Re-examining the relationship between patience, risk-taking, and human capital investment across countries*

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

October 12, 2023

Abstract

Hanushek et al. (2022) show that students in countries in which people are more patient and less risk-taking perform better in the Programme for International Student Assessment (PISA) test. In this paper, we probe the robustness of this study. Our narrow replication shows that the results are mostly robust to alternative model specifications. Our broad replication shows that the main results are robust to measuring student performance with data from The Trends in International Mathematics and Science Study (TIMSS) and The Progress in International Reading Literacy Study (PIRLS) instead of PISA.

Keywords: economic preferences, human capital, replication

* 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 paper is a result of the Replication Games held at Deakin University in Melbourne Australia. For more details, see https://i4replication.org/description.html. An earlier version of this paper called "*A Comment on" Patience, Risk-Taking, and Human Capital Investment Across Countries" by Hanushek et al.(2021)*" is available at http://hdl.handle.net/10419/273433 (de Gendre, Feld, and Salamanca 2023). 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 et al. (2022) use data from the Global Preference Survey (GPS, see: Falk et al. 2018; Falk et al. 2016) and the Programme for International Student Assessment (PISA) and to test how country-level measures of patience and risk-taking are correlated with students PISA test scores. They find that patience is positively correlated with test score and risk-taking is negatively correlated with test scores and show that these effects account for two-thirds of the cross-country variation in student test achievement. 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 migrant students. This analysis yields similar results. Finally, they use country-level aggregate data to descriptively link more patience and less risk-taking to higher parental investments and "residual investments" (which combines unmeasured inputs and differences in the productivity of measured and unmeasured inputs). They further link more patience to higher school inputs.

Before Hanushek et al. (2022) we knew that a large share of the between-country variation in PISA test-scores can be explained by student characteristics, family backgrounds, home inputs, resources, teachers, and institutions (Fuchs and Wößmann 2008). However, we knew little about the deeper structural determinants of between-country differences in student performance. Hanushek et al. (2022) make progress on this front by a 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.

However, findings from one individual study should be taken with a grain of salt because many studies do not replicate (Camerer et al. 2016; Camerer et al. 2018; Open Science Collaboration 2015). Studies can fail to replicate because the original results were wrong or because of p-hacking

(Brodeur et al. 2016; Brodeur et al. 2023). While we have no particular reason to distrust Hanushek et al. (2022)'s motives, it is possible that they have also been influenced, maybe subconsciously, by the goal of finding statistically significant and publishable results. Even if results are correct for the specific context of the study, it is not clear whether they hold more broadly.[1] For example, Hanushek et al. (2022) only use one data source to measure student achievement. If their findings hold more broadly, we should see similar results with other measures of student achievement. It is therefore up to the scientific community to probe the robustness of Hanushek et al. (2022)'s findings. This is what we are doing in this paper.

We conduct a narrow replication by testing whether Hanushek et al. (2022)'s results are robust to decisions about imputation, weighting, operationalization of dependent variables, choice of control variables, and the inclusion of high-leverage observations. We also conduct a wide replication by testing whether the main results change if we use test score data from The Trends in International Mathematics and Science Study (TIMSS) and The Progress in International Reading Literacy Study (PIRLS) instead of PISA, across Grade 4 and Grade 8 students, and across each wave of TIMSS (1995-2019) and PIRLS (2001-2021).

Overall, we find that the results of Hanushek et al. (2022) are robust. Our narrow replication shows that results are largely robust to changes in empirical specifications. Only one of the tested coefficients of interest was meaningfully affected by a change to the empirical specification. In the migrant analysis, the estimated effect of risk-taking remains qualitatively similar but is no longer statistically significant when we exclude controls for whether the student's country of origin is part of the OECD. All other empirical decisions do not meaningfully affect the statistical

[1] See Eronen and Bringmann (2021) for a discussion on the importance of identifying relatively constant and stable empirical facts ("phenomena") and for an example of how to empirically estimate if an empirical fact can be considered universal or near-universal (de Gendre et al. 2023).

significance or magnitude of the coefficients of interest. Our wide replication shows that results are robust to using student achievement data from TIMSS and PIRLS instead of PISA. The results also hold for different subjects, grade-levels, and waves.

2 Data used in Hanushek et al. (2022)

Hanushek et al. (2022) use PISA data to measure student achievement. PISA is a study which aims at creating internationally comparable measures of student achievement in math, science, and reading using random samples of 15-year-old students in several countries. In their main analysis, Hanushek et al. (2022) use data from seven PISA waves (from 2000 to 2018) covering 49 countries.

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 student's math test score, any one of which should give a good approximation of a student's math ability. In their main analysis, Hanushek et al. (2022) use the first of these plausible values for student's math ability as a measure of student achievement. The plausible values are scaled to approximate a normal distribution with a mean of 500 points and a standard deviation of 100 points (OECD 2019). Hanushek et al. (2022) divide the plausible values by 100 so that coefficients can be interpreted in terms of standard deviations.

Hanushek et al. (2022) use GPS data to measure patience and risk-taking. The GPS measures each of these preferences at the individual level with a combination of one qualitative survey question and one hypothetical choice question. The answers to both questions are then combined to a single preference measure using weights from a validation procedure (Falk et al. 2018; Falk et al. 2016). In their analysis, the main explanatory variables are the country-averages of patience and risk-taking. Both measures are re-standardized to have a mean of zero and standard

deviation of 1 in the analysis sample. Hanushek et al. (2022) provide a replication package including raw original data, analysis files, and working datasets available at https://doi.org/10.3886/E153101V2.

3. Narrow replication: robustness to different empirical specification choices

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. The results from Column 3 of Table 1 from Hanushek et al. (2022) suggest that 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. Both coefficients of interest are statistically significant at the 1% level.

While the authors' specification seems reasonable to us, we believe other researchers could have chosen similarly defensible specifications (see Appendix B for more details on the original empirical specification). We therefore test how robust the results are to other reasonable specifications.[2] We identified five areas where reasonable alternative decisions would have been possible.

Weights – The original 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., 2022, page 2295). We checked whether the results depend on the weighting scheme by estimating specifications without weights.

[2] We also tested whether the material provided in the authors' replication package allows us to produce Tables 1-4 of the original paper. This exercise revealed no issues of computational 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.

Controls for student background – The original regression includes controls for being a first-generation immigrant and a second-generation immigrant. First-generation immigrants are defined as foreign-born students with two foreign-born parents. Second-generation immigrant students are students who were born in the country where they sat the PISA test and have two foreign-born parents (OECD 2023). Another way to account for students' background is to control for whether they are foreign born. Foreign-born students would not be classified as immigrants if at least one of their parents is born in the country of the test. We test how controlling for students' background affects the results by estimating specifications which additionally include a foreign-born dummy or include a foreign-born dummy instead of the first and second-generation migrant dummies. Because the foreign-born dummy variable includes imputed values, we also include a dummy variable flagging imputed values for this variable whenever we include the foreign-born dummy.

Imputations – The original regression includes imputed values for first-generation immigrant status, second-generation immigrant status, age, and gender.[3] Excluding all imputed values reduces the sample size by 3.4% (from 1,992,276 to 1,925,530). The original regression also includes dummy variables flagging whether a value has been imputed for age, gender, and first-generation immigrant status. However, the original regression does not include a dummy variable flagging imputed values for second-generation immigrant status despite this variable also containing imputed values. We asked the authors of the original study about this decision and they told us that including this imputation dummy would have been more precise. To test whether decisions about imputations affect the results, we estimate specifications that additionally include a dummy

[3] Hanushek et al. (2022) impute missing values in these variables at the individual level by replacing them with their country-by-waive weighted means. We closely follow their imputation process in our wide replication exercise (Section 4).

variable flagging imputed values for second-generation immigrant status, and specifications without any imputed values.

Plausible values –The authors performed their analysis using the first provided plausible value as the 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 –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 $h_i = x_i'(X'X)^{-1}x_i'$. We test whether results are robust to excluding the 1% of observations with the highest leverage.

To be able to identify any influential methodological choices, we estimate specifications with different combinations of these decisions (e.g., without weights and plausible value 1, without weights and plausible value 2). Stata do-files for these robustness checks are available in the Open Science Foundation repository at https://doi.org/10.17605/OSF.IO/KGT8Z.

Table 1 shows that the patience coefficient is very stable across all different specifications. The point estimates range from 1.209 SD to 1.232 SD; all point estimates are statistically significant at the 1% level. Table 1 further shows that 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 1% level. While we see that point estimates from specifications without weighting are less negative throughout, this empirical choice does not qualitatively affect the results. Overall, we find that *Key finding #1* is robust to several alternative empirical decisions.

Table 1. Robustness of the main findings in Table 1 in Hanushek et al. (2022) to alternative model specification choices

Outcome: Standardized PISA test scores in math																	
					Specification:												
								Controls for:				Incl. imputed flags & values for:					
Patience coef.	std. err.	Risk-taking coef.	std. err.	Obs.	Plausible value	Uses weights	Incl. top 1% leverage obs.	Foreign-born	1st & 2nd gen. migrant parents			Foreign-born	1st gen. migrant parents	2nd gen. migrant parents		Female & age
1.209***	(0.129)	-0.936***	(0.204)	1,992,276	#2		✓		✓	✓			✓			✓
1.209***	(0.129)	-1.235***	(0.182)	1,972,342	#1	✓			✓	✓			✓			✓
1.210***	(0.129)	-0.935***	(0.204)	1,992,276	#1		✓		✓	✓			✓			✓
1.210***	(0.129)	-0.934***	(0.203)	1,992,276	Avg.		✓		✓	✓			✓			✓
1.212***	(0.129)	-1.238***	(0.181)	1,972,342	#2	✓			✓	✓			✓			✓
1.212***	(0.129)	-1.237***	(0.181)	1,972,342	Avg.	✓			✓	✓			✓			✓
1.219***	(0.131)	-1.230***	(0.185)	1,925,530	#1	✓	✓		✓	✓						
1.220***	(0.132)	-1.232***	(0.184)	1,992,276	#1	✓	✓	✓	✓	✓		✓	✓	✓		✓
1.221***	(0.132)	-1.233***	(0.184)	1,992,276	#1	✓	✓		✓	✓			✓	✓		✓
1.221***	(0.132)	-1.234***	(0.184)	1,925,530	#2	✓	✓		✓	✓						
1.221***	(0.131)	-1.232***	(0.184)	1,925,530	Avg.	✓	✓		✓	✓						
1.221***	(0.132)	-1.235***	(0.183)	1,992,276	#2	✓	✓	✓	✓	✓		✓	✓	✓		✓
1.221***	(0.132)	-1.233***	(0.183)	1,992,276	Avg.	✓	✓	✓	✓	✓		✓	✓	✓		✓
1.222***	(0.132)	-1.236***	(0.183)	1,992,276	#2	✓	✓		✓	✓			✓	✓		✓
1.222***	(0.132)	-1.234***	(0.183)	1,992,276	Avg.	✓	✓		✓	✓			✓	✓		✓
1.226* **	**(0.132)**	**-1.241*** **	**(0.184)**	**1,992,276**	**#1**	✓	✓		✓	✓			✓			✓
1.231***	(0.133)	-1.236***	(0.185)	1,992,276	#1	✓	✓	✓				✓				✓
1.232***	(0.133)	-1.238***	(0.184)	1,992,276	#2	✓	✓	✓				✓				✓
1.232***	(0.133)	-1.237***	(0.184)	1,992,276	Avg.	✓	✓	✓				✓				✓

This table shows regression coefficients of measures of patience and risk-taking on student PISA test. Values of patience and risk-taking are measured as country-level averages of these preferences from the GPS. All specifications include control variables for female, student's age, and PISA wave fixed effects. Rows show both coefficients under different specifications and are sorted in ascending order in the size of the coefficient on patience. The authors' main specification is highlighted in **bold**. *Weights refer to PISA sampling weights. Leverage is calculated based on all observations from unweighted regressions. Standard errors clustered at the country level are in parentheses. ***, ** and * mark estimates statistically different from zero at the 1, 5 and 10 percent significance level.*

3.2 Probing the robustness of Key Finding #2

Hanushek et al. (2022) argue that their Key Finding #1 is because of culture rather than country-level confounding factors. To make this argument, they focus on a sub-sample of students with a

migration background and use patience and risk-taking measures in their country of origin as main explanatory variables (see Appendix B for more details on the original empirical specification).

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. Column 3 of Table 2 in Hanushek et al. (2022) shows that being a migrant from a country in which people are 1 SD more patient is associated with an 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 (see Table 2, Column 3, page 2300). The estimated effect of patience is statistically significant at the 1% level, the estimated effect of risk-taking is statistically significant at the 5% level.

Again, we find the authors' decisions reasonable, but we believe that there are other reasonable specifications. We test the robustness of their findings to using different plausible values and the exclusion of high-leverage observations (see Section 3.1). We also test whether the results affected by the inclusion of a dummy variable indicating whether a migrant student's country of origin is part of the OECD.

Table 2 shows that the patience coefficients are similar across all 10 specifications. The point estimates range from 0.699 SD to 0.967 SD and all coefficients are statistically significant at the 1% level. Excluding the OECD dummy leads to around 20% smaller coefficients. However, this change does not affect the qualitative conclusion of these estimates. Table 2 further shows the risk-taking coefficients have the same sign in all 10 specifications. The coefficients range from -0.135 SD to -0.299 SD. However, the magnitude of the coefficients and statistical significance crucially depend on whether the OECD country of residence dummy is included as a control

variable. When the OECD dummy is included the point estimates range from -0.294 SD to -0.299 SD and all estimates are statistically significant at the 5% level. When it is excluded, point estimates range from -0.135 SD to -0.147 SD and none are significant even at the 10% level. Overall, we find that in *Key Finding #2* the conclusions regarding patience are robust to different specifications, but that the conclusions regarding risk-taking depend on whether the OECD dummy is included as a control.

Table 2. Robustness of the main findings in Table 2 in Hanushek et al. (2022) to alternative model specification choices

Outcome: Standardized PISA test scores in math					Specification:		
Country-of-origin patience coef.	std. err.	Country-of-origin risk-taking coef.	std. err.	Obs.	Plausible value used	Controls for OECD dummy	Incl. top 1% leverage obs.
0.699***	(0.154)	-0.145	(0.145)	80,398	#1		✓
0.704***	(0.155)	-0.147	(0.146)	79,595	#1		
0.704***	(0.155)	-0.135	(0.144)	80,398	Avg.		✓
0.711***	(0.155)	-0.137	(0.144)	79,595	Avg.		
0.719***	(0.160)	-0.142	(0.151)	80,398	#2		✓
0.726***	(0.161)	-0.144	(0.152)	79,595	#2		
0.931* **	**(0.116)**	**-0.294** **	**(0.122)**	**80,398**	**#1**	✓	✓
0.937***	(0.116)	-0.296**	(0.122)	79,595	#1	✓	
0.949***	(0.114)	-0.290**	(0.119)	79,595	Avg.	✓	
0.967***	(0.119)	-0.299**	(0.127)	79,595	#2	✓	

This table shows regression coefficients of measures of patience and risk-taking on student PISA test scores in a sample of first- and second-generation migrants. Values of patience and risk-taking are measured as country-level averages of these preferences from the GPS, and are assigned to students based on their country of origin. All specifications include controls for being a female student, student age, imputation dummies for female and age, and country-of-residence-by-wave fixed effects. Rows show both coefficients under different specifications and are sorted in ascending order in the size of the coefficient on patience. The authors' main specification is highlighted in **bold**. *Leverage is calculated based on specifications without sampling weights. Standard errors clustered at the country level are in parentheses. ***, ** and * mark estimates statistically different from zero at the 1, 5 and 10 percent significance level.*

It is unclear whether the OECD dummy is a necessary control for identifying causal effects of patience and risk-taking on test scores. On the one hand, it could capture omitted variables that are correlated with preferences in the migrant students' countries of origin and their test scores in

the host countries. On the other hand, it could be a "bad control" (see e.g., Cinelli, Forney, and Pearl 2022) if, for example, part of the effect of risk-taking would work via students' country of origin joining the OECD. However, even without this control variable, the point estimates of the risk-taking coefficients are sizable. There might economically significant effects of risk-taking that we would only be able to reliably detect with larger samples.

4. Wide replication using data from TIMSS and PIRLS instead of PISA

We conduct a wide replication by using data from TIMSS and PIRLS instead of PISA (see Appendix A for more details on these datasets and the included countries). TIMSS and PIRLS differ from PISA in a two important ways. First, they differ in terms of which kinds of students they sample. PISA samples 15-year-old students (typically in Grade 9 or 10) irrespective of their grade-level; TIMSS samples Grade 4 and Grade 8 students and PIRLS samples Grade 4 students irrespective of their age; PIRLS samples Grade 4 students irrespective of their age. Second, they differ in the design of the tests. PISA focuses on measuring cognitive and problem-solving skills using applications in math, science, and reading. TIMSS focuses on measuring content knowledge in math and science (e.g., algebra, geometry, chemistry, biology) and how that knowledge is applied by students (Mullis and Martin 2017). PIRLS measures reading literacy skills and how those skills are used by students in a variety of contexts—from reading for pleasure to reflecting on text, gathering information to perform a task or following instructions (Mullis and Martin 2019). Our wide replication therefore allows us to test whether Hanushek et al. (2022)'s results hold in a sample of younger students and for tests that focus on subject knowledge rather than problem-solving skills.

For this wide replication, we focus on the Key Finding #1 (see above). We cannot replicate Key Finding #2 because PIRLS and some waves of TIMSS lack information on the country of

origin of students and their parents. This data limitation also prevents us from estimating the same specification as in the original paper, which controls for the first-generation and second-generation status of students. Instead, in our analyses using TIMSS and PIRLS, we control for whether students are foreign born and a dummy for whether this information was imputed. We show in Appendix Table C1 that this choice of control variables does not affect our findings in the waves of TIMSS where we can estimate the same specification as Hanushek et al. (2022).

Table 3. Replication of Table 1 in Hanushek et al. (2022) using the TIMSS and PIRLS datasets

Outcome:	Original study estimates on PISA test scores in			Standardized TIMSS/PIRLS test scores in:			Standardized TIMSS/PIRLS test scores on grade:	
	Math	Science	Reading	Math	Science	Reading	4	8
Patience	1.226***	1.121***	1.110***	1.091***	1.070***	0.937***	0.959***	1.169***
	(0.132)	(.121)	(.114)	(0.166)	(0.151)	(0.140)	(0.141)	(0.168)
Risk-taking	-1.241***	-1.169***	-1.133***	-1.543***	-1.640***	-0.954***	-1.125***	-1.632***
	(0.184)	(.180)	(.198)	(0.227)	(0.251)	(0.199)	(0.200)	(0.218)
Obs.	1,992,276	1,992,276	1,950,722	1,950,724	1,950,718	910,585	1,765,081	1,096,127
PISA	✓	✓	✓					
TIMSS				✓	✓		✓	✓
PIRLS						✓	✓	
Grade 4				✓	✓	✓	✓	
Grade 8				✓	✓			✓

*This table shows regression coefficients of measures of patience and risk-taking on student math and science test scores from the TIMSS and student reading scores from PIRLS data, contrasted against estimates for math, science, and reading from PISA reported in the original study. We standardized original TIMSS and PIRLS test scores by subtracting 500 and dividing by 100. Values of patience and risk-taking are measured as country-level averages of these preferences from GPS. All specifications also include controls for being a female student, student age, foreign-born status, imputation dummies for female, age and foreign-born, and wave fixed effects. Standard errors clustered at the country level in parentheses. ***, ** and * mark estimates statistically different from zero at the 1, 5 and 10 significance level.*

Hanushek et al. (2022) estimate Key Finding #1 with PISA math scores in their main analysis and show that their results are robust to using PISA science and PISA reading scores (see Hanushek et al. (2022)'s Appendix, page A12). We reproduced these results in Columns 1-3 of Table 3. Columns 4-6 of Table 3 show that the Key Finding 1 is robust to using TIMSS math scores, TIMSS science scores, and PIRLS reading scores as dependent variables. All coefficients

of interests are of roughly similar magnitude as the original estimates and are statistically significant at the 1% level. Appendix Table C2 shows that Key Finding #1 also holds separately for Grade 4 and Grade 8 students.

Hanushek et al. (2022) use data from seven PISA waves (2000, 2003, 2006, 2009, 2012, 2015 and 2018) in their main analysis and show in their Online Appendix (p.A12) that their results are similar if only looking at PISA 2015, the first PISA wave after the GPS data was collected in 2012. We show in appendix Table C2 that all coefficients of interest in the 11 waves in which either TIMSS or PIRLS data was collected are of similar magnitude and 20 out of the 22 coefficients are statistically significant at the 1% level. Overall, our results confirm the robustness of Key Finding #1.

4. Conclusion

We have probed Hanushek et al. (2022)'s results with different specifications and in different datasets and find strong evidence for the robustness of first key finding: students in countries that are more patient and less risk-taking score higher on standardized tests. Testing several different empirical specifications also left us more confident that migrant students from countries with higher levels of patience indeed perform better on standardized tests in their host countries. However, we are now less certain about Hanushek et al. (2022)'s estimated effect of risk-taking in the migrant analysis. While the point estimates in all our robustness checks suggest that migrant students from countries with higher levels of risk-taking score worse on standardized tests, the statistical significance of this estimate depends on the choice of control variables. We hope that future high-powered replications will resolve this open question.

References

Brodeur, Abel, Scott E Carrell, David N Figlio, and Lester R Lusher. 2023. Unpacking p-hacking and publication bias. National Bureau of Economic Research.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2016. "Star wars: The empirics strike back." *American Economic Journal: Applied Economics* 8 (1):1-32.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, and Taizan Chan. 2016. "Evaluating replicability of laboratory experiments in economics." *Science* 351 (6280):1433-1436.

Camerer, Colin F, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, and Thomas Pfeiffer. 2018. "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." *Nature human behaviour* 2 (9):637-644.

Cinelli, Carlos, Andrew Forney, and Judea Pearl. 2022. "A crash course in good and bad controls." *Sociological Methods & Research*:00491241221099552.

de Gendre, Alexandra, Jan Feld, and Nicolás Salamanca. 2023. A Comment on" Patience, Risk-Taking, and Human Capital Investment Across Countries" by Hanushek et al.(2021). I4R Discussion Paper Series.

de Gendre, Alexandra, Jan Feld, Nicolás Salamanca, and Ulf Zölitz. 2023. "Same-sex role model effects in education." *Working paper series/Department of Economics* (438).

Eronen, Markus I, and Laura F Bringmann. 2021. "The theory crisis in psychology: How to move forward." *Perspectives on Psychological Science* 16 (4):779-788.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global evidence on economic preferences." *The Quarterly Journal of Economics* 133 (4):1645-1692.

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde. 2016. "The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences."

Fuchs, Thomas, and Ludger Wößmann. 2008. *What accounts for international differences in student prformance? A re-examination using PISA data*: Springer.

Hanushek, Eric A, Lavinia Kinne, Philipp Lergetporer, and Ludger Woessmann. 2022. "Patience, risk-taking, and human capital investment across countries." *The Economic Journal* 132 (646):2290-2307.

Mullis, Ina VS, and Michael O Martin. 2017. *TIMSS 2019 Assessment Frameworks*: ERIC.

Mullis, Ina VS, and Michael O Martin. 2019. "PIRLS 2021 Aessessment Frameworks." https://pirls2021.org/frameworks/.

OECD. 2019. "How PISA results are reported: What is a PISA score?" *Program. Int. Student Assess. Result from PISA 2018* 1 (41).

OECD. 2023. "Annex A1. Construction of indices." accessed October 6. https://www.oecd-ilibrary.org//sites/0a428b07-en/index.html?itemId=/content/component/0a428b07-en#s101.

Open Science Collaboration. 2015. "Estimating the reproducibility of psychological science." *Science* 349 (6251):aac4716.

Appendix

Appendix A. TIMSS and PIRLS data

We obtained publicly-available TIMSS and PIRLS original data files through the TIMSS and PIRLS international database (https://timssandpirls.bc.edu/databases-landing.html). The TIMSS 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 TIMSS files for waves 2003, 2007, 2011, 2015 and 2019 are provided in SPSS or SAS format, and so are the PIRLS files. 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. (2022).

Hanushek et al. (2022)'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 and 40 countries in PIRLS for which GPS data are also available. Of those countries, 35 countries are sampled in all three surveys and are used in the analyses of both the original paper and this replication study, 6 countries are only in TIMSS and PIRLS (Botswana, Egypt, Ghana, Iran, Pakistan and South Africa) and 4 are only in PISA (Costa Rica, Mexico, Peru, and Vietnam). The list of those countries is presented in Table A1.

Table A1. Countries used in Hanushek et al. (2022, Table 1) and in own conceptual replication using TIMSS and PIRLS

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

Country name	TIMSS	PIRLS	PISA
United Arab Emirates	1	1	1
United Kingdom (England, Scotland, Nothern Ireland)	1	1	1
United States	1	1	1
Vietnam			1
Total:	**50**	**40**	**49**

Appendix B. Details on the empirical specification for Main Findings 1 and 2

Empirical model to estimate Main Finding #1

To estimate the relationship between patience, risk-seeking, and student test achievement, Hanushek et al. (2022) use an ordinary least squares (OLS) regression to estimate the following model:

$$T_{ict} = \beta_1 Patience_c + \beta_2 Risk_c + \alpha_1 B_{ict} + \mu_t + \epsilon_{\text{ict}},$$

where T_{ict} is the standardized first plausible value of math ability of student *i* in country *c* at time *t*, $Patience_c$ is the standardized average level of patience of all respondents in country c in the GPS, $Risk_c$ is the equivalent standardized average of the risk-seeking, B_{ict} is a vector of control variables consisting of one female student dummy, student's age in years, one dummy variable for whether the student is a first-generation immigrant, and one dummy variable for whether the student is a second-generation immigrant, one dummy variable indicating whether the value of the female student dummy was imputed, one dummy variable indicating whether student age was imputed, and one dummy variable indicating whether first-generation immigration status was imputed. μ_t is a time fixed effect, which is accounted for by the inclusion of dummy variables for the different PISA waves. 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., 2022, page 2295). Standard errors are robust to clustering at the country level.

Empirical model to estimate Main Finding #2

To estimate the effect of patience and risk-taking on student achievement, Hanushek et al. (2022) use OLS regressions to estimate the following model:

$$T_{ioct} = \delta_1 Patience_o + \delta_2 Risk_o + \gamma_1 B_{ioct} + \theta_c \times \mu_t + \epsilon_{ioct},$$

where T_{ioct} is the standardized first plausible value of the math ability of migrant student i from country of origin o living in country of residence c at time t. $Patience_o$ and $Risk_o$ are the standardized country-level averages of patience and risk-taking in the migrant student's country of origin. B_{ioct} is a vector of control variables consisting of one female student dummy, student age in years, one dummy variable indicating whether the female dummy was imputed, one dummy variable indicating whether age was imputed, and one dummy variable indicating whether the migrant student's country of origin is part of the OECD. $\theta_c \times \mu_t$ are country-of-residence-wave fixed effects which are held constant in the regression by the inclusion of dummies for PISA wave-country of residence interaction terms.

Appendix C. TIMSS and PIRLS Analyses: Additional Tables

Table C1. The relationship between patience and risk-taking and math test scores in TIMSS across alternative ways to control for immigration background

Outcome:	Standardized TIMSS test scores in math		
Patience	1.096***	1.096***	1.091***
	(0.170)	(0.170)	(0.166)
Risk-taking	-1.560***	-1.561***	-1.543***
	(0.234)	(0.234)	(0.227)
Obs.	1,950,724	1,950,724	1,950,724
HKLW specification	✓	✓	
Missing flag for 2nd gen. migrant		✓	
Foreign-born flag and missing flag			✓

*This table shows regression coefficients of measures of patience and risk-taking on student math test scores from TIMSS data. We standardized original TIMSS test scores by subtracting 500 and dividing by 100. Values of patience and risk-taking are measured as country-level averages of these preferences from the GPS, and are assigned to students based on their country of origin. All specifications also include controls for female student, student age, imputation dummies for female and age, and wave fixed effects. The HKLW specification refers to the main model specification in Hanushek et al. (2022). Standard errors clustered at the country level in parentheses. ***, ** and * mark estimates statistically different from zero at the 1, 5 and 10 significance level.*

Table C2. The relationship between patience, risk-taking and test scores across waves of TIMSS and PIRLS

Outcome:	Standardized TIMSS/PIRLS average test scores in year:										
	1995	1999	2001	2003	2006	2007	2011	2015	2016	2019	2021
Patience	0.934***	1.219***	0.947***	1.354***	1.044***	1.004***	0.842***	0.917***	0.699***	1.135***	1.132***
	(0.202)	(0.238)	(0.173)	(0.238)	(0.186)	(0.217)	(0.136)	(0.197)	(0.146)	(0.273)	(0.227)
Risk-taking	-0.698	-1.980***	-2.018***	-2.286***	-1.420***	-1.700***	-1.112***	-0.929***	-0.911***	-1.451***	-0.749*
	(0.478)	(0.316)	(0.354)	(0.278)	(0.211)	(0.264)	(0.185)	(0.237)	(0.172)	(0.379)	(0.375)
Obs.	230,166	125,393	108,241	215,177	130,214	272,756	592,595	354,662	202,655	381,807	247,542
TIMSS	✓	✓		✓		✓	✓	✓		✓	
PIRLS			✓		✓		✓		✓		✓
Grade 4	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Grade 8	✓	✓		✓		✓	✓	✓		✓	

*This table shows regression coefficients of measures of patience and risk-taking on student math and science test scores from TIMSS and student reading scores from PIRLS data. We standardized original TIMSS and PIRLS test scores by subtracting 500 and dividing by 100. Each column shows the results for a different wave of the studies, which also means that the subject matter and grade changes across columns too. Values of patience and risk-taking are measured as country-level averages of these preferences from GPS. Average test scores are measured as the student-level standardized average of math and science test scores for the TIMSS data and the reading score for the PIRLS data. All specifications also include controls for female student, student age, foreign-born status, imputation dummies for female, age and foreign-born, and wave fixed effects. Standard errors clustered at the country level in parentheses. ***, ** and * mark estimates statistically different from zero at the 1, 5 and 10 significance level.*