Title: Common Genetic Variants Associated with Cognitive Performance Identified Using Proxy-Phenotype Method

Authors: All authors and their affiliations appear at the end of the paper*

*Correspondence to: Daniel Benjamin (db468@cornell.edu)
or Philipp Koellinger (p.d.koellinger@uva.nl).

Abstract: We identify common genetic variants associated with cognitive performance using a two-stage approach, which we call the “proxy-phenotype method.” First, we conduct a genome-wide association study of educational attainment in a large sample ($N = 106,736$), which produces a set of 69 “education-associated single-nucleotide polymorphisms (SNPs).” Second, using independent samples ($N = 24,189$), we measure the association of these education-associated SNPs with cognitive performance. Three SNPs (rs1487441, rs7923609, rs2721173) are significantly associated with cognitive performance after correction for multiple hypothesis testing. In an independent sample of older Americans ($N = 8,652$), we also show that a polygenic score derived from the education-associated SNPs is associated with memory and absence of dementia. Convergent evidence from a set of bioinformatics analyses implicates four specific genes ($KNCMA1$, $NRXN1$, $POU2F3$, $SCRT$). All of these are associated with a particular neurotransmitter pathway involved in synaptic plasticity, the main cellular mechanism for learning and memory.

Significance Statement: We identify several common genetic variants associated with cognitive performance using a two-stage approach: we conduct a genome-wide association study of educational attainment to generate a set of candidates, then we measure the association of these variants with cognitive performance. In older Americans, we find that these variants are jointly associated with cognitive health. Bioinformatics analyses implicate a set of genes that are associated with a particular neurotransmitter pathway involved in synaptic plasticity, the main cellular mechanism for learning and memory. In addition to the substantive contribution, this paper also serves to demonstrate a “proxy-phenotype” approach to discovering common genetic variants that is likely to be useful for many phenotypes of interest to social scientists (such as personality traits).
Introduction: Twin and family studies have shown that at least a moderate share of variation in most facets of cognitive performance (i.e., performance by healthy individuals on cognitive tests) is associated with genetic factors (1, 2). However, despite considerable interest and effort, research to date has largely failed to identify common genetic variants associated with cognitive performance phenotypes (3–5), with the exception of APOE which predicts cognitive decline in older individuals (6–8). Existing studies have relied on one of two research strategies. The first is a candidate-gene design, in which researchers test a small number of genetic variants for association with the phenotype of interest, typically based on hypotheses derived from the known biological functions of the candidate genes. The candidate-gene associations that have been reported with cognitive performance (9), however, fail to replicate when larger samples are used (3). The second research strategy is a genome-wide association study (GWAS), in which researchers atheoretically test hundreds of thousands of single-nucleotide polymorphisms (SNPs) for association with the phenotype and apply a threshold for “genome-wide” statistical significance—typically $5 \times 10^{-8}$—in order to account for multiple-hypothesis testing. For physical and medical phenotypes, GWASs have identified many novel associations that replicate (10). GWASs on cognitive performance, however, have not yet identified any genome-wide significant associations (4, 5).

Here, we apply an alternative, two-stage research strategy, which we call the proxy-phenotype method. In the first stage, we conduct a GWAS on a “proxy phenotype” to identify a relatively small set of SNPs that are associated with the proxy phenotype. In the second stage, these SNPs serve as candidates that are tested in independent samples for association with the phenotype of interest, at a significance threshold corrected for the number of proxy-associated SNPs. In the study reported here, our phenotype of interest is cognitive performance, for which we use Spearman’s measure of general cognitive ability (usually abbreviated to $g$; it is the general factor measured by a battery of diverse cognitive tests (4)). Our proxy phenotype is educational attainment, as measured by self-reported years of schooling.

Rietveld et al. (11) had suggested the strategy of using SNPs associated with educational attainment as “empirically-based candidate genes” for association with cognitive performance; here we conduct that analysis and further develop the methodology for doing so. The SI Appendix contains our formal framework, building on that in (11), as well as power calculations under a range of assumptions. According to the framework, educational attainment is a good proxy phenotype for cognitive performance because cognitive performance is strongly genetically influenced and causally affects educational attainment. The genetic correlation between the two traits does not have straightforward implications for the statistical power to identify specific SNPs influencing cognitive performance; nonetheless, the high genetic correlation (estimated to be roughly 0.65 or higher (12–14)) may also provide a suggestive justification for the approach.

Results: In our first stage, we conducted a GWAS of educational attainment in a pooled “Education Sample” of 106,736 individuals. We used the same data, analysis protocol, and quantitative years-of-schooling measure as (11), except that we omit cohorts with high-quality measures of cognitive performance; we instead include these cohorts in the subsequent “Cognitive Performance Sample.” We chose our “inclusion threshold” of $p < 10^{-5}$ for selecting candidate SNPs based on ex ante power calculations whose goal was to maximize the number of
true positives among the candidates (see SI Appendix). Pruning for linkage disequilibrium the 927 SNPs that reach this threshold resulted in 69 approximately independent SNPs (see SI Appendix).

In our second stage, we tested these 69 “education-associated SNPs” for association with cognitive performance in the Cognitive Performance Sample, which comprises 24,189 genotyped subjects from 11 cohorts (see SI Appendix section 2). The specific cognitive tests differ across cohorts, but the cognitive performance measure in every cohort is calculated as Spearman’s g (see SI Appendix); previous research has found that g from different test batteries are highly correlated, especially if the batteries have many tests, or if the test is specifically constructed to measure g (15–17). We tested each SNP individually for association with cognitive performance using ordinary least squares, controlling for sex, age, and (depending on the cohort) at least four principal components of the genome-wide data (to reduce confounding from population stratification). At the cohort level, the analyses were conducted according to a prespecified plan that we preregistered on the Open Science Framework (see https://osf.io/z7fe2/). The cohorts’ results were then meta-analyzed using an inverse-variance weighting scheme. Two independent teams of analysts crosschecked and verified the results.

To confirm that the education-based first stage identifies reasonable candidate SNPs for cognitive performance, Figure 1 plots the standardized regression coefficients from the regression of years-of-schooling on the education-associated SNPs in the Education Sample (with the reference allele chosen to ensure the coefficient is positive) against the standardized coefficients from the second-stage regression of cognitive performance on the SNPs in the Cognitive Performance Sample. The direction of the effect coincides in 53 out of 69 cases (two-sided binomial test, \( p = 9.10 \times 10^{-6} \)), indicating that this is a good context for applying the proxy-phenotype method. We were surprised that the correlation between the effect size on educational attainment and the effect size on cognitive performance is negative (\( \rho = -0.25; \ p = 0.03 \)), although not significantly after dropping a possible outlier, the bottom-most point of the figure (\( \rho = -0.14 \ p = 0.26 \)). Within our theoretical framework, a negative correlation suggests that SNPs that affect cognitive performance more strongly tend to affect other factors that matter for educational attainment (such as personality traits) less strongly, and vice-versa (see SI Appendix).

To provide a benchmark for evaluating our list of education-associated candidate SNPs, we generated (via a pre-specified algorithm) a list of “theory-based” candidate SNPs for cognitive performance drawn from published findings in the candidate-gene literature (see SI Appendix). (This list does not include the SNPs comprising the APOE haplotype because these SNPs were not available in the cohort GWAS results.) After applying the same pruning procedure as for the education-associated SNPs, our list of theory-based SNPs contains 24 independent SNPs, of which only one is in a genomic region close to an education-associated SNP. Figure 2 overlays Q–Q plots for the theory-based and education-associated candidates. The education-associated candidates taken altogether are more strongly associated with cognitive performance than would be expected by chance (\( z = 5.98, \ p = 1.12 \times 10^{-9} \)). Whereas a visual inspection of the plot suggests that the theory-based candidates exhibit some association with cognitive performance, we cannot reject the null hypothesis for any SNP individually, nor for all of them taken together (\( z = 1.19, \ p = 0.12 \)).
The top three education-associated SNPs—rs1487441, rs7923609, and rs2721173—show clear separation from the others in Figure 2 and are significantly associated with cognitive performance after Bonferroni correction for multiple hypothesis testing (see Table 1). Consistent with the negative correlation in Figure 1, these SNPs are different from the three SNPs that reached genome-wide significance for association with educational attainment in the (11) analyses. After adjusting the SNPs’ estimated effect sizes (each $R^2 \approx 0.0006$) for the winner’s curse, we estimate each as $R^2 \approx 0.0002$ (see SI Appendix), or in terms of coefficient magnitude, each additional reference allele for each SNP is associated with $\approx 0.02$ standard-deviation increase in cognitive performance (or 0.3 points on the typical “IQ” scale). This $R^2 \approx 0.0002$ is about the same as the $R^2$ for the known SNP associations with educational attainment (11) but far smaller than the largest effect sizes for complex physical traits such as height ($R^2 \approx 0.004$) and BMI ($R^2 \approx 0.003$) (18, 19).

Power calculations we report in the SI Appendix help shed light on why the proxy-phenotype method succeeded in identifying SNPs even though GWA studies to date on cognitive performance have not. A GWAS in our Cognitive Performance Sample of $N = 24,189$—which is larger than the largest GWA studies ($N = 17,989$ in Benyamin et al. (2014) and $N = 3,511$ in Davies et al. (2011))—would have had power 0.06% to identify a SNP whose association has $R^2 = 0.0002$. In contrast, our proxy-phenotype approach had power 12%. Given this power and the rather stringent significance threshold (0.05/69 $\approx .00072$), Bayesian calculations using reasonable assumptions regarding priors suggest that the posterior probabilities that these three SNPs are associated with cognitive performance are high (see SI Appendix).

Turning from specific SNPs to the set of all 69 education-associated SNPs, we assess the explanatory power of a linear polygenic score that aggregates their coefficients (see SI Appendix). In pooled results from four family-based cohorts (4,463 individuals in total), we find that the score is significantly associated with cognitive performance ($p = 8.17 \times 10^{-4}$), with $R^2$ ranging approximately from 0.2% to 0.4% across samples. Using only within-family variation, the pooled coefficient has the same sign but is smaller and has a larger standard error ($p = 0.36$). Thus we cannot rule out that some of the score’s explanatory power is due to population stratification, although even without stratification, the non-significance of the within-family coefficient is not surprising given the low power of this test (see SI Appendix).

Next, we explore whether educational attainment might serve as a proxy phenotype for cognitive-health phenotypes (as opposed to cognitive performance in the normal range). Our sample comprises 8,652 European-descent individuals over the age of 50 from the Health and Retirement Study (HRS) (see SI Appendix). We confirm that, for the 60 out of 69 SNPs available in the HRS data, the direction of the effects on educational attainment generally coincides with the direction of the effects on the two cognitive-health phenotypes we study: “total word recall,” which is a test for memory problems (two-sided binomial test, $p = 0.0067$); and “total mental status,” which is a battery that screens for early signs of dementia ($p = 0.0775$).

Next, we obtain the weights for a polygenic score by conducting a de novo meta-GWAS analysis of educational attainment just as in the first stage described above, but this time excluding the HRS from the Education Sample.

Figure 3 shows that the score is associated with both of the cognitive-health phenotypes. The strength of the protective effect is approximately constant across age categories from age 50 to 80, and becomes weaker for total word recall after age 80. These associations are essentially unaffected when we control for up to 20 principal components of the genome-wide data,
suggesting that the associations are not driven by population stratification (20). The $R^2$ of these associations range roughly 0.2%-0.4% (similar magnitudes as in the analysis of cognitive performance in the family-based cohorts). When we control for years of schooling, the estimated effect of the score falls roughly in half but remains statistically significant (see SI Appendix). The score is not associated with cognitive decline (i.e., the change in a cognitive phenotype across longitudinal survey waves), except for total word recall after age 80.

Finally, we used the 14 (out of 69) education-associated SNPs that are nominally significantly associated with cognitive performance ($p < .05$) to explore possible biological pathways in a set of bioinformatic analyses (see SI Appendix). Two of the 14 SNPs are in gene deserts, but the other 12 are in close vicinity to at least one gene predicted (based on its expression profile) to be involved in the nervous system (see SI Appendix). Among the most promising genes across these loci are KNCMA1, NRXN1, POU2F3, and SCRT, all of which are predicted to be involved in a glutamate neurotransmission pathway (labeled in REACTOME as “unblocking of NMDA receptor, glutamate binding, and activation”) that is involved in synaptic plasticity, a cellular mechanism for learning and memory. Using different methods (but some overlapping data), this same pathway has previously been implicated in human cognitive performance (21).

**Discussion:** This paper makes two contributions. First, we demonstrate that the “proxy-phenotype method” generates positive findings in a domain in which neither candidate-gene nor GWAS approaches have so far made substantial progress. Similar approaches have sometimes been used in prior work (e.g., to find rare structural variants associated with cognition; (22)), and there is existing work focused on the related idea of increasing statistical power in GWAS by analyzing correlated phenotypes jointly (23, 24).

We propose that the proxy-phenotype method, if systematically applied in social-science genetics, could be a useful complement to traditional gene discovery methods (such as GWAS) in cases where it affords greater statistical power. In the present case, it does so because (i) much larger genotyped samples are available for educational attainment than for cognitive performance, and (ii) some genetic variants are likely to be associated with educational attainment due to their more direct, stronger relationships with cognitive performance. For the same reasons, educational attainment might similarly serve as a proxy phenotype for personality traits such as persistence and self-control. In other contexts, the proxy-phenotype method may be better powered for different reasons. For example, for behavioral phenotypes with substantial measurement error—such as smoking, drinking, exercise, or eating habits—the proxy phenotype could be a medical outcome associated with the behavior (e.g., pulmonary disease for smoking, cirrhosis for alcohol consumption). We also note that, while our analysis plan specified that cohorts look up a relatively small set of education-associated SNPs in their existing GWAS results on cognitive performance, researchers with access to full GWAS results on the phenotype of interest could implement a more powerful version of the proxy-phenotype method. For example, first-stage results on the proxy phenotype could inform priors that are updated using GWAS results on the phenotype of interest.

We caution that the proxy-phenotype method (like theory-based candidate-SNP approaches) could generate an unacceptably high rate of false positives if it were applied when underpowered and if results were reported selectively. To avoid this, we propose a set of “best practices” that proxy-phenotype studies should follow: researchers should (a) conduct power calculations ex ante to justify the use of the method for a particular phenotype of interest, and report these
calculations in the SI; (b) circulate an analysis plan to all cohorts prior to conducting any analysis, and register the plan in a public repository; (c) commit to publishing all findings from the study, including null results; and (d) conduct Bayesian calculations of the credibility of any findings. We followed these procedures in this paper. While replication of findings in an independent cohort would be ideal, we anticipate that it will often be infeasible given the unavailability of genotyped samples that may motivate the proxy-phenotype approach in the first place.

The second contribution of this paper is to identify common genetic variants associated with cognitive phenotypes. Knowing the three significant SNPs is not useful for predicting any particular individual’s cognitive performance because the effect sizes are far too small, but it does enable follow-up research—e.g., pinpointing the causal variants and then conducting knock-out experiments in animals—that may ultimately shed light on biological pathways underlying cognitive variation. The polygenic scores constructed from our results may prove useful for studying gene-environment interactions. In future work, the magnitude of explained variance will increase as researchers gain access to datasets with even larger first-stage samples. Our results suggest that such scores hold promise for eventually identifying individuals whose cognitive health at older ages is at greatest risk, which could allow for appropriate preparation and (if possible) preventative intervention.

**Materials and Methods:** See SI Appendix for all details on the samples and methods.

**Acknowledgments:** This research was carried out under the auspices of the Social Science Genetics Association Consortium (SSGAC), a cooperative enterprise among medical researchers and social scientists that coordinates genetic association studies for social science variables. Data for our analyses come from many studies and organizations, some of which are subject to an MTA (see SI Appendix). Results from the meta-analysis, the complete biological annotation, and a FAQ document describing the findings of this paper are available at the website of the consortium, http://www.ssgac.org.

The formation of the SSGAC was made possible by an EAGER grant from the NSF and a supplemental grant from the NIH/OBSSR (SES-1064089). This research was funded in part by the Ragnar Söderberg Foundation (E9/11), the Swedish Research Council (412-2013-1061), and by the NIA/NIH through grants P01-AG005842, P01-AG005842-20S2, P30-AG012810, and T32-AG000186-23. For a full list of acknowledgments, see SI Appendix.
References:


Table 1. The SNPs significantly associated with cognitive performance after Bonferroni correction (for full results see Table S4). The chromosome and basepair position are from the NCBI genome annotation (build 36), and the nearest gene from the SCAN database. “Allele frequency” refers to the Cognitive Performance Sample.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Effective allele</th>
<th>Nearest gene</th>
<th>Year of Schooling (Education Sample)</th>
<th>Cognitive Performance (Cognitive Performance Sample)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Allele frequency</td>
<td>Standardized coefficient</td>
<td>P-value</td>
<td>Standardized coefficient</td>
</tr>
<tr>
<td>rs1487441</td>
<td>A</td>
<td>0.473</td>
<td>0.026</td>
<td>1.78×10⁻⁹</td>
</tr>
<tr>
<td>rs7923609</td>
<td>A</td>
<td>0.521</td>
<td>-0.021</td>
<td>1.06×10⁻⁶</td>
</tr>
<tr>
<td>rs2721173</td>
<td>T</td>
<td>0.473</td>
<td>-0.020</td>
<td>8.61×10⁻⁶</td>
</tr>
</tbody>
</table>
**Figure 1.** The relationship between standardized coefficients from the first-stage regression of years of schooling on the education-associated SNPs in the Education Sample (x-axis) and standardized coefficients from the second-stage regression of cognitive performance on these SNPs in the Cognitive Performance Sample (y-axis). The reference allele is chosen such that the coefficient on years of schooling is positive. Each point represents one of the 69 education-associated SNPs. (The cloud of points is bounded away from zero effect on years of schooling because the criterion for including a SNP was its reaching $p < 10^{-5}$ in the GWAS on years of schooling in the Education Sample.) Since the standard deviation of years of schooling is approximately 3, a coefficient of 0.03—a typical size for a years-of-schooling standardized coefficient—means that each reference allele is associated with an increase of $0.03 \times 3 \approx 0.09$ years of educational attainment. In conventional “IQ” units that have a standard deviation of 15, a standardized regression coefficient on cognitive performance of 0.03 corresponds to $\approx 0.45$ “IQ points.”
Figure 2. Q–Q plot for a regression of cognitive performance on the education-associated SNPs (the dark points) with 95% confidence interval around the null hypothesis (the darkly shaded region); and Q–Q plot for a regression of cognitive performance on the theory-based SNPs (the light points) with 95% confidence interval around the null hypothesis (the lightly shaded region). The table shows the nominal effect sizes and $p$-values for the three labeled SNPs, which are statistically significantly associated with cognitive performance after Bonferroni correction (for testing the 69 education-associated SNPs).
Figure 3. Coefficients from regression of standardized cognitive phenotype (Total Word Recall or Total Mental Status) on standardized polygenic score within age category, controlling for sex and clustering standard errors by individual (see SI Section 14 for details). Error bars show ±1 standard error.
Authors:

Affiliations:
1 Department of Applied Economics, Erasmus School of Economics, Erasmus University Rotterdam, 3000 DR Rotterdam, The Netherlands
2 Department of Epidemiology, Erasmus Medical Center, Rotterdam 3000 CA, The Netherlands
3 Boston Children's Hospital, Boston, Massachusetts 02115, United States of America
4 Broad Institute of the Massachusetts Institute of Technology and Harvard, Massachusetts 02142, United States of America
5 Harvard Medical School, Boston, Massachusetts 02115, United States of America
6 Estonian Genome Center, University of Tartu, Tartu 51010, Estonia
7 Centre for Cognitive Ageing and Cognitive Epidemiology, The University of Edinburgh, Edinburgh EH8 9JZ, Scotland, United Kingdom
8 Queensland Brain Institute, The University of Queensland, Brisbane, Queensland 4072, Australia
9 Department of Psychology, Union College, Schenectady, New York 12308, United States of America
10 Icelandic Heart Association, Kopavogur 201, Iceland
11 Faculty of Pharmaceutical Sciences, University of Iceland, 107 Reykjavik, Iceland
12 Framingham Heart Study, National Heart, Lung, and Blood Institute, Framingham, Massachusetts 01702, United States of America
13 Department of Psychology, Harvard University, Cambridge, Massachusetts 02138, United States of America
14 Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455-0344, United States of America
15 Department of Functional Genomics, VU University Amsterdam and VU Medical Center, 1081 HV Amsterdam, the Netherlands
39 Medical Research Council Centre for Causal Analyses in Translational Epidemiology, School of Social and Community Medicine, University of Bristol, Bristol BS8 2PR, United Kingdom

40 King’s College London, Medical Research Council Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, London SE5 8AF, United Kingdom.

41 Department of Psychology, The University of Edinburgh, Edinburgh EH8 9JZ, Scotland, United Kingdom

42 Department of Economics, Stockholm School of Economics, Stockholm 113 83, Sweden

43 University of Queensland Diamantina Institute, The University of Queensland, Princess Alexandra Hospital, Brisbane, Queensland 4102, Australia

44 Department of Economics, Cornell University, Ithaca, New York 14853, United States of America

45 Center for Experimental Social Science, Department of Economics, New York University, New York, New York 10012, United States of America

46 Faculty of Economics and Business, University of Amsterdam, Amsterdam 1018 TV, The Netherlands

* These authors contributed equally