0.038) |
| Observations | 4429 | 4429 | 4429 |
| $R^2$ | 0.212 | 0.244 | 0.244 |
| Adjusted $R^2$ | 0.212 | 0.244 | 0.244 |
| Change of model fit after extensions from *hireg* in Stata | | | |
| $R^2$ change | | 0.032 | 0.000 |
| F(df) change | | 187.750(1,4426) | 0.120(1,4425) |
| p-values | | 0.000 | 0.729 |

Source: BCS70 (CLS, n.d.). Heteroscedasticity-robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

**Figure A2: The share of those with positive OLS residuals from Equation 1 by gender**




Source: BCS70 (CLS, n.d.). The share of those with positive residuals estimated in linear models according to Equation 1 and plotted with the 95% confidence intervals of the means. N=4,429.