INSTITUTE for REPLICATION

No. 111
I4R DISCUSSION PAPER SERIES

A Comment on Wu, Zhang, Wang (2023)

Alejandro Abarca

Felipe Juan

Ke Lyu

Alexa Prettyman

Idil Tanrisever

April 2024

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 111

A Comment on Wu, Zhang, Wang (2023)

Alejandro Abarca[1], Felipe Juan[2], Ke Lyu[3], Alexa Prettyman[4], Idil Tanrisever[5]

[1]Oregon State University, Corvallis/USA
[2]Howard University, Washington D.C./USA
[3]University of Nevada, Reno/USA
[4]Towson University, Towson/USA
[5]University of California, Irvine/USA

APRIL 2024

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

A comment on Wu, Zhang, Wang (2023)

Alejandro Abarca (Oregon State University)
Felipe Juan (Howard University)
Ke Lyu (University of Nevada, Reno)
Alexa Prettyman (Towson University)
Idil Tanrisever (University of California, Irvine)

January 6, 2024

Abstract

Wu et al. (2023) estimate the effect of classroom seating arrangements in China using a randomized control trial with two treatment schemes. The first treatment scheme involves seating high and low achieving students together, and the second treatment involves this same seating arrangement with financial incentives for the high-achieving students, if their deskmates' test scores improved. All statistically significant impacts come from the incentivized treatment scheme. Wu et al. (2023) find that low-achieving students sitting next to incentivized high-achieving students perform 0.24 SD (p-value=0.018) better on math exams. In addition, being assigned to the incentive treatment scheme increased extraversion and agreeableness for low and high achieving students. Lastly, they do not find much evidence of peer effects on test scores nor personality traits. This study is computationally reproducible using their provided replication package. We ran their code using Stata 14, 17, and 18. After running their replication package, we further investigated Tables 2-5. The main conclusions are generally robust to various coding decisions. Notably, in investigating the peer effects, when we change the specification to also control for the difference in baseline scores between the student and their deskmate, we find that the more dissimilar deskmates are at baseline, the bigger the peer effects.

1. Introduction

Wu et al. (2023) estimate the effect of classroom seating arrangements in China during the 2015-2016 school year using a randomized control trial with two treatment schemes. The first treatment scheme involves seating high and low achieving students together (MS), and the second treatment involves this same seating arrangement with financial incentives for the high-achieving students (MSR), if their deskmates' test scores improved. The final sample sizes in the control, first treatment, and second treatment are 574, 634, and 594, respectively.

In the current paper, all statistically significant impacts come from the incentivized treatment scheme. Regarding academic outcomes, Wu et al. (2023) find that low-achieving students sitting next to incentivized high-achieving students perform 0.24 SD (p-value=0.018) better on math exams (Wu et al. 2023 Table 3). This estimate comes from their model that includes controls. When they exclude controls, the estimate is only marginally significant (p-value=0.10). We check the robustness of these results two ways. First, rather than running separate regressions for the lower-track and higher-track students, we run the regressions with an indicator variable for higher-track students. We find being assigned to a MSR class increases math test scores by 0.17 SD (statistically significant with 95% confidence). Second, we check the robustness of the main estimate by varying the controls. For example, we control for household income, distance to the head teacher, and remove health variables. We find that the estimate for math varies from 0.19 to 0.24 SD (statistically significant between the 10 to 5% significance level).

Another set of outcomes of interest are the "Big Five" personality traits (Wu et al. 2023 Table 4). Being assigned to the incentive treatment scheme statistically increased extraversion and agreeableness for low and high achieving students. Upon examining the code, we found that some raw variables used to construct the "Big Five" personality traits are miscoded. After fixing this error, the results still hold. We also modified the way these variables were coded. This modification implies that being assigned to the MRS treatment group only statistically increased extraversion by 5% among the lower-track students.

Lastly, Wu et al. (2023) find little evidence of peer effects (Table 5). We extend on their analysis by first estimating the peer effects on Chinese and Math scores separately rather than the total average. We find that under the original econometric specification, there are no peer effects. However, we also re-estimate this model and control for the difference in baseline scores between the student and their deskmate. Under this new specification, we find very strong and significant peer effects on test scores. This is a valuable contribution to the original paper and suggests that the more dissimilar deskmates are at baseline, the bigger the peer effects. These results hold for students in both treatments considered in the study. We estimate this specification for the "Big Five" personality traits and find no statistically significant peer effects. This "dissimilar effect" only applies to test scores.

2. Reproducibility

This study is computationally reproducible using their provided replication package. We ran their code using Stata 14, 17, and 18. In addition, we modified code and analyses to test the robustness of the results in Tables 2-5. Generally, the conclusions are robust to these decisions.

Upon investigating variables in the final dataset provided online, we found that a handful (6 out of 120) of the variables used to construct the “Big Five” personality traits seem to be miscoded. That is, they should take values from 1 to 5, but some are -1 and 0, which we suspect were supposed to be missing. Anywhere between 5 to 144 students were recoded; however, this correction does not impact the reproducibility of the results because only a handful of values were -1 and most of the corrections changed from 0 to missing. Turns out, 0 and missing are treated similarly when constructing the “Big Five” personality traits.

3. Replication

Our replication exercises consist of adding additional statistics, modifying variable construction, and varying controls in analyses. This section discusses these exercises in context to the different analyses.

3.1 Summary Statistics

While replicating the summary statistics, we decided to add T-Tests between the control and treatment schemes. We find statistically significant differences in the “Big Five” personality traits between the control group and the MS treatment group. These discrepancies could challenge the assumption of random assignment, suggesting that the control group and the MS treatment group may not be balanced.

3.2 Achievement effect estimates and robustness checks

Table 1 shows the regressions with an indicator variable for higher-track students instead of separate regressions for the lower-track and higher-track students. The table shows similar results as the main analysis (Wu et al. 2023 Table 3), with significant results for the average score and the mathematics score. In fact, the paper separates these groups to portray that the effects are more pronounced for the lower-track students.

Table 2 changes the control variables used in the main analysis. Columns (1) and (4) use an additional control variable for distance between student and head teacher, which is equivalent to distance to school in rural China. Including this variable changes the estimates significantly. For the average score outcome, the point estimate becomes insignificant when this control variable is added. While the mathematics score outcome remains significant, it is only significant at the 10 percent level, compared to the 5 percent significance in the main analysis. Columns (2) and (5) eliminate the health variable, since self-reported health may be biased. Removing this variable changes the outcome for average score such that the point estimate is no longer statistically significant. However, the outcome for mathematics score is robust to this change. Lastly, we use

parents' income instead of education for columns (3) and (6) and we find that the results are robust to using parents' income instead of parents' education.

3.3 The "Big Five" effects and robustness

Table 3 shows the results after we reconstruct the "Big Five" personality traits. The authors construct these variables by summing 12 intermediate variables that range from 1 to 5. Alternatively, we construct these variables by taking the mean of these 12 intermediate variables. We do this to reduce the variation in the final scale and to adjust for missingness. In doing this, we only find a statistically significant effect on extraversion among the MSR classes for both the lower-track and upper-track students. Among the lower-track students, we also find a statistically significant effect on agreeableness.

3.4 Peer effects and robustness checks

With regards to Wu et al. (2023) Table 5 "Peer effects in the mixed-seating classes", we first expanded on the results of column 1 by running the estimates for Chinese and Math scores separately and not just the total average z-scores. In addition, instead of the baseline score, we controlled for the difference in the baseline distribution. We show these results in Table 4 below.

With the original specification, we find no peer effects in either Chinese nor Math Z-scores. However, when controlling the difference in baseline performance, we find very significant and positive peer effects in both types of scores. Furthermore, the difference in baseline performance is also very significant and positive. This finding suggests that peers that initially were more dissimilar in performance can create better test score outcomes. It is worth noting that these results hold for both types of treatment and across the two tracks considered in the study.

Lastly, we re-estimated the models for the five personality traits considered in the paper. We do not find peer effects, nor is the difference in baseline performance significant for any of the dependent variables. Therefore, positive and significant peer effects conditional on differences at baseline are only found for the test scores and not the personality traits.

3.5 Representativeness of the sample

This randomized control trial was conducted in impoverished rural areas, which raises questions about the representativeness of the total sample. The effects observed in these poor areas may not be applicable in more developed regions. We examined the heterogeneity among students from low- and high-income households and found that the treatment effect of the MSR classes is significantly lower for the high-income group.

4. Conclusion

The paper is well constructed, the provided replication package is easy to run and follow, and the variables are clearly labeled. The code to produce Tables 1-9 runs and reproduces the results in the paper. We focused the robustness replication exercises on Tables 2-5. We varied the construction of some variables and the set of controls. Through these exercises, overall, the

main conclusions hold. Other replication exercises might want to start with Table 6, focus on the Appendix Tables, do an attrition analysis, and investigate the results when correcting for imbalance in the “Big Five” personality traits.

References

Wu, Jia, Junsen Zhang, and Chunchao Wang. 2023. "Student Performance, Peer Effects, and Friend Networks: Evidence from a Randomized Peer Intervention." American Economic Journal: Economic Policy, 15 (1): 510-42.

Tables

Table 1 – Effects of interventions on the Students' Test Scores - Combined Samples

	Average Score		Chinese Score		Mathematics Score	
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
MS	-0.039	-0.025	-0.056	-0.041	-0.025	-0.018
	[0.097]	[0.056]	[0.081]	[0.046]	[0.140]	[0.076]
MSR	0.075	0.091*	-0.018	-0.010	0.154	0.174**
	[0.078]	[0.051]	[0.075]	[0.037]	[0.111]	[0.070]
Controls	NO	YES	NO	YES	NO	YES
P-value of intervention in the lower track (MS=MSR)	0.014	0.044	0.640	0.508	0.196	0.014
Observations	1,802	1,802	1,802	1,802	1,802	1,802

Notes: This table reports the regression estimates of treatment effects of MS and MSR interventions on students' endline average scores and Chinese and mathematics test scores. The estimated equations are specified by equation (1) in Wu et al. (2023) for odd columns and equation (2) in Wu et al. (2023) for even columns. The table combines the samples for lower-track students and upper tract students and uses an indicator variable for the high-track group. Controls include the corresponding own baseline personality trait, gender, age, height, health status, hukou registration status, minority status, father's education, mother's education, and whether the student's household has a computer or a car. Robust standard errors clustered at the class level are reported in brackets.
Significant at the ***[1%] **[5%] *[10%] level.
Comparison to Table 3 in Wu et al. (2023).

Table 2 – Effects of interventions on the Students' Test Scores - Modifying Controls

	Average score			Mathematics score		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Panel A. Lower-track students						
MS	0.015	-0.004	0.002	0.016	-0.021	-0.013
	[0.112]	[0.086]	[0.084]	[0.139]	[0.106]	[0.105]
MSR	0.093	0.127	0.138*	0.192*	0.224**	0.239**
	[0.084]	[0.082]	[0.079]	[0.109]	[0.099]	[0.097]
Controls	YES	YES	YES	YES	YES	YES
P-value of intervention in the lower track (MS=MSR)	0.34	0.0657	0.0517	0.105	0.0066	0.0053
Observations	901	901	901	901	901	901
Panel B. Upper-track students						
MS	-0.058	-0.068	-0.068	-0.024	-0.039	-0.038
	[0.058]	[0.056]	[0.056]	[0.097]	[0.086]	[0.086]
MSR	0.066	0.033	0.042	0.12	0.098	0.104
	[0.069]	[0.057]	[0.058]	[0.104]	[0.083]	[0.081]
Controls	YES	YES	YES	YES	YES	YES
P-value of intervention in the lower track (MS=MSR)	0.0835	0.1583	0.1296	0.2096	0.1757	0.1534

Observations	901	901	901	901	901	901

Notes: This table reports the regression estimates of treatment effects of MS and MSR interventions on students' endline average scores and mathematics test scores. The estimated equations are specified by equation (2) in Wu et al. (2023). Panel A reports the estimated results for lower-track students, and panel B reports estimated results for upper-track students. Controls include the corresponding own baseline personality trait, gender, age, height, health status, hukou registration status, minority status, father's education, mother's education, and whether the student's household has a computer or a car. Columns (1) and (4) use an additional control variable for distance between student and head teacher, which is equivalent to distance to school in rural China. Columns (2) and (5) eliminate the health variable. Columns (3) and (6) replace father's and mother's education variable with father's and mother's income. Robust standard errors clustered at the class level are reported in brackets.
Significant at the ***[1%] **[5%] *[10%] level.
Comparison to Table 3 in Wu et al. (2023).

Table 3 – Effects of interventions on the "Big Five" Personality Traits - Reconstructed "Big Five"

	Extraversion (1)	Agreeableness (2)	Openness (3)	Neuroticism (4)	Conscientiousness (5)
Panel A. Lower-track students					
MS	0.058	0.139	-0.002	0.075	0.092
	[0.071]	[0.109]	[0.053]	[0.062]	[0.073]
MSR	0.186**	0.212**	0.092	0.038	0.103
	[0.069]	[0.105]	[0.066]	[0.057]	[0.066]
Controls	YES	YES	YES	YES	YES
P-value of intervention in the lower track (MS=MSR)	0.155	0.545	0.187	0.597	0.895
Mean of the dependent variable for students in control classes	3.18	2.80	3.34	3.18	3.17
Observations	901	901	901	901	901
Panel B. Upper-track students					
MS	-0.054	-0.018	0.044	0.005	-0.018
	[0.057]	[0.111]	[0.066]	[0.051]	[0.063]
MSR	0.140**	0.154	0.026	-0.027	0.054
	[0.057]	[0.106]	[0.042]	[0.056]	[0.063]
Controls	YES	YES	YES	YES	YES

P-value of intervention in the lower track (MS=MSR)	0.009	0.079	0.114	0.617	0.366
Mean of the dependent variable for students in control classes	3.22	2.84	3.33	3.21	3.23
Observations	901	901	901	901	901

Notes: This table reports the regression estimates of treatment effects of MS and MSR interventions on students' "big five' personality traits. The dependent variable is the mean of "big five" personality traits surveyed in the endline questionnaire. The estimated equation is equation (2). Panel A reports the estimated results for lower-track students, and panel B reports estimated results for upper-track students. Controls include the corresponding own baseline personality trait, gender, age, height, health status, hukou registration status, minority status, father's education, mother's education, and whether the student's household has a computer or a car. Robust standard errors clustered at the class level are reported in brackets.
Significant at the ***[1%] **[5%] *[10%] level.
Comparison to Table 4 in Wu et al. (2023).

Table 4 - Peer effects in the mixed-seating classes for Chinese and Math scores

	Chinese Z-scores		Math Z-scores	
	(1)	(2)	(3)	(4)
Panel A: Lower-rank students in MS classes				
Deskmate's baseline performance	-0.011	0.624***	-0.082	0.479**
	[0.130]	[0.154]	[0.111]	[0.171]
Difference in baseline performance		0.634***		0.561***
		[0.075]		[0.085]
Class-by-height-group fixed effect	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	317	317	317	317
Panel B: Upper-rank students in MS classes				
Deskmate's baseline performance	-0.005	0.548***	-0.032	0.631***
	[0.020]	[0.084]	[0.036]	[0.105]
Difference in baseline performance		0.553***		0.663***
		[0.081]		[0.100]
Class-by-height-group fixed effect	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	317	317	317	317
Panel C: Lower-rank students in MSR classes				
Deskmate's baseline	-0.169	0.450**	0.088	0.698***

performance				
	[0.114]	[0.202]	[0.064]	[0.141]
Difference in baseline performance		0.620***		0.609***
		[0.108]		[0.110]
Class-by-height-group fixed effect	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	297	297	297	297
Panel D: Lower-rank students in MSR classes				
Deskmate's baseline performance	0.001	0.357*	0.033	0.704***
	[0.034]	[0.172]	[0.050]	[0.121]
Difference in baseline performance		0.355*		0.672***
		[0.171]		[0.115]
Class-by-height-group fixed effect	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	297	297	297	297

Notes: Columns 1 and 3 show the estimates of baseline of the deskmate´s baseline performance on endline average Chinese and mathematics test scores conditional on the initial performance of the student. Columns 2 and 4 control for the difference in baseline performance of deskmates and show the estimate of baseline of the deskmate´s baseline performance on endline average Chinese and mathematics test scores. Controls include gender, age, height, health status, hukou registration status, minority status, father's education, mother's education, and whether the student's household has a computer or a car. Robust standard errors clustered at the class level are reported in brackets.
Significant at the ***[1%] **[5%] *[10%] level.
Comparison to Table 5 in Wu et al. (2023).