

PAPER

Measurement Sensitivity in Heterogeneous Treatment Effect Analysis: Evidence from the Project STAR Experiment

Mika Takechi¹, Shigeyuki Hamori^{1,2}  

¹Yamato University,
Suita, Japan

²Kobe University, Kobe, Japan

hamori@econ.kobe-u.ac.jp

ABSTRACT

This paper examines how the construction and timing of outcome and heterogeneity measures influence the estimation of treatment heterogeneity in randomized educational experiments. Using data from the Project STAR experiment as a laboratory setting, we compare treatment-effect estimates across alternative definitions of student achievement and student heterogeneity. Baseline specifications using contemporaneous Grade 3 achievement measures suggest positive interaction effects between small-class assignment and measured student heterogeneity. However, these interaction effects become substantially weaker and generally lose statistical significance when heterogeneity is measured using pre-treatment achievement scores from kindergarten, grade 1, and grade 2. Similar patterns emerge across reading and mathematics outcomes as well as across alternative cross-domain specifications. The results indicate that conclusions regarding treatment heterogeneity are highly sensitive to the timing, domain, and construction of heterogeneity measures. More broadly, the findings demonstrate that measurement sensitivity represents an important issue in empirical data analysis of heterogeneous treatment effects. The results highlight the importance of robustness assessment under alternative measurement constructions before drawing substantive conclusions in experimental and policy evaluation studies.

KEYWORDS

data analysis, treatment heterogeneity, measurement sensitivity, outcome construction, policy evaluation

1 INTRODUCTION

Randomized controlled trials (RCTs) have become one of the most influential approaches for evaluating educational policies and interventions. By ensuring that treatment assignment is independent of individual characteristics, randomization provides a credible basis for identifying average treatment effects [1] and [2]. In education

Takechi, M., Hamori, S. (2026). Measurement Sensitivity in Heterogeneous Treatment Effect Analysis: Evidence from the Project STAR Experiment. *IETI Transactions on Data Analysis and Forecasting (iTDAF)*, 4(3), pp. 34–46. <https://doi.org/10.3991/itdaf.v4i3.62410>

Article submitted 2026-05-21. Revision uploaded 2026-06-25. Final acceptance 2026-07-17.

© 2026 by the authors of this article. Published under CC-BY.

research, however, attention has increasingly shifted beyond average effects toward the analysis of treatment heterogeneity, focusing on how the effects of educational interventions may vary across students and learning environments [3] and [4].

A growing methodological literature has emphasized that treatment effects may vary substantially across individuals and contexts, leading to increased interest in heterogeneous treatment-effect estimation and subgroup analysis [5], [6], and [7]. At the same time, recent research has highlighted the importance of measurement choices in empirical evaluation. Alternative constructions of outcomes, scales, and performance measures may generate different substantive conclusions even when the underlying intervention remains unchanged [8]. Related concerns also arise when researchers examine multiple outcomes or alternative outcome definitions within the same experiment, raising issues of outcome selection and multiple inference [9]. These concerns suggest that conclusions regarding treatment heterogeneity may depend not only on identification strategies but also on how outcomes and heterogeneity measures are defined and constructed.

This paper examines a fundamental challenge in the treatment heterogeneity literature: the sensitivity of estimated interaction effects to the construction and timing of both outcome and heterogeneity measures. We argue that estimated treatment heterogeneity may depend critically on how educational outcomes and student heterogeneity are defined and measured. In particular, when observed outcomes reflect interactions between educational interventions and student characteristics, alternative constructions of outcomes and heterogeneity measures may generate substantially different empirical conclusions, even within the same randomized experiment. Consequently, randomized assignment alone does not necessarily guarantee robust inference regarding heterogeneous treatment effects.

These issues are especially relevant in the evaluation of educational achievement, where learning outcomes are inherently multidimensional. Reading, mathematics, and composite achievement measures may capture different aspects of student performance and therefore imply different patterns of estimated treatment heterogeneity. More generally, when treatment heterogeneity is evaluated using alternative outcome constructions, empirical conclusions may depend importantly on how outcomes and latent student characteristics are measured.

An additional challenge arises when heterogeneity measures are constructed using variables that may themselves be affected by treatment assignment. The causal inference literature has long emphasized the distinction between pre-treatment covariates and post-treatment variables, noting that post-treatment measures can complicate causal interpretation and introduce additional sources of bias [10] and [11]. Recent work also highlights the value of pre-treatment covariates in improving the credibility and precision of treatment-effect estimation [12]. These concerns are particularly relevant in educational settings, where measures of student ability are often observed after educational interventions have already taken place.

To investigate these issues empirically, we revisit the Project STAR experiment, one of the most influential randomized evaluations in education economics [2]. The experiment has generated a large literature examining both short-run and long-run consequences of class-size reductions, as well as recent efforts to revisit the STAR data using modern empirical approaches [13] and [14]. Using the STAR data, we compare estimated treatment heterogeneity across alternative outcome definitions and heterogeneity measures.

The paper contributes to the literature in two ways. First, it provides evidence that conclusions regarding treatment heterogeneity can be highly sensitive to the timing, domain, and construction of heterogeneity measures. Using data from the Project STAR experiment as a laboratory setting, we show that interaction effects

observed under contemporaneous achievement measures largely disappear when heterogeneity is measured using pre-treatment achievement scores.

Second, the paper highlights a broader methodological issue in experimental and policy evaluation research. While randomized assignment ensures unbiased estimation of average treatment effects, it does not necessarily guarantee stable inference regarding treatment heterogeneity. The results demonstrate that estimated interaction effects may vary substantially across alternative definitions of outcomes, achievement domains, and heterogeneity measures. Consequently, careful attention to measurement construction is essential when interpreting heterogeneous treatment effects in empirical applications.

More broadly, the paper uses the Project STAR experiment as a laboratory for examining how alternative constructions of outcomes and heterogeneity measures affect the interpretation of heterogeneous treatment effects. The contribution therefore lies not in reassessing class-size effects per se, but in highlighting the measurement sensitivity of treatment-heterogeneity estimates within randomized experiments.

The remainder of the paper is organized as follows. Section 2 presents the conceptual framework. Section 3 describes the data and empirical strategy. Section 4 reports the empirical results and sensitivity analyses. Section 5 concludes.

2 CONCEPTUAL FRAMEWORK AND EMPIRICAL STRATEGY

2.1 Empirical framework

Educational outcomes are inherently multidimensional, and the measured effects of educational interventions may therefore depend on how student achievement is defined and evaluated. Recent research has emphasized that alternative outcome constructions and measurement choices can influence empirical conclusions [8]. Motivated by these concerns, this paper examines whether estimated treatment heterogeneity varies across alternative outcome constructions within the same randomized educational experiment.

To investigate this issue empirically, we estimate both a conventional treatment-effect specification and an augmented specification that allows for interaction between treatment assignment and student heterogeneity.

Let Y_i denote observed student achievement and let T_i be an indicator for assignment to a small class. The baseline specification is given by

$$Y_i = \alpha + \gamma T_i + \epsilon_i \quad (1)$$

where Y_i denotes the observed outcome variable, T_i is an indicator for small-class assignment, γ captures the average treatment effect under random assignment, and ϵ_i represents unobserved determinants of student achievement.

Although random assignment ensures that treatment is independent of student characteristics, observed educational outcomes may still reflect interaction between educational interventions and individual heterogeneity. To account for this possibility, we estimate the following interaction specification:

$$Y_i = \alpha + \gamma_0 T_i + \delta(T_i \times Ability_i) + \epsilon_i \quad (2)$$

where $Ability_i$ represents a proxy for student heterogeneity constructed from alternative standardized test scores, γ_0 captures the baseline treatment effect, δ measures

the extent to which treatment effects vary systematically with student heterogeneity, and ϵ_i denotes unobserved determinants of observed achievement outcomes.

Comparing equations (1) and (2) allows us to evaluate whether estimated treatment effects remain stable across alternative outcome measures. A non-zero estimate of δ indicates that the effects of small-class assignment differ systematically across students with different levels of measured heterogeneity.

Importantly, the empirical analysis considers multiple definitions of both achievement outcomes and heterogeneity measures. This enables us to examine whether estimated treatment heterogeneity depends on how educational achievement is measured.

2.2 Outcome construction and interaction effects

The framework implies that estimated treatment effects may vary across alternative educational outcome measures when educational interventions interact with heterogeneous student characteristics.

We evaluate this implication in two ways. First, we examine whether the interaction coefficient δ in equation (2) differs systematically across specifications. Second, we compare treatment-effect estimates obtained from equations (1) and (2). If these estimates vary across alternative outcome measures, this would suggest that conclusions regarding educational treatment heterogeneity depend importantly on how student achievement is defined and measured.

2.3 Composition effects

A further implication of the framework is that estimated treatment effects may vary across groups of students with different characteristics. To examine this possibility, we conduct subsample analyses based on the distribution of *Ability_i*.

Specifically, we partition the sample into low- and high-heterogeneity groups and estimate the baseline specification separately within each sub-sample:

$$Y_i = \alpha + \gamma^{(g)}T_i + \epsilon_i, \quad g \in \{Low, High\}, \quad (3)$$

where $\gamma^{(g)}$ denotes the estimated treatment effect within heterogeneity group g , and ϵ_i represents unobserved determinants of student achievement.

Comparing estimates of $\gamma^{(g)}$ across groups allows us to assess whether apparent treatment heterogeneity remains stable across alternative constructions of student heterogeneity. Systematic differences across specifications would suggest that conclusions regarding treatment heterogeneity depend on how student heterogeneity is measured.

2.4 Interpretation

Taken together, equations (1)–(3) provide a framework for examining whether conclusions regarding treatment heterogeneity remain stable across alternative educational outcome and heterogeneity measures.

Importantly, the analysis does not question the validity of random assignment itself. Rather, it examines how alternative definitions of educational achievement may influence the interpretation of heterogeneous treatment effects within the same randomized experiment.

3 DATA

3.1 Data source

The empirical analysis uses data derived from the Tennessee Student/Teacher Achievement Ratio (STAR) experiment, one of the most widely studied randomized evaluations in education. The original Project STAR randomly assigned students and teachers to small classes, regular classes, and regular classes with a teacher aide, allowing for causal identification of class size effects.

In this paper, we use a cleaned and publicly available version of the STAR dataset distributed through the Rdatasets repository. The dataset contains individual-level observations on students, including test scores, demographic characteristics, and class assignments across multiple grade levels.

The Project STAR experiment has been extensively studied in the literature on class-size effects [2] and [15]. More recently, the STAR data have also been used to examine treatment heterogeneity, long-run educational outcomes, and methodological issues in causal inference. In the present study, we use the STAR experiment as a laboratory for examining how alternative constructions of outcomes and heterogeneity measures influence the interpretation of heterogeneous treatment effects.

3.2 Sample construction

The analysis focuses on third-grade (grade 3) observations, where test score data are most complete and commonly used in the literature. We define the treatment indicator as assignment to a small class in grade 3. Observations with missing information on class assignment or test scores are excluded. The final sample consists of 5,967 students.

Focusing on grade 3 outcomes also facilitates comparison between contemporaneous and pre-treatment measures of student heterogeneity, which is central to the paper's examination of measurement sensitivity in heterogeneous treatment-effect estimation.

3.3 Outcome variables

Student performance is measured using alternative definitions of grade 3 standardized test scores. Specifically, we consider three specifications: (i) a composite outcome measure constructed as the average of reading and mathematics scores, (ii) mathematics scores as the outcome variable, and (iii) reading scores as the outcome variable.

Using multiple outcome definitions allows us to examine whether estimated treatment heterogeneity is sensitive to the construction and interpretation of empirical outcome measures.

3.4 Ability measure

To capture heterogeneity in student characteristics, we construct proxies for student heterogeneity using standardized reading and mathematics scores in grade 3, depending on the specification considered. The variables are standardized to have mean zero and unit variance.

Because the empirical analysis examines alternative combinations of outcome variables and heterogeneity measures, the ability proxy is not interpreted as a direct measure of innate ability. Instead, it serves as an empirical proxy for latent student heterogeneity that may interact with treatment assignment.

While the baseline specifications use contemporaneous achievement measures to construct heterogeneity proxies, subsequent sensitivity analyses employ pre-treatment achievement measures observed in kindergarten, grade 1, and grade 2. This allows us to evaluate whether estimated treatment heterogeneity remains robust across alternative definitions, domains, and timing of heterogeneity measures.

Accordingly, the purpose of the empirical analysis is not to establish a causal interpretation of the heterogeneity proxies themselves, but rather to examine how alternative constructions of outcomes and heterogeneity measures influence the estimation and interpretation of heterogeneous treatment effects within a randomized experiment.

3.5 Summary statistics

Table 1 reports summary statistics for the main variables used in the analysis, including reading and mathematics test scores, class assignment, and the standardized heterogeneity measures constructed from student test scores. The distribution of student outcomes and treatment assignment is broadly consistent with existing studies using the Project STAR experiment.

Table 1. Summary statistics

Variable	Mean	SD	Min	Max
Reading score	615.52	38.54	499.00	775.00
Mathematics score	617.85	39.89	487.00	774.00
Small class	0.32	0.47	0.00	1.00
Reading heterogeneity	0.00	1.00	−3.02	4.14
Mathematics heterogeneity	0.00	1.00	−3.28	3.91

Notes: Reading and mathematics scores are grade 3 standardized test scores from Project STAR. Small class is an indicator variable equal to one if the student is assigned to a small class. Reading and mathematics heterogeneity measures are standardized test-score indices.

4 RESULTS

4.1 Baseline estimates

Table 2 reports baseline estimates under alternative outcome definitions. Across all specifications, assignment to a small class has a positive and statistically significant effect on student performance. The estimated treatment effect ranges from approximately 7.32 to 8.43 points depending on the outcome definition.

At the same time, the variation across specifications suggests that estimated treatment effects are sensitive to how student performance is measured. While the conventional baseline estimates remain broadly consistent with the existing Project STAR literature, the differences across outcome definitions motivate the subsequent analysis of treatment heterogeneity and outcome construction.

Table 2. Baseline estimates under alternative outcome definitions

	Composite Outcome	Mathematics Outcome	Reading Outcome
Small class	7.771***	7.320***	8.428***
	(1.043)	(1.111)	(1.079)
Constant	614.172***	615.509***	612.835***
	(0.575)	(0.620)	(0.595)
Observations	5,967	5,967	5,967
Adj. R^2	0.009	0.007	0.01

Notes: The dependent variable is alternatively defined as (i) the composite average of reading and mathematics scores, (ii) mathematics scores, and (iii) reading scores in grade 3. Small class is an indicator variable equal to one if the student is assigned to a small class. Heteroskedasticity-robust standard errors are reported in parentheses. *** $p < 0.01$.

4.2 Interaction with student heterogeneity

Table 3 presents the interaction specifications under alternative definitions of outcome variables and heterogeneity measures. The estimated interaction patterns differ substantially across specifications.

Under the composite outcome specification, the interaction term between small class assignment and student heterogeneity is positive and marginally significant, indicating that estimated treatment effects increase with student heterogeneity. A similar positive interaction pattern is observed under the reading outcome specification. By contrast, when mathematics scores are used as the outcome variable, the interaction coefficient becomes substantially smaller and statistically insignificant.

These results suggest that empirical evidence on treatment heterogeneity is sensitive to how outcomes and heterogeneity measures are defined. More broadly, the findings indicate that estimated interaction effects are not invariant to alternative measurement constructions even under randomized assignment.

Table 3. Interaction between small classes and student heterogeneity

	(1) Composite	(2) Mathematics	(3) Reading
Small Class (γ_0)	2.726***	0.794	3.089***
	(0.686)	(0.733)	(0.709)
Student Heterogeneity	27.928***	29.577***	28.097***
	(0.433)	(0.443)	(0.451)
Small Class $\times$ Heterogeneity (δ)	1.388*	0.412	1.444*
	(0.729)	(0.760)	(0.747)
Observations	5,967	5,967	5,967
Adjusted R^2	0.553	0.556	0.557

Notes: The dependent variable is alternatively defined as (i) the composite average of reading and mathematics scores, (ii) mathematics scores, and (iii) reading scores. Student heterogeneity is measured using standardized scores corresponding to each specification. Heteroskedasticity-robust standard errors are reported in parentheses. The interaction coefficients provide preliminary evidence that estimated treatment effects may vary with measured student heterogeneity. However, the interpretation of these interaction effects should be treated with caution because the heterogeneity measures are contemporaneous with the outcome variables. Subsequent sensitivity analyses examine the robustness of these findings using alternative pre-treatment measures of student heterogeneity. *** $p < 0.01$, * $p < 0.1$.

4.3 Interpretation and measurement considerations

The differences in estimated treatment heterogeneity across alternative outcome specifications highlight the importance of measurement choices in empirical evaluations of educational interventions. While some specifications generate positive interaction effects between small-class assignment and measured student heterogeneity, these patterns are not consistently observed across alternative outcome definitions, achievement domains, or heterogeneity measures.

One possible interpretation is that different achievement measures capture distinct dimensions of student learning and therefore map educational interventions into observed outcomes in different ways. Reading and mathematics assessments may reflect different cognitive skills, learning processes, and responses to classroom environments. As a result, estimated interaction effects may vary even when the underlying treatment assignment remains unchanged.

Importantly, the additional sensitivity analyses using pre-treatment achievement measures indicate that the observed interaction effects are not robust across alternative constructions of student heterogeneity. In several specifications, interaction effects that appear under contemporaneous achievement measures become substantially weaker or disappear entirely when pre-treatment measures are employed. This finding suggests that conclusions regarding treatment heterogeneity depend importantly on how student ability is defined and measured.

Therefore, the results should not be interpreted as definitive evidence of skill complementarity or differential treatment responsiveness among high-ability students. Rather, they illustrate that empirical evidence on treatment heterogeneity may be sensitive to measurement choices, including the timing, domain, and construction of both outcome and heterogeneity measures. From this perspective, the findings reinforce the importance of conducting robustness analyses before drawing strong conclusions regarding heterogeneous treatment effects in educational and policy evaluation settings.

4.4 Composition effects across student heterogeneity groups

Table 4. Composition effects under alternative outcome definitions

	Composite		Mathematics		Reading	
	Low	High	Low	High	Low	High
Small class	3.176** (1.243)	6.256*** (1.209)	1.394 (1.265)	4.203*** (1.270)	3.728*** (1.172)	6.625*** (1.225)
Constant	591.787*** (0.653)	636.557*** (0.701)	593.434*** (0.666)	639.360*** (0.761)	590.525*** (0.630)	636.053*** (0.722)
Observations	2,917	3,050	2,920	3,047	2,914	3,053
Adj. R^2	0.002	0.008	0.000	0.003	0.003	0.009

Notes: This table reports estimated treatment effects for sub-samples categorized by low and high levels of student heterogeneity. Heteroskedasticity-robust standard errors are reported in parentheses. While treatment effects are generally larger among students classified in the high-heterogeneity group, the interpretation of these differences depends on how student heterogeneity is measured. As discussed in the subsequent sensitivity analyses, the evidence for treatment heterogeneity becomes substantially weaker when pre-treatment measures of student ability are used. *** $p < 0.01$, ** $p < 0.05$.

Table 4 reports composition effects across student heterogeneity groups under alternative outcome definitions. Across all specifications, the estimated effect of small class assignment is consistently larger for students in the high-heterogeneity group than for those in the low-heterogeneity group.

In particular, for composite and reading scores, treatment effects appear larger among students classified in the high-heterogeneity group. However, as shown in the subsequent sensitivity analyses, these patterns depend importantly on how student heterogeneity is measured. When pre-treatment achievement measures are used, the evidence for treatment heterogeneity becomes substantially weaker and generally loses statistical significance.

By contrast, under the mathematics specification, treatment effects are smaller and less precisely estimated. Taken together, these results suggest that apparent differences in treatment responsiveness may be sensitive to outcome construction and measurement choices rather than reflecting stable underlying heterogeneity.

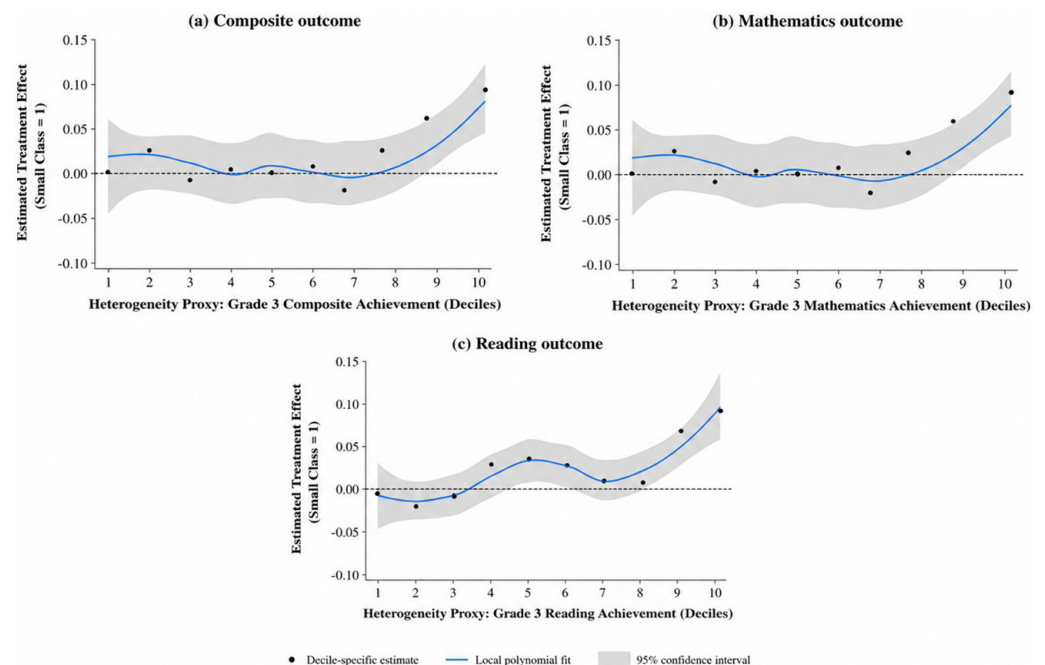


Fig. 1. Estimated treatment effects under alternative outcome definitions using grade 3 achievement as the heterogeneity proxy

Each panel plots estimated treatment effects (small class = 1) against deciles of the heterogeneity proxy measured in Grade 3.

(a) Composite outcome: average of reading and mathematics scores.

(b) Mathematics outcome: standardized mathematics score.

(c) Reading outcome: standardized reading score.

Notes: The dots represent estimation results for each decile of the heterogeneity proxy. The blue line shows a local polynomial fit, and the grey shaded area indicates 95% confidence intervals clustered at the classroom level. All models include baseline covariates as described in Section 2.

Figure 1 provides graphical illustrations of estimated treatment effects across student heterogeneity deciles under alternative outcome definitions. The estimated relationship between student heterogeneity and treatment effects differs substantially across specifications.

Under the composite outcome specification shown in panel (a), estimated treatment effects tend to increase with measured student heterogeneity. However, the corresponding relationship appears considerably weaker under the mathematics

specification in panel (b). A positive association is again observed under the reading specification in panel (c), although the magnitude and statistical significance of these patterns depend on how heterogeneity is measured. Taken together, the figures illustrate that estimated treatment heterogeneity is sensitive to outcome construction and measurement choices.

Overall, the figures suggest that empirical evidence on treatment heterogeneity depends importantly on how outcome variables and heterogeneity measures are constructed and interpreted.

4.5 Sensitivity to the timing and construction of heterogeneity measures

A central concern is that the heterogeneity measures used in the baseline analysis are constructed from contemporaneous grade 3 achievement scores, which may themselves be affected by treatment assignment. To address this concern, we conduct a series of additional analyses using achievement measures observed prior to the grade 3 outcome period.

Specifically, we reconstruct the heterogeneity variable using kindergarten, grade 1, and grade 2 test scores in both reading and mathematics. We then re-estimate the interaction specification across alternative combinations of outcome variables, achievement domains, and measurement periods.

The results reveal a striking pattern. The positive interaction effects reported in the baseline specifications become substantially weaker when pre-treatment achievement measures are used. Across most specifications, the interaction between class-size treatment and student heterogeneity is statistically insignificant. This finding holds for both reading and mathematics outcomes and remains robust when heterogeneity measures are constructed from either same-domain or cross-domain achievement scores.

We further examine whether the estimated interaction effects vary according to the timing of outcome measurement. Using kindergarten achievement as the heterogeneity measure, we estimate treatment heterogeneity separately for grade 1, 2, and 3 outcomes. The resulting interaction coefficients remain unstable across specifications and generally fail to achieve conventional levels of statistical significance.

Taken together, these findings suggest that the evidence for treatment heterogeneity is highly sensitive to the timing and construction of heterogeneity measures. While contemporaneous achievement measures may generate apparent interaction effects, these patterns largely disappear when heterogeneity is measured using pre-treatment achievement. The results therefore indicate that conclusions regarding differential treatment effects should be interpreted with caution, as estimated heterogeneity may depend importantly on measurement choices rather than reflecting stable underlying differences in treatment responsiveness.

Table 5 reports the results obtained using alternative pre-treatment measures of student heterogeneity. Across all specifications, the estimated interaction coefficients are considerably smaller and less stable than those reported in the baseline analysis. In most cases, the interaction terms are statistically insignificant and vary substantially across alternative achievement domains and measurement periods. These findings indicate that the apparent treatment heterogeneity observed under contemporaneous achievement measures is not robust to alternative constructions of student ability.

Overall, the results suggest that conclusions regarding differential treatment effects depend importantly on how student heterogeneity is measured.

While some specifications produce positive interaction coefficients, the estimated effects are neither statistically robust nor consistent across alternative outcome and ability measures.

Table 5. Sensitivity of estimated treatment heterogeneity to alternative ability measures

Outcome	Ability Measure	Interaction	Std. Error	<i>p</i> -Value
Read3	ReadK	0.251	1.277	0.844
Read3	Read1	0.169	1.041	0.871
Read3	Read2	−0.563	0.885	0.525
Read3	MathK	1.034	1.467	0.481
Read3	Math1	−0.017	1.182	0.989
Read3	Math2	−0.418	0.987	0.672
Math3	MathK	2.013	1.520	0.186
Math3	Math1	0.331	1.199	0.783
Math3	Math2	−0.990	0.999	0.322
Math3	ReadK	0.797	1.502	0.595
Math3	Read1	0.772	1.190	0.516
Math3	Read2	−0.525	1.063	0.622

Notes: The table reports the coefficient on the interaction term between small-class assignment and student heterogeneity. Student heterogeneity is alternatively measured using reading and mathematics achievement scores observed at kindergarten (K), grade 1, and grade 2. None of the interaction coefficients remains statistically significant at the 5% level. The results indicate that estimated treatment heterogeneity is highly sensitive to the timing, domain, and construction of ability measures.

5 CONCLUSION

This paper has examined whether evidence of treatment heterogeneity remains robust to alternative constructions of outcome and heterogeneity measures within a randomized educational experiment. Using data from the Project STAR experiment, we investigated how estimates of treatment heterogeneity vary across different definitions of student achievement and ability.

The baseline analysis suggested the presence of positive interaction effects between small-class assignment and measured student ability. However, additional analyses revealed that these interaction effects are highly sensitive to the timing and construction of ability measures. When contemporaneous grade 3 achievement scores are used to measure heterogeneity, positive interaction effects emerge in several specifications. In contrast, when heterogeneity is measured using pre-treatment achievement scores from kindergarten, grade 1, or 2, the estimated interaction effects become substantially weaker and generally lose statistical significance. Similar patterns arise across reading and mathematics outcomes as well as across alternative cross-domain specifications.

These findings suggest that evidence of treatment heterogeneity may not be as robust as implied by baseline specifications alone. While randomized assignment provides credible identification of average treatment effects, conclusions regarding heterogeneous treatment effects can depend importantly on how student ability is defined and measured. In this sense, the interpretation of treatment heterogeneity involves not only identification issues but also measurement issues.

More broadly, the paper contributes to the growing literature on heterogeneous treatment effects and policy evaluation. Recent methodological advances have enabled researchers to estimate increasingly complex forms of treatment heterogeneity. Our results suggest that careful attention should also be paid to the construction of outcome and heterogeneity measures, as alternative measurement choices may generate substantially different empirical conclusions even within the same randomized experiment.

Several limitations should be acknowledged. The analysis focuses on a single educational intervention and a specific set of achievement measures. Future research could examine whether similar patterns arise in other experimental settings, alternative policy interventions, and more flexible approaches to measuring latent ability and student heterogeneity.

Taken together, the findings suggest that measurement choices may play a central role in the interpretation of heterogeneous treatment effects, even within well-designed randomized experiments. Although the empirical analysis is based on the Project STAR experiment, the underlying issue is not specific to education. Whenever treatment heterogeneity is assessed using constructed outcome measures or proxy variables for latent characteristics, alternative measurement choices may generate substantially different conclusions. These findings underscore the importance of evaluating measurement sensitivity as part of empirical data analysis before drawing substantive conclusions about heterogeneous treatment effects.

6 CONFLICT OF INTEREST

The authors declare no conflict of interest.

7 DATA AVAILABILITY

The data used in this study are publicly available from the Rdatasets repository: <https://vincentarelbundock.github.io/Rdatasets/>

8 FUNDING

This work was supported by JSPS KAKENHI Grant Number 25K05043.

9 REFERENCES

- [1] G. W. Imbens and J. D. Angrist, "Identification and estimation of local average treatment effects," *Econometrica*, vol. 62, no. 2, pp. 467–475, 1994. <https://doi.org/10.2307/2951620>
- [2] A. B. Krueger, "Experimental estimates of education production functions," *Quarterly Journal of Economics*, vol. 114, no. 2, pp. 497–532, 1999. <https://doi.org/10.1162/003355399556052>
- [3] J. J. Heckman, J. A. Smith, and N. Clements, "Making the most out of programme evaluations and social experiments," *Review of Economic Studies*, vol. 64, no. 4, pp. 487–535, 1997. <https://doi.org/10.2307/2971729>
- [4] E. Duflo, R. Hanna, and S. P. Ryan, "Incentives work: Getting teachers to come to school," *American Economic Review*, vol. 102, no. 4, pp. 1241–1278, 2012. <https://doi.org/10.1257/aer.102.4.1241>

- [5] S. Athey and G. W. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7353–7360, 2016. <https://doi.org/10.1073/pnas.1510489113>
- [6] S. Athey and G. W. Imbens, "The state of applied econometrics: Causality and policy evaluation," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 3–32, 2017. <https://doi.org/10.1257/jep.31.2.3>
- [7] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018. <https://doi.org/10.1080/01621459.2017.1319839>
- [8] T. N. Bond and K. Lang, "The sad truth about happiness scales," *Journal of Political Economy*, vol. 27, pp. 1629–1640, 2019. <https://doi.org/10.1086/701679>
- [9] M. L. Anderson, "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1481–1495, 2008. <https://doi.org/10.1198/016214508000000841>
- [10] C. E. Frangakis and D. B. Rubin, "Principal stratification in causal inference," *Biometrics*, vol. 58, no. 1, pp. 21–29, 2002. <https://doi.org/10.1111/j.0006-341X.2002.00021.x>
- [11] H. C. Kraemer, G. T. Wilson, C. G. Fairburn, and W. S. Agras, "Mediators and moderators of treatment effects in randomized clinical trials," *Archives of General Psychiatry*, vol. 59, no. 10, pp. 877–883, 2002. <https://doi.org/10.1001/archpsyc.59.10.877>
- [12] W. Lin, "Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique," *Annals of Applied Statistics*, vol. 7, no. 1, pp. 295–318, 2013. <https://doi.org/10.1214/12-AOAS583>
- [13] R. Chetty, J. N. Friedman, N. Hilger, E. Saez, D. W. Schanzenbach, and D. Yagan, "How does your kindergarten classroom affect your earnings? evidence from Project STAR," *Quarterly Journal of Economics*, vol. 126, no. 4, pp. 1593–1660, 2011. <https://doi.org/10.1093/qje/qjr041>
- [14] S. Dynarski, J. Hyman, and D. W. Schanzenbach, "Experimental evidence on the effect of childhood investments on postsecondary attainment and degree completion," *Journal of Policy Analysis and Management*, vol. 32, no. 4, pp. 692–717, 2013. <https://doi.org/10.1002/pam.21715>
- [15] J. D. Angrist and V. Lavy, "Using Maimonides' rule to estimate the effect of class size on scholastic achievement," *Quarterly Journal of Economics*, vol. 114, no. 2, pp. 533–575, 1999. <https://doi.org/10.1162/003355399556061>

10 AUTHORS

Mika Takechi is an Associate Professor of English education at Yamato University. Her research interests include English language pedagogy, second language acquisition, and the economics of education (E-mail: takechi.mika@yamato-u.ac.jp).

Shigeyuki Hamori is a Professor of Economics at Yamato University and Professor Emeritus at Kobe University. His research interests include financial econometrics, connectedness analysis, systemic risk, and applied time-series econometrics (E-mail: hamori@econ.kobe-u.ac.jp).