Received: 12 October 2023 | Revised: 1 February 2024 | Accepted: 3 February 2024

DOI: 10.1002/jae.3045

REPLICATION

WILEY Journal of APPLIED ECONOMETRICS

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

Alexandra de Gendre[1] | **Jan Feld**[2] | **Nicolás Salamanca**[1]

[1]University of Melbourne, Australia

[2]Victoria University of Wellington, Wellington, New Zealand

Correspondence
Jan Feld, Victoria University of Wellington, Australia.
Email: jan.feld@vuw.ac.nz

Summary

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 most of the results are 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

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 patience is positively correlated with test score and risk-taking is negatively correlated with test scores, and they show 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.

* 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: Jan Feld. 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, & Salamanca, 2023). Stata do-files of the analysis reported in this paper are available at https://doi.org/10.15456/jae.2024044.2004572080

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 & 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 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, for example, because the original results were wrong 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 a control for whether the student's country of origin is part of the OECD. All other empirical decisions do not meaningfully affect the statistical 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 a subject (e.g. math) based on students' answers to the subject-component of the PISA test. The PISA data contains up to 10 different plausible values for each student's subject-specific test score, any one of which should give a good approximation of a student's subject-specific ability (the number of available plausible values is, depending on the PISA wave, either 5 or 10). In their main analysis, Hanushek et al. (2022) use the first of these plausible values for students' 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., 2016; Falk et al., 2018). In their analysis, the main explanatory variables are the country-averages

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

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.

3.1.1 | 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 check whether the results depend on the weighting scheme by estimating specifications without weights.

3.1.2 | 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 dummy variable flagging imputed values for this variable whenever we include the foreign-born dummy.

3.1.3 | 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

[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.

[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).

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)	−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***	(0.204)	1,992,276	MA		✓		✓	✓		✓		✓
1.212***	(0.129)	−1.238***	(0.181)	1,972,342	MA	✓			✓	✓		✓		✓
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***	(0.131)	−1.233***	(0.184)	1,925,530	MA	✓	✓		✓	✓				
1.222***	(0.132)	−1.234***	(0.183)	1,992,276	MA	✓	✓	✓	✓	✓	✓	✓	✓	✓
1.223***	(0.132)	−1.235***	(0.183)	1,992,276	MA	✓	✓		✓	✓		✓	✓	✓
1.226* **	**(0.132)**	**−1.241*** **	**(0.184)**	**1,992,276**	**#1**	✓	✓		✓	✓		✓		✓
1.231***	(0.133)	−1.236***	(0.185)	1,992,276	#1	✓	✓	✓			✓			✓
1.233***	(0.133)	−1.238***	(0.184)	1,992,276	MA	✓	✓	✓			✓			✓

Note: *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**. MA indicates that coefficients were calculated using the model averaging procedure described in Section 3.1. Weights refer to PISA sampling weights. Leverage is calculated based on all observations from unweighted regressions that are otherwise identical to the author's main specification. Standard errors clustered at the country level are in parentheses.
*mark estimates statistically different from zero at the 10% significance level.
**mark estimates statistically different from zero at the 5% significance level.
***mark estimates statistically different from zero at the 1% significance level.

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

3.1.4 | 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 combining five of the provided plausible values via model averaging, as instructed by the PISA data analysis manuals (e.g. OECD, 2009).[4] Model averaging consists of estimating the parameters of interest (the patience and risk-taking coefficients and their standard errors) using each plausible value as the dependent variable, averaging the parameter estimates across models to arrive at the final estimates, and adjusting the standard errors. This adjustment ensures the final standard errors reflect, in addition to the traditional sampling error, the additional uncertainty resulting from using a measure of ability constructed from students' answers to different questions in a test. We implement this estimator using Stata's user-written `pv` command (Macdonald, 2019).

3.1.5 | 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 with model averaging). Stata do-files for these robustness checks are available at https://doi.org/10.15456/jae.2024044.2004572080.

Table 1 shows that the patience coefficients are very stable across all different specifications. The point estimates range from 1.209 SD to 1.233 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 SD to −1.241 SD 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. Our results also suggest that none of the coefficient estimates for either patience or risk-taking in Table 1 are statistically significantly different from Hanushek et al.'s main estimates.[5] Overall, we find that *Key finding #1* is robust to several alternative empirical decisions.

[4]The Hanushek et al. (2022) data includes five plausible values for math test scores for waves 2000, 2003, 2006, 2009 and 2012, and 10 plausible values for waves 2015 and 2018. We therefore only the first five plausible values in our analyses since we estimate models which combine all waves and some key parameters (e.g., year fixed effects) would not be identified for all plausible values, making model averaging impossible. This limitation implies that our (implicit) estimate of the sampling error from students answering different test questions is less precise than if we would have been able to use all 10 plausible values.

[5]To properly answer these questions, we would need to calculate the t-statistics $t_i = \left(\widehat{\beta}_{HKLW} - \widehat{\beta}_i\right)\left(SE^2_{\widehat{\beta}_{HKLW}} + SE^2_{\widehat{\beta}_i} + 2Cov\left(\widehat{\beta}_{HKLW}, \widehat{\beta}_i\right)\right)^{-\frac{1}{2}}$ where $\widehat{\beta}_{HKLW}$ is the preferred estimate of Hanushek et al. (2022) and $\widehat{\beta}_i$ is the estimate of our $i = 1...I$ alternative specification. This would involve constructing covariance matrices for Hanushek et al.'s main specification and each alternative specification, which quickly becomes cumbersome and is technically challenging when estimates come from model averaging, use cluster-robust standard errors, or are produced using probability weighting. Instead, by making the simplifying assumption $Cov\left(\widehat{\beta}_{HKLW}, \widehat{\beta}_i\right) = 0$, we can calculate the t-statistic from reported coefficient and standard error estimates and compare this to critical values from the standard normal distribution. We make this simplification and note that the results of these tests are only suggestive. The estimates of this exercise for Tables 1 and 2 are produced in the results log file available at https://doi.org/10.17605/OSF.IO/KGT8Z.

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 model averaging instead of the first plausible value, 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 seven specifications. The point estimates range from 0.699 SD to 0.948 SD and all coefficients are statistically significant at the 1% level. Excluding the OECD dummy leads to around 0.2 SD smaller coefficients. However, this change does not affect the qualitative conclusion of these estimates, and the differences between the preferred estimates of Hanushek et al. (2022) and our alternative estimates are never statistically significant at conventional levels.

Table 2 further shows the risk-taking coefficients have the same sign in all seven specifications. The coefficients range from −0.136 SD to −0.296 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 −0.291 SD to −0.296 SD and all estimates are statistically significant at the 5% level. When it is excluded, point estimates range from −0.136 SD to −0.147 SD and

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-originpatience coef.	std. err.	Country-of-originrisk-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)	−0.136	(0.145)	80,398	MA		✓
0.710***	(0.155)	−0.138	(0.146)	79,595	MA		
0.931* **	**(0.116)**	**−0.294** **	**(0.122)**	**80,398**	**#1**	✓	✓
0.937***	(0.116)	−0.296**	(0.122)	79,595	#1	✓	
0.948***	(0.115)	−0.291**	(0.121)	79,595	MA	✓	

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**. MA indicates that coefficients and standard errors are calculated using the model averaging procedure described in Section 3.1. Leverage is calculated based on all observations from unweighted regressions that are otherwise identical to the author's main specification. Standard errors clustered at the country level are in parentheses.
*mark estimates statistically different from zero at the % significance level.
**mark estimates statistically different from zero at the 5% significance level.
***mark estimates statistically different from zero at the 1% significance level.

none are significant, even at the 10% level. Nevertheless, and just as in the case of patience, none of the differences between the original authors' preferred estimates and our alternative estimates are statistically significant at conventional levels. 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.

It is unclear whether the OECD dummy is an appropriate control for identifying the causal effects of patience and risk-taking on test scores. On the one hand, being from an OECD country might be a "confounder" affecting preferences of migrant students and the test scores in their host countries. If this is the case, including the OECD dummy would be appropriate. On the other hand, it might be a "mediator" if part of the effect of preferences comes via students' country of origin joining the OECD. For example, having a more risk-taking population may increase a country's probability of joining the OECD and being part of this organization may affect students' test scores even after they migrate to another country. If this is the case, including an OECD dummy is not appropriate because it would "control away" part of the effect that we are interested in (for a more extensive discussion on "bad controls" see Cinelli et al. (2022) and Cunningham (2021)). However, even without this control variable, the point estimates of the risk-taking coefficients are sizable. There might be 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 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 & 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 & 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).

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. 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.

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

	Original study estimates on PISA test scores in			Standardized TIMSS/PIRLS test scores in:			Standardized TIMSS/PIRLS test scores on grade:	
Outcome:	**Math**	**Science**	**Reading**	**Math**	**Science**	**Reading**	**4**	**8**
Patience	1.226***	1.121***	1.110***	1.091***	1.070***	0.937***	0.961***	1.150***
	(0.132)	(0.121)	(0.114)	(0.166)	(0.151)	(0.140)	(0.142)	(0.166)
Risk-taking	−1.241***	−1.169***	−1.134***	−1.543***	−1.640***	−0.954***	−1.127***	−1.603***
	(0.184)	(0.180)	(0.198)	(0.227)	(0.251)	(0.199)	(0.200)	(0.218)
Obs.	1,992,276	1,992,276	1,950,722	1,950,724	1,950,724	910,587	1,765,181	1,096,130
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. The dependent variables in the first three columns are the first plausible values of students' math, science, or reading test scores, respectively. The remaining specifications show results using model averaging based on the first five plausible values from TIMSS or PIRLS (see Section 3.1). We do not use model averaging in the first three specifications because we wanted to exactly replicate the results of Hanushek et al. (2022). Furthermore, their replication package only includes the first plausible values for science and reading. However, our narrow replication suggests that using the first plausible value instead of model averaging does not meaningful affects the results. 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.
*mark estimates statistically different from zero at the 10% significance level.
**mark estimates statistically different from zero at the 5% significance level.
***mark estimates statistically different from zero at the 1% significance level.

5 | CONCLUSION

We have probed Hanushek et al.'s (2022) 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 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.'s (2022) 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.

ACKNOWLEDGEMENTS
Open access publishing facilitated by Victoria University of Wellington, as part of the Wiley - Victoria University of Wellington agreement via the Council of Australian University Librarians.

DATA AVAILABILITY STATEMENT
Stata do-file of the analysis reported in this paper are available at https://doi.org/10.15456/jae.2024044.2004572080. We do not have permission to share the data used in this paper.

REFERENCES
Brodeur, A., Carrell, S. E., Figlio, D. N., & Lusher, L. R. (2023). *Unpacking p-hacking and publication bias*. National Bureau of Economic Research.

Brodeur, A., Lé, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. *American Economic Journal: Applied Economics*, *8*(1), 1–32. https://doi.org/10.1257/app.20150044

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*. https://doi.org/10.1177/00491241221099552

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale university press.

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

de Gendre, A., Feld, J., Salamanca, N., & Zölitz, U. (2023). Same-sex role model effects in education. Working paper series/Department of Economics (438).

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2016). The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences.

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. https://doi.org/10.1093/qje/qjy013

Fuchs, T., & Wößmann, L. (2008). *What accounts for international differences in student prformance? A re-examination using PISA data*. Springer. https://doi.org/10.1007/978-3-7908-2022-5_10

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. https://doi.org/10.1093/ej/ueab105

Macdonald, K. (2019). PV: Stata module to perform estimation with plausible values.

Mullis, I. V. S., & Martin, M. O. (2017). TIMSS 2019 Assessment Frameworks: ERIC.

Mullis, I. V. S., & Martin, M. O. (2019). PIRLS 2021 Aessessment Frameworks. https://pirls2021.org/frameworks/

OECD. (2009). *PISA data analysis manual: SPSS and SAS*. OECD Publishing Paris.

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.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: de Gendre, A., Feld, J., & Salamanca, N. (2024). Re-examining the relationship between patience, risk-taking, and human capital investment across countries. *Journal of Applied Econometrics*, 1–9. https://doi.org/10.1002/jae.3045