OSF: Pre-registration

Preliminary Title: Too early to declare a general law of social mobility and heritability for education

Preliminary author list: Damien Morris

Introduction

I submitted a letter to the editor at PNAS responding to an article by Engzell and Tropf (2019).[1] In response to my draft letter the reviewers made several helpful comments and recommendations.[a]

In my original letter,[b] I had indicated that the regression of heritability estimates on parent-offspring correlations performed by the authors had not accounted for differences in precision between twin estimates from small and large samples. Reviewer 1 responded by saying that, "because the data to conduct the weighted regression are provided by Engzell and Tropf, the author should simply conduct the regression they are saying should have been fitted and compare the result with the original." I now plan to do this as part of the resubmission using inverse-variance precision weights described below in the methods section. I also plan to perform several other analyses, prompted by Reviewer suggestions, which are outlined further below.

Methods

Data

The data used for this analysis will be the same data used by Engzell and Tropf (2019)[1] to produce their main results in Figure 1B and Table 1, that is ACE estimates taken from the meta-analysis of twin studies by Branigan et al (2013)[2] and parent-offspring correlations and variances taken from the World Bank's Global Database on Intergenerational Mobility (GDIM).[3] I have obtained the specific data as published on the authors' OSF page (https://osf.io/c549j/).

Analyses

All analyses will be performed in the open-source statistical programming language R. I have applied the same protocols to merge the data outlined in the authors' Methods section to e.g. create smoothed variables, create destandardised ACE estimates, exclude non-representative samples, and set negative parameter estimates to zero, etc. I have already successfully reproduced the authors' main results and, in addition, successfully reproduced all of the supplementary results (from their Table S3) which I have attempted.

Having performed this, I will now proceed to fit a series of Weighted Least Squares (WLS) regression models to the same data, using an inverse variance weighting scheme. Inverse variance weights for heritability estimates will be calculated as the reciprocal of the sampling

---

[a] Email from Editor at PNAS with reviewers comments received on June 24th 2020

[b] Preprint published at PsyArXiv here: doi.org/10.31234/osf.io/vkn7m on June 3rd 2020

variance of each estimate (i.e. 1/*Var*(h^2)). To calculate the sampling variation in the Falconer's heritability estimates (*Var*(h^2)), we use a method devised by Branigan et al (2013)[2] which was algebraically derived from the method for deriving the sampling variances for correlations described in Borenstein et al (2009):[4]

$$Var(h^2) \approx 4\left(\frac{(1-r_{MZ}^2)^2}{n_{MZ}-1}+\frac{(1-r_{DZ}^2)^2}{n_{DZ}-1}\right)$$

Using these precision-weights, I will plot the WLS regression slope of (unscaled) heritability on (uncaled) parent-offspring correlation for direct comparison against the OLS regression slope in Figure 1B of Engzell and Tropf. Plots will be produced using the ggplot2 R-package. [5]

I will then perform a WLS multiple regression which includes gender as a covariate, and which scales heritability estimates and parent-offspring correlations so as to obtain standardised regression coefficients that can be compared directly with Engzell and Tropf's Table 1 results. Cluster-robust standard errors clustered by country and decade of birth will also be calculated for this purpose. These WLS multiple regression models will initially be produced using the lm.cluster() command from the miceadds R-package,[6] incorporating the "weights" argument. These clustered SEs will also be corroborated using the vcovCL() command from the sandwich package[7] after running WLS multiple regressions using the lm() command in base-R.

During my reproduction attempt, I found that lm.cluster() and vcovCL() returned the same parameter estimates, standard errors and t-values as the authors, but produced different p-values. However, I successfully obtained the same p-values once the t-values from lm.cluster() or vcovCL() were checked against t-tables for a two-sided t-test using the pt() command in base-R, using degrees of freedom equivalent to the number of clusters minus 1. I will therefore report these same manually calculated p-value for my WLS multiple regression models.

Supplementary analyses

I plan to perform additional analyses which I intend to publish as supplements on the OSF page associated with this pre-registration document. Space constraints in the letter and deadlines for resubmission make it unlikely that the following analyses can be included in my main submission.

I plan to extend the WLS analysis described above for estimated heritability (h^2) to estimated shared environmental influence (c^2) and non-shared environmental influence (e^2) as well. That is, I will visually compare WLS with OLS simple regression slopes for the unscaled variables, and also compare WLS with OLS standardised regression coefficients from Table 1 of Engzell and Tropf[1] in the scaled variables after controlling for gender and clustering standard errors.

The sampling variance of c^2 will be estimated using the following equation from Branigan et al:[2]

$$Var(c^2) \approx 4\left(\frac{(1-r_{DZ}^2)^2}{n_{DZ}-1}\right)+\frac{(1-r_{MZ}^2)^2}{n_{MZ}-1}.$$

Following Branigan et al, I will also assume that the sampling variance of e^2 is equivalent to the sampling variance of the monozygotic twin correlation (r_{MZ}), which can be estimated using the following formula for the variance of a correlation:

$$Var(r) \approx \frac{(1 - r^2)^2}{n - 1}$$

Analyses that were not pre-registered

As has previously been noted, some preliminary analyses have already been performed to ensure I could reproduce the author's original results in R before proceeding (the original authors published code for their analysis but it was in Stata). Additionally, before preparing this pre-registration I have already performed an additional piece of analysis on Reviewer 2's request.[c]

In my original letter, I had indicated that the parent-offspring correlations reported for Norwegian twins by Heath et al (1985)[8] were much higher than the values that Engzell and Tropf had used from the GDIM[3], and that this discrepancy reduces confidence in the headline association between heritability and social mobility. Reviewer 2 suggested that I should demonstrate how Engzell and Tropf's Table 1 results for heritability were affected when these alternative Norwegian parent-offspring correlations were used, and also suggested I plot how the regression slope was affected as compared to Figure 1 in Engzell and Tropf.[1]

I performed some preliminary analyses as proof of concept, providing Reviewer 2 with a revised OLS regression coefficient for heritability and a revised p-value (as generated by the lm.cluster model, but not manually adjusted after t-values were looked up with pt() command).[d] Revised standardised regression coefficients for c2 and e2 were also calculated at this time, as well as for $h^{2\prime}$, $c^{2\prime}$ and $e^{2\prime}$ the destandardised variance components. These analyses were already performed prior to submitting this pre-registration, but let the record show that this was only performed in response to Reviewer 2's specific request, with few researcher degrees of freedom available.

Only the revised, standardised h^2 results are likely to appear in the letter. I intend to publish the full Table 1 results with revised estimates for standardised and destandardised variance components on the OSF page associated with this pre-registration document (alongside the results for the pre-registered supplementary analyses already described above).

[c] In addition to their formal review, received on June 24th 2020, Reviewer 2 sent a private communication to make the same recommendation on 15th June 2020. I responded with provisional results on the 16th June 2020.
[d] I responded with provisional results on the 16th June 2020.

References

1. Engzell, P. & Tropf, F. C. Heritability of education rises with intergenerational mobility. *Proc. Natl. Acad. Sci.* **116**, 25386–25388 (2019).
2. Branigan, A. R., McCallum, K. J. & Freese, J. Variation in the Heritability of Educational Attainment: An International Meta-Analysis. *Soc. Forces* **92**, 109–140 (2013).
3. What is the Global Database on Intergenerational Mobility (GDIM)? *World Bank* https://www.worldbank.org/en/topic/poverty/brief/what-is-the-global-database-on-intergenerational-mobility-gdim.
4. Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. *Introduction to Meta-Analysis*. (John Wiley & Sons, 2011).
5. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*.
6. Robitzsch, A., Grund, S. & Henke, T. *miceadds: Some Additional Multiple Imputation Functions, Especially for 'mice'*. (2020).
7. Zeileis, A. Econometric Computing with HC and HAC Covariance Matrix Estimators. *J. Stat. Softw.* **11**, 1–17 (2004).
8. Heath, A. C. *et al.* Education policy and the heritability of educational attainment. *Nature* **314**, 734–736 (1985).