INSTITUTE for REPLICATION

No. 48
I4R DISCUSSION PAPER SERIES

A Comment on “Patience, Risk-Taking, and Human Capital Investment Across Countries” by Hanushek et al. (2021)

Alexandra de Gendre

Jan Feld

Nicolás Salamanca

July 2023

I4R Discussion Paper Series

I4R DP No. 48

A Comment on "Patience, Risk-Taking, and Human Capital Investment Across Countries" by Hanushek et al. (2021)

Alexandra de Gendre[1], Jan Feld[2], Nicolás Salamanca[3]

[1]University of Melbourne/Australia, Life Course Centre (LCC), and IZA Bonn/Germany
[2]University of Wellington/New Zealand, and IZA Bonn/Germany
[3]Melbourne Institute, University of Melbourne/Australia, Life Course Center (LCC), and IZA Bonn/Germany

July 2023

ISSN: 2752-1931

A comment on "Patience, risk-taking, and human capital investment across countries" by Hanushek et al. (2021)*

Alexandra de Gendre (University of Melbourne)
Jan Feld (Victoria University of Wellington)
Nicolás Salamanca (University of Melbourne)

[June 29, 2023]

Abstract

Hanushek et al. (2021) test how country-level measures of patience and risk-taking from the Global Preference Survey predict student performance on the Programme for International Student Assessment (PISA) math test. They find that country-level patience positively predicts math test scores and country-level risk-taking negatively predicts math test scores. They find similar results when holding country of residence characteristics constant and focusing on the preferences of the country of origin of migrants. We have checked the computational reproducibility and find that the data and analysis script provided by the authors allowed us to exactly reproduce the main tables in the paper. We also checked the robustness replicability by testing how robust the results are to decisions about imputation, weighting, operationalization of dependent variables, choice of control variables, and the inclusion of highly leveraged observations. We see that results are generally robust, though statistical significance of the risk-taking coefficient in the migrant analysis hinges on whether a control for OECD country of residence is included. Finally, we check the conceptual replicability of the results by using data from the Trends in International Mathematics and Science Study (TIMSS) instead of PISA – a different dataset with a different standardized test. This exercise shows that their results are robust to expanding the analysis to countries participating in both PISA and TIMSS.

* de Gendre: Department of Economics, The University of Melbourne, LCC and IZA, a.degendre@unimelb.edu.au; Feld: School of Economics and Finance, Victoria University of Wellington and IZA, jan.feld@vuw.ac.nz; Salamanca: Melbourne Institute: Applied Economics & Social Research, The University of Melbourne, LCC and IZA, n.salamanca@unimelb.edu.au. Corresponding author: Alexandra de Gendre. This comment is a result of the Replication Games held at Deakin University in Melbourne Australia. For more details, see https://i4replication.org/description.html. Stata do-files the analysis reported in this paper are available in the Open Science Foundation Repository at https://doi.org/10.17605/OSF.IO/KGT8Z

1. Introduction

Hanushek, Kinne, Lergetporer, and Woessmann (2021) use data from the Programme for International Student Assessment (PISA) and the Global Preference Survey (GPS, from Falk et al., 2018) to test how country-level measures of patience and risk-taking are correlated with students PISA test scores. They find that patience is positively related to test score and risk-taking is negatively related to test scores and show that these effects account for two thirds of the cross-country variation in student test scores. They address concerns about potential confoundedness in a second analysis that includes country of residence fixed effects and leverages the variation in country-level patience and risk-taking in the country of origin of migrants, finding similar results. Finally, they descriptively link more risk-taking to lower parental investments and "residual investments" (which combines unmeasured inputs and differences in the productivity of measured and unmeasured inputs). The paper makes an important contribution in exploring the potential for cultural factors to explain country-level differences in student skills. Prior to this study, the drivers of these country-level differences were poorly understood. Educational resources had previously been shown to be bad predictors of these differences, and more impactful factors, such as parental investments and engagement, still only provided a partial explanation since their structural determinants were still poorly understood. The paper makes progress on both these fronts by showing that patience and risk-taking can explain a large part of these country-level differences in student achievement, and by showing these preferences can be key proximate determinants of human capital investments that are key for skill formation.

In this comment, we probed the robustness of the main results in Hanushek et al. (2021). We test whether the material provided in the authors' replication package allows us to produce Tables 1-4 of the original paper and check for Tables 1-2 if the analysis in the Stata do-file matches the description of the analysis in the paper. We further test whether two of their key results are robust to sensible changes of the empirical specification. Finally, we test whether the main results change if we use test score data from The Trends in International Mathematics and Science Study (TIMSS) instead of PISA.

Overall, we find the results of Hanushek et al. (2021) are very robust. The replication package allowed us to produce Tables 1-4 and the numbers in these tables matched those from the tables in the published paper. For Tables 1-2 we did not find any meaningful differences between the analysis in the Stata do-file and the description of the analysis in the paper. The results are also largely robust to changes in empirical specifications. Only one of the tested coefficients was meaningfully affected by whether the regression included one control variable. In specifications that do not include this control variable, point estimates remain qualitatively similar but are no longer statistically significant at conventional levels. Results are also robust to using math test scores from TIMSS instead of PISA.

2. Reproducibility

Hanushek et al. (2021) and provide a replication package including raw original data, analysis files, and working datasets (available at https://doi.org/10.3886/E153101V2). We checked the reproducibility of their analysis in two ways. First, we re-ran all Stata do-files to test whether we can re-create Tables 1-4 from the provided raw data. Second, we checked whether these tables are identical to Tables 1-4 reported in the published paper. While the reproduced tables may be identical, the analysis may differ from the description in the paper. For example, the paper may describe one way of clustering the standard errors while the actual analysis is using a different way. We therefore checked whether the analysis in Table 1-2 is identical to the analysis described in the published paper.

Overall, we identified no issues of reproducibility. We were able to reproduce Tables 1-4 from the provided raw data. These tables were - except for some minor formatting differences- identical to the ones shown in the published paper.

For Tables 1-2 we also found no meaningful difference between the Stata do-files and the description of the analysis in the paper. However, we did notice three imprecisions in the description of the analysis.

OECD dummy. The notes of Table 2 mentions that regressions include a dummy for OECD country of origin. However, this control is not discussed in the description of the empirical model in Section 3.1.

Means of Item non-response. When reading the published paper, it appears that the specification shown in Table 1 Column (8) and Table 2 Column (7) include means of item non-response at the country level. However, these appear to be country-by-wave means of item non-response.

Migration status. The description of the empirical model in Section 2.1 as well as the notes of Tables 1 describe the models include controls for student gender, age, and migration status. From checking the do-file, we can see the regressions include two variables for migration status (migra1 and migra2). However, these variables are not labelled. After reading the code, it became clear that these are child-level flags for whether the child is a first- or second-generation migrant.

3. Replication

In this section, we check the robustness of two main findings in Hanushek et al. (2021). We broadly decided on our approach to the robustness checks and replication exercise (e.g., which main tables to focus on, which decisions to look into carefully)before looking at the program of the original authors, but we made specific decisions (e.g., to consider the robustness of findings to the inclusion or exclusion of the OECD dummy variable) after looking at the program. We sent our board plan of analysis to the organizers of the replication games that gave rise to this comment but did not pre-register our robustness checks.

3.1 Probing the robustness of key finding #1

Key finding #1: Country-level patience positively predicts students' math test scores and country-level risk-taking negatively predicts students' math test scores.

This finding is taken from Table 1, Column (3) on page 2296. The patience coefficient is 1.226 standard deviations (SD) with a standard error (SE) of 0.132 SD; the risk-taking coefficient is -1.241 SD (SE 0.184 SD). These estimates shows that a 1 SD increase in country-level patience is associated with a 1.226 SD increase in PISA math test scores and a 1 SD increase in country-level risk-taking is associated with a 1.241 SD decrease in math test scores.

The dependent variable of the specification reported in the paper is the first plausible value for students' math ability. The PISA team use Item Response Theory modelling to estimate students' latent ability in Mathematics based on students' answers to the math-component of the

PISA test. The PISA data contain 10 different plausible values for each students' math test score, any one of which should give a good approximation of a child's ability. The main explanatory variables of interest are country level measures of patience and risk-taking as measured with the GPS. Control variables include a female dummy, students' age in years, two dummy variables describing students' migration status, three dummy variables indicating whether the student age, student gender, or student migration status was imputed, as well as fixed effects for the year of the wave of the PISA test. The regression uses sampling weights. More specifically, "*All regressions are weighted by students' sampling probabilities within countries and give equal weight to each country*" (Hanushek et al., 2021, page 2295). Standard errors are robust to clustering at the country level.

While the authors' specification seems reasonable to us, we believe other researchers might have chosen similarly defensible specifications. We therefore test how robust the results are to other reasonable specifications. More specifically, we identified four decisions of the authors where reasonable alternatives are possible.

Weights –We checked whether the results depend on this weighting scheme.

Imputations – The original regression includes imputed values for students' migration status, age, and gender. We test whether excluding all imputed values affects the analysis. Excluding imputed values reduces the sample size by 2.5% (from 1,992,276 to 1,942,635).

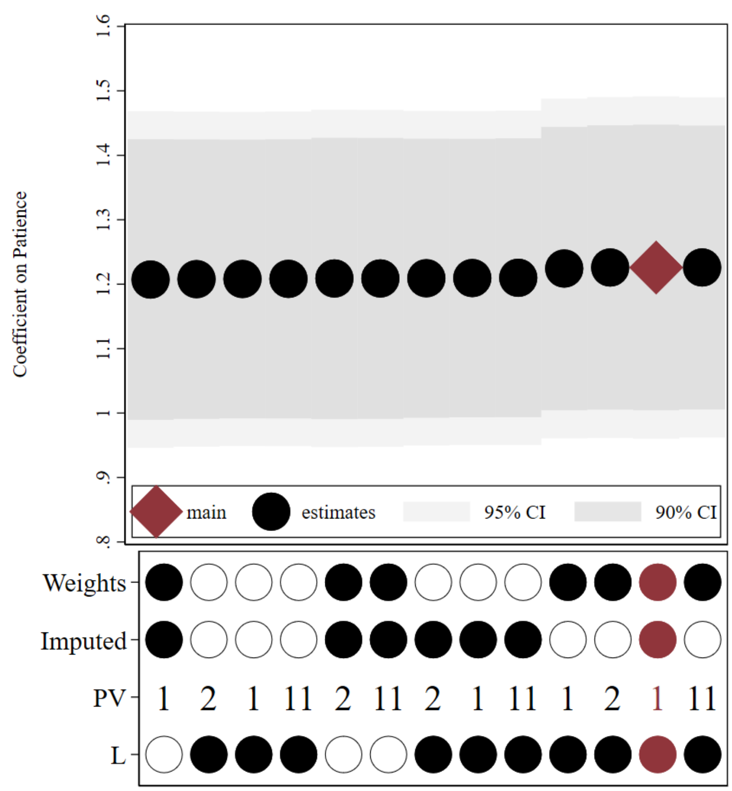
Plausible values –The authors performed their analysis using the first provided plausible value as dependent variable. We test whether the results are robust to using the second provided plausible value or the average of all 10 provided plausible values.

High-leverage observations – We test whether results are robust to excluding the 1% of observations with highest leverage. High leverage observations are observations with extreme or outlier values of the independent variables, which have the potential to heavily influence a regression fit. We identify the leverage of observation i, h_i, as the corresponding diagonal element in the projection or hat matrix, such that $h_i = x_i'(X'X)^{-1}x_i'$.

To be able to identify any influential methodological choices, we estimate specifications with different combinations of these decisions (e.g., without weights but with imputed values, without weights and without imputed values). We provide Stata do-files for these robustness checks in the Open Science Foundation Repository available at https://doi.org/10.17605/OSF.IO/KGT8Z. For brevity, we show the key point estimates using a specification curves figures and more details on each estimate (e.g., standard error, p-value) in Table A1 in the appendix.

Figure 1 shows the patience coefficient is very stable across all different specifications. The point estimates range from 1.207 SD to 1.226 SD; all point estimates are statistically significant at the 0.1% level.

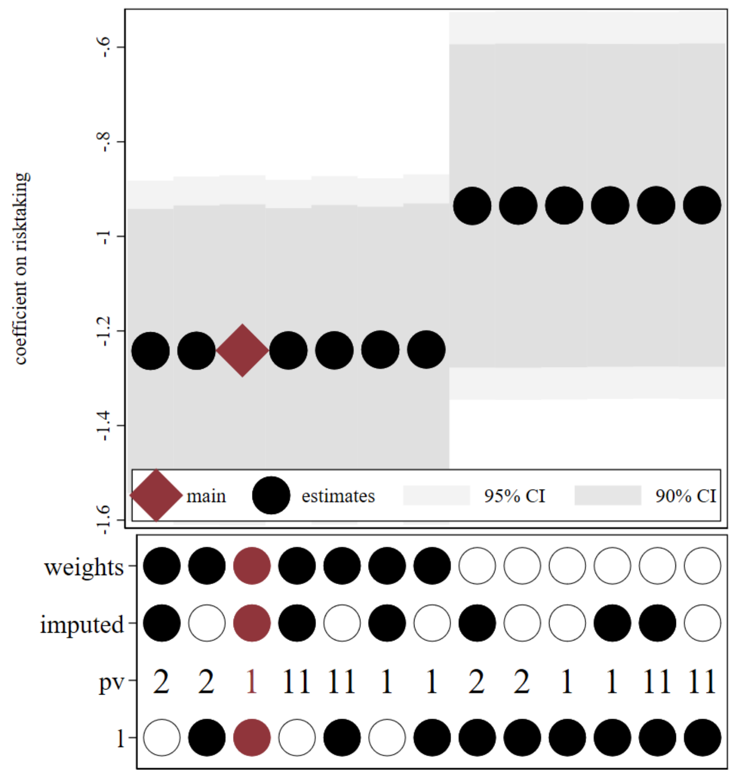
Figure 1: Specification curve for patience coefficient (Main result #1)



Note: This figure shows the estimated patience coefficients and their 90% and 95% confidence intervals from the main specification (red diamond) and 12 robustness specifications (black circles). The bars below the panel show details about the specifications. Weight indicates whether the robustness specification used weights (black circle = yes, white circle = no), imputed indicates whether the regression included imputed observations, pv indicates which plausible value was used as the dependent variable (1 = plausible value 1, 2 = plausible value 2, 11 = average of all 10 plausible values provided in the data), l indicates whether the regression included the 1% highest leveraged observations. For more details, see Table A1 in the appendix.

Figure 2 shows the risk-taking coefficient is negative and statistically significant in all specifications. The point estimates range from -0.934 to -1.241 and all coefficients are statistically significant at the 0.1% level (see Table A1). While we see that point estimates from specifications without weighting are less negative throughout, this empirical choice does qualitatively affect the results.

Figure 2: Specification curve for risk-taking coefficient (Main result #1)



Note: This figure shows the estimated risk-taking coefficients and their 90% and 95% confidence intervals from the main specification (red diamond) and 11 robustness specifications (black circles). The bars below the panel show details about the specifications. Weight indicates whether the robustness specification used weights (black circle = yes, white circle = no), imputed indicates whether the regression included imputed observations, pv indicates which plausible value was used as the dependent variable (1 = plausible value 1, 2 = plausible value 2, 11 = average of all 10 plausible values provided in the data), l indicates whether the regression included the 1% highest leveraged observations. For more details, see Table A1 in the appendix.

3.2 Probing the robustness of key finding #2

Key finding #2: Holding host country characteristics constant, migrant students from countries in which people are more patient score better on standardized math tests and migrants from countries in which people are more risk-taking score worse on standardized math tests.

This finding is taken from Table 2, Column (3) on page 2300. The patience coefficient is 0.931 SD (SE 0.116 SD) and the risk-taking coefficient is -0.294 SD (SE0.122). These estimates suggest that being a migrant from a country in which people are 1 SD more patient is associated with a 0.931 SD increase in students' math test scores and being from a country in which people are 1 SD more risk-taking is associated with a 0.294 SD decrease in students' math test scores.

The dependent variable in the published paper is the first plausible value for students' math score. The main independent variables of interest are the country averages of patience and risk-taking of the migrants' country of origin. Control variables include a female student dummy, students' age in years, two dummy variables indicating whether student gender or age was imputed, a dummy variable indicating whether migrants' country of origin is part of the OECD, as well as country-wave fixed effects. Standard errors are clustered at the migrants' country of origin.

Again, we find the authors' decisions seem reasonable, but we believe that there are other reasonable specifications. We identified four decisions of the authors where reasonable alternatives are possible.

Imputations – The original regression includes imputed values for students' gender. We estimate specifications with and without imputed observations. Only 7 of 80,398 observations include imputed values.[1]

Plausible values – The authors performed their analysis using the first provided plausible value as dependent variable. We test whether the results are robust to using the second provided plausible value or the average of all 10 provided plausible values.

OECD dummy – The specification includes a control variable for whether migrants country of origin is part of the OECD or not. We test whether including it matters for the results.

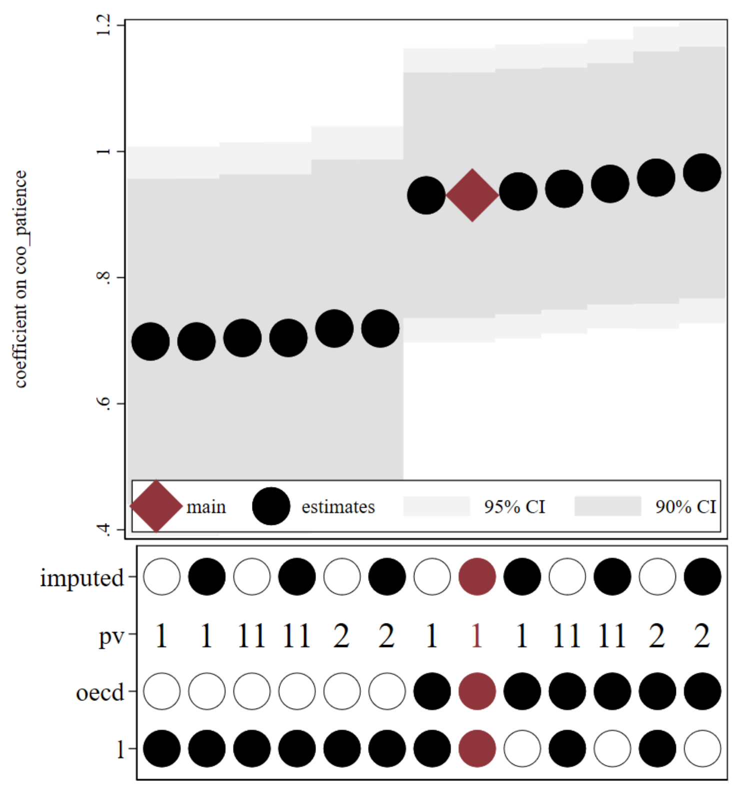
High-leverage observations – We test whether results are robust to excluding the 1% of observations with highest leverage. High leverage observations are observations with

[1] We only checked how many observations were imputed when writing the report. If we would have known beforehand that only 7 observations are imputed, we would not have done this robustness check. However, we have decided to show the results for completeness.

extreme or outlier values of the independent variables, which have the potential to heavily influence a regression fit. We again identify the leverage of observation i as $h_i = x_i'(X'X)^{-1}x_i'$.

Figure 3 shows that the patience coefficients are very similar across all 13 specifications. The point estimates range from 0.699 SD to 0.967 SD and all coefficients are statistically significant at the 0.1% level (see Table A2). Excluding the OECD dummy leads to smaller coefficients. However, this change does not affect the qualitative conclusion of these estimates.

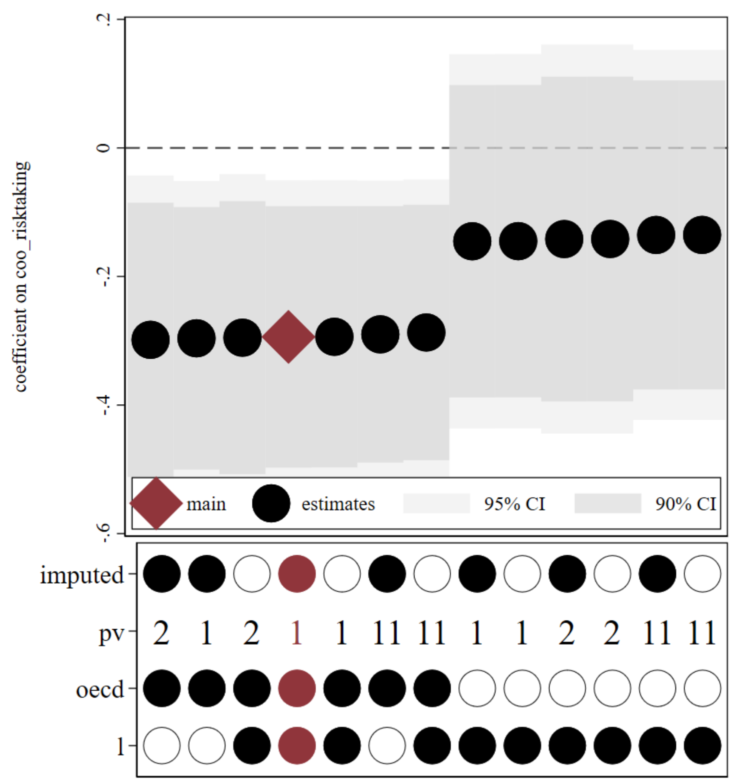
Figure 3: Specification curve for patience coefficient (Main result #2)



Note: This figure shows the estimated patience coefficients and their 90% and 95% confidence intervals from the main specification (red diamond) and 12 robustness specifications (black circles). The bars below the panel show details about the specifications. Imputed indicates whether the regression included imputed observations (black circle = yes, white circle = no), pv indicates which plausible value was used as the dependent variable (1 = plausible value 1, 2 = plausible value 2, 11 = average of all 10 plausible values provided in the data), oecd indicates whether the specification included a dummy variable indicating that the migrants' country of origin is part of the OECD, l indicates whether the regression included the 1% highest leveraged observations. For more details, see Table A2 in the appendix.

Figure 4 shows that the risk-taking coefficients have the same sign in all 13 specifications. The coefficients range from -0.135 SD to -0.299 SD (see Table A2). However, the magnitude of the coefficients and statistical significance depends on the inclusion of the OECD dummy. With this dummy the point estimates range from -0.287 SD to -0.299 SD and all estimates are statistically significant at the 5% level. Without this dummy, point estimates range from -0.135 SD to -0.299 SD and none are significant even at the 10% level.

Figure 4: Specification curve for risk-taking coefficient (Main result #2)



Note: This figure shows the estimated risk-taking coefficients and their 90% and 95% confidence intervals from the main specification (red diamond) and 12 robustness specifications (black circles). The bars below the panel show details about the specifications. Imputed indicates whether the regression included imputed observations (black circle = yes, white circle = no), pv indicates which plausible value was used as the dependent variable (1 = plausible value 1, 2 = plausible value 2, 11 = average of all 10 plausible values provided in the data), oecd indicates whether the specification included a dummy variable indicating that the migrants' country of origin is part of the OECD, l indicates whether the regression included the 1% highest leveraged observations. For more details, see Table A2 in the appendix.

4. Conceptual replication

Finally, we check the conceptual replicability of the results by using data from TIMSS instead of PISA. TIMSS data differ from PISA data in a few important ways. Firstly, while PISA samples 15 year old students irrespective of their grade-level within schools of participating countries, TIMSS samples Grade 4 and Grade 8 pupils irrespective of their age or repeater status, within schools. Secondly, PISA is an assessment programme run every three years by the OECD since 2000, while TIMSS runs every four years since 1995. Lastly, PISA focuses on measuring cognitive and problem-solving skills using applications in numeracy and literacy, whereas TIMSS focuses on knowledge of concepts in Math and Science and how that knowledge is used by students. Therefore, this conceptual replication answers the question of whether the main findings replicate in a sample of younger students, and with a test that measures mathematical concepts rather than problem-solving skills.

For this conceptual replication task, we focused on Table 1 from Hanushek et al. (2021). We obtained publicly-available TIMSS original data files through the TIMSS international database (https://timssandpirls.bc.edu/databases-landing.html). The files for waves 1995 and 1999 are provided in .DAT format and require the user to build a data dictionary to convert those files into a format fit for analysis. The files for waves 2003, 2007, 2011, 2015 and 2019 are provided in SPSS or SAS format. We use Stata for our analyses, and therefore convert all files to Stata .DTA format. We provide our own code to perform those steps, prepare the data and produce Table 1 following Hanushek et al. (2021).

Hanushek et al. (2021)'s analyses are based on 49 countries participating in the PISA assessment for which GPS data are also available. There are 50 countries participating in TIMSS for which GPS data are also available; of those countries, 44 overlap with the countries used in Hanushek et al. (2021), 6 countries are only in TIMSS (Botswana, Egypt, Ghana, Iran, Pakistan and South Africa) and 5 are only in PISA (Brazil, Costa Rica, Mexico, Peru, and Vietnam). The list of those countries is presented in Appendix.

We focus on replicating the authors' preferred estimate, presented in col. 3, Table 1 in Hanushek et al. (2021). Their preferred specification regresses the first plausible value in Math (standardized) on the country-level means of patience and risk-taking, controlling linearly for female status, age, first-generation migration status, second-generation migration status, and dummies for missing values in age, female status, and first-generation immigration status. In addition, the authors add fixed effects for survey years, account for survey weights and cluster standard errors at the country level. Our own specifications are identical to their benchmark specification using TIMSS data. Table 1 presents our findings for all countries available in TIMSS, for countries that overlap with PISA and for countries that do not overlap with PISA.

Table 1: Conceptual replication of Hanushek et al. (2021, Table 1, Col. 3) using TIMSS.

	Benchmark estimate	TIMSS (all 50 countries)	TIMSS (44 overlapping countries)	TIMSS (6 non-overlapping countries)
	(1)	(2)	(3)	(4)
Patience:				
Coef. Est.	**1.225926**	.953727	.8715074	-1.635801
Sd. Err.	**.1320491**	.1905004	.2148762	.8155331
p-value	**0.000**	0.000	0.000	0.101
Risk-Taking:				
Coef. Est.	**-1.241304**	-1.337247	-1.298366	.4722979
Sd. Err.	**.1841879**	.2378276	.4083088	.2386361
p-value	**0.000**	0.000	0.003	0.105

Note: The benchmark estimate highlighted in **bold** is the preferred estimate as indicated by the authors of Hanushek et al. (2021); this estimate is in col. 3, Table 1. This estimate corresponds to a regression of the standardized first plausible value in Math on the country-level means of patience and risk-taking, controlling linearly for female status, age, first-generation migration status, second-generation migration status, and dummies for missing values in age, female status, and first-generation immigration status. In addition, the authors add fixed-effects for survey years, account for survey weights and cluster standard errors at the country level. Our own specifications are identical to their benchmark specification using TIMSS data.

Our conceptual replication exercise indicates that the original findings by Hanushek et al. (2021) are broadly replicable. Cols (2) and (3) in Table 1 are of same sign, similar magnitude and statistical significance as the benchmark estimate from the original study. Estimates do differ in three important ways in the subsample of 6 countries that are only available in TIMSS and not in PISA (Col 4). In this subsample of countries 1) we find an opposite sign for both patience and risk-taking coefficients; 2) the coefficient on risk-taking is of less than half the magnitude of the benchmark estimate; and 3) our coefficients are not statistically significant. We do not view this as evidence that the paper is not replicable since these are based on a much smaller sample and only based on data from 6 countries, whereas the broader point of the paper is made with many more countries for which the main results seem to be very robust. One could also argue that the difference between our estimates and the benchmark estimates could stem from the large cultural differences between the countries that are included in each analysis; the PISA-only countries included in the benchmark estimate are almost all from Latin America, whereas the TIMSS-only countries included in Col. (4) are all African or Middle-Eastern countries. However, without more data it is difficult to disentangle country sample differences in culture from sampling error.

4. Conclusion

Overall, our work reaffirms the main results of Hanushek et al. (2021) in terms of the computational reproducibility of their findings, the robustness of their main results to different specification choices, and the replicability of their main findings using a completely different international student assessment dataset. Our findings lend confidence to the scientific quality of the paper and the likelihood that the results reflect true relationships between patience and risk-taking and students' achievement across the world. To further strengthen the scrutiny of this paper, future replicators could focus on three additional exercises:

- Use Bayesian shrinkage methods to produce Best Linear Unbiased Predictors (BLUPs) of country-level patience and risk-taking preferences. BLUPs would account for the fact that preferences are more accurately measured for some countries than others due to sampling error.
- Check for potential non-linear dependences between patience and risk-taking. Hanushek et al. (2021) do check for potential interactions between these two preferences (pg. 2295, footnote 4) but there is always the possibility of non-linear dependencies, which could change the interpretation o the main findings in important ways.
- Extend the migrant analysis by providing conditional rather than unconditional expectations as country-of-origin preferences (e.g., by imputing the patience of a Tunisian higher educated female migrant as the average patience of higher educated Tunisian women). This would improve identification since, as the authors remark, "[i]dentification in the migrant analysis depends on the extent to which the national preferences of the country of origin provide a good proxy for the students' and families' actual preferences." (p. 2301). Even better would be to construct specific preference profiles of potential migrants based on observable characteristics, though this might be unfeasible even with the GPS data.

References

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, *133*(4), 1645-1692.

Hanushek, E. A., Kinne, L., Lergetporer, P., & Woessmann, L. (2022). Patience, risk-taking, and human capital investment across countries. *The Economic Journal*, *132*(646), 2290-2307.

Appendix

Table A1: More details on coefficients shown in Figures 1-2

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Specification				patience			risk-taking	
weights	imp.	pv	lev	coef.	se	p-value	coef.	se	p-value
1	1	1	0	1.207	0.130	0.0000	-1.239	0.180	0.0000
0	0	2	1	1.208	0.129	0.0000	-0.935	0.204	0.0000
0	0	1	1	1.208	0.129	0.0000	-0.935	0.204	0.0000
0	0	11	1	1.208	0.129	0.0000	-0.934	0.204	0.0000
1	1	2	0	1.209	0.130	0.0000	-1.242	0.179	0.0000
1	1	11	0	1.209	0.130	0.0000	-1.241	0.179	0.0000
0	1	2	1	1.209	0.129	0.0000	-0.936	0.204	0.0000
0	1	1	1	1.210	0.129	0.0000	-0.935	0.204	0.0000
0	1	11	1	1.210	0.129	0.0000	-0.934	0.203	0.0000
1	0	1	1	1.224	0.131	0.0000	-1.239	0.184	0.0000
1	0	2	1	1.226	0.131	0.0000	-1.242	0.183	0.0000
1	**1**	**1**	**1**	**1.226**	**0.132**	**0.0000**	**-1.241**	**0.184**	**0.0000**
1	0	11	1	1.226	0.131	0.0000	-1.241	0.183	0.0000

Note: Specification from main paper highlighted in **bold**.

Table A2: More details on coefficients shown in Figures 3-4

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					patience			risk-taking	
imp.	pv	oecd	lev.	coef.	se	p-value	coef.	se	p-value
0	1	0	1	0.699	0.154	0.0000	-0.145	0.145	0.3224
1	1	0	1	0.699	0.154	0.0000	-0.145	0.145	0.3220
0	11	0	1	0.704	0.155	0.0000	-0.135	0.144	0.3504
1	11	0	1	0.704	0.155	0.0000	-0.135	0.144	0.3501
0	2	0	1	0.719	0.160	0.0000	-0.142	0.151	0.3522
1	2	0	1	0.719	0.160	0.0000	-0.142	0.151	0.3520
0	1	1	1	0.930	0.116	0.0000	-0.294	0.122	0.0189
1	**1**	**1**	**1**	**0.931**	**0.116**	**0.0000**	**-0.294**	**0.122**	**0.0188**
1	1	1	0	0.937	0.116	0.0000	-0.296	0.122	0.0185
0	11	1	1	0.941	0.115	0.0000	-0.287	0.119	0.0189
1	11	1	0	0.949	0.114	0.0000	-0.290	0.119	0.0183
0	2	1	1	0.959	0.120	0.0000	-0.295	0.127	0.0237
1	2	1	0	0.967	0.119	0.0000	-0.299	0.127	0.0229

Note: Specification from main paper highlighted in **bold**.

Table A3: Countries used in Hanushek et al. (2022, Table 1) and in own conceptual replication using TIMSS.

Country name	PISA	TIMSS
Algeria	1	1
Argentina	1	1
Australia	1	1
Austria	1	1
Bosnia and Herzegovina	1	1
Botswana	0	1
Brazil	1	0
Canada	1	1
Chile	1	1
Colombia	1	1
Costa Rica	1	0
Croatia	1	1
Czech Republic	1	1
Egypt	0	1
Estonia	1	1
Finland	1	1
France	1	1
Georgia	1	1
Germany	1	1
Ghana	0	1
Greece	1	1
Hungary	1	1
Indonesia	1	1
Iran	0	1
Israel	1	1
Italy	1	1
Japan	1	1
Jordan	1	1
Kazakhstan	1	1
South Korea	1	1
Lithuania	1	1
Mexico	1	0
Moldova	1	1
Morocco	1	1
Netherlands	1	1
Pakistan	0	1
Peru	1	0
Philippines	1	1
Poland	1	1
Portugal	1	1
Romania	1	1
Russian Federation	1	1
Saudi Arabia	1	1
Serbia	1	1
South Africa	0	1
Spain	1	1
Sweden	1	1
Switzerland	1	1
Thailand	1	1

Country name	**PISA**	**TIMSS**
Turkey	1	1
Ukraine	1	1
United Arab Emirates	1	1
United Kingdom (England, Scotland, Nothern Ireland)	1	1
United States	1	1
Vietnam	1	0
Total countries	*49*	*50*