High dosage tutoring in pre-vocational secondary education: Experimental evidence from Amsterdam*

Joppe de Ree† Mario A. Maggioni‡ Bowen Paulle§
Domenico Rossignoli¶ Dawid Walentek‖

March 11, 2021

Abstract

We present first results from an experimental evaluation of a high dosage math tutoring program implemented in a secondary school for pre-vocational education located in a low-income neighborhood in Amsterdam, the Netherlands. From a pool of 98 students in their first year of secondary school (aged 12-13) 49 were randomly assigned to receive 2-on-1 tutoring. Treated students received one class period of personalized tutoring daily in the first 16 weeks of the school year. Outcome test scores were observed at the end of the first semester and at the end of the school year. We find treatment effects of 0.44 ($p < 0.01$) and 0.72 ($p < 0.01$) control group standard deviations after one semester on a verbal math test (with math word problems) and a nonverbal math test (math without textual context) respectively. The results indicate substantial learning gains. On the nonverbal math test, treated students gained statistically significantly more in one semester than control students gained in the entire school year.
JEL: H0, I20, I24, I26, J24

*For their various types of contributions to this paper, the authors would like to thank the following individuals and organizations: Students, teachers and other staff of the school that participated in the research project, Amsterdam University Fund, International Marketmaker's Combination (IMC), The Bridge Learning Interventions Foundation, Saga Education, Mike Goldstein, Jonathan Guryan, participants of the sociology seminar and the QMSS seminar at the University of Amsterdam and all the research assistants and others – including Isabel Speelman, Shelby Sissing, David van der Duin – who supported our team.

†Erasmus University Rotterdam. Corresponding author: joppederee@gmail.com
‡Università Cattolica Milano
§University of Amsterdam
¶Università Cattolica Milano
‖University of Warsaw

1 Introduction

Early tracking is a key feature of many (European) education systems. In the Netherlands, 12 year olds are assigned to different educational programs or tracks depending to their perceived abilities (determined in part by standardized achievement tests). Tracking can be effective as teachers can target instruction to a more homogeneous group.[1] However, the practice of early tracking is controversial. Critics highlight that it can reduce peer effects, generate stigma and fuel a culture of low expectations for students assigned to the lower tracks.[2] This is particularly problematic if track assignment does not reflect a student's true potential. Students who are tracked too low are held back by a slower pace of learning in these lower tracks. This problem is likely exacerbated when students remain tracked below their potential for multiple years.[3]

This research aims at measuring *unrealized potential* in lower pre-vocational tracks in Dutch secondary education, especially in low-income (urban) areas. This paper presents experimental evidence from one of the first high dosage tutoring (HDT) interventions in secondary education carried out outside of the United States. The HDT intervention was developed and implemented by The Bridge Learning Interventions, henceforth The Bridge, a nonprofit organization in the Netherlands.[4] The program aimed at substantially improving the math skills of students attending a secondary school for lower and mid-level pre-vocational education, located in the low-income area of *Nieuw-West* in Amsterdam. Student who start secondary education in this school at age 12 are typically substantially behind average grade level.

[1] See for example Duflo, Dupas, and Kremer (2011).

[2] See for example Berends (1995), Bram Spruyt and Kavadias (2015), Van Houtte (2016) and Garlick (2018).

[3] It is plausible that the potentially harmful effects of early educational tracking dis-proportionally affect low-income groups. See for example Hanushek and Woessman (2006), Dronkers and Korthals (2016) and Sund (2013)

[4] The Bridge aims to reduce inequalities of opportunity by supporting disadvantaged youth through high dosage tutoring interventions (`https://tbli.nl/`).

The Bridge HDT was modeled after the successful Saga Education program.[5] Saga Education and other tutoring initiatives have shown promising results, both in terms of treatment effects and for potential scalability.[6] The HDT program was implemented for most of the first semester of the school year (approximately 16 weeks) and consisted of daily sessions lasting 50 minutes.[7] Before the start of the program, selected students were assigned (nonrandomly) to small groups of two students and one tutor. These small groups (or teams) would remain together for the duration of the program. Tutoring took place during regular school hours and selected students would miss one instructional hour of their other courses (excluding regular math) each day.[8] Student achievement was assessed at three points in time with two standardized computer-based tests for math and one for reading comprehension. The two math tests we used differ in their reliance on language. We used a verbal math test which relies heavily on math word problems and a nonverbal math test which is without any textual context. Otherwise the difficulty levels of these tests are similar.

We find highly significant treatment effects of 0.72 control group standard deviations for the nonverbal math test and 0.44 control group standard deviations for the verbal math test after 1 semester. The difference in magnitude between the two treatment effects is mostly driven by the portion (1/3) of students in our sample who are assigned to the lowest pre-vocational track, i.e. *Praktijkonderwijs* or practical education. For this group we find strong treatment effects on the nonverbal test, but not on the verbal test. This result suggests that the language used in the verbal math test provides a barrier for a sizable

[5]See Guryan, Ludwig, Bhatt, Cook, Davis, Dodge, Farkas, Fryer Jr., Mayer, Pollack, and Steinberg (2021)

[6]See e.g. Guryan, Ludwig, Bhatt, Cook, Davis, Dodge, Farkas, Fryer Jr., Mayer, Pollack, and Steinberg (2021) and Kraft (2015), as well as Nickow, Oreopoulos, and Quan (2020) and Pellegrini, Lake, Neitzel, and Slavin (2021) for overviews of experimental research on tutoring interventions. Recently, Carlana and La Ferrara (2021) find effects of an online tutoring intervention during school lockdowns in Italy, in March and April of 2020.

[7]The typical duration of the Bridge HDT program is one school year, but due to budgetary constraints The Bridge had to experiment with a half year model.

[8]To avoid students falling behind in a particular domain, these courses were rotated.

percentage of our experimental sample.

We observe modest test score decay after the program ended. Between the end of the program (at the end of the first semester) and the end of the school year, the treatment effects reduced from 0.72 to 0.51 and from 0.44 to 0.28 for the nonverbal math test and the verbal math test respectively.[9] To assess positive or negative spillover effects into other domains we also measure the effects on reading comprehension. The results on the reading comprehension test are positive, albeit smaller, and statistically insignificant. The statistical tests on the reading scores however are not sufficiently powered to detect reasonable effect sizes.

Overall the results suggest substantial improvements in achievement after 16 weeks of intensive tutoring. On the nonverbal math test, treated students gained statistically significantly more in 1 semester than control students in an entire school year ($p < 0.01$). This result demonstrates that this group of low-achieving students has the potential to learn significantly more than usual in a specified amount of time. The HDT program reduces the gap between what these students need to thrive, and the educational opportunities that are typically afforded to them.

The rest of the paper proceeds as follows. Section 2 briefly describes the program's main characteristics. Section 3 outlines the research design. Section 4 presents the results of the analysis and Section 5 concludes.

2 The high dosage tutoring intervention

The HDT program of The Bridge was modeled after the Saga Education HDT program. Saga Education is one of the primary service providers for HDT in the US.[10] The Bridge has worked closely with Saga Education in the process of developing The Bridge HDT. There

[9] These amount to a rate of decay of about 30% per semester.

[10] `https://www.sagaeducation.org/our-story`

are, however, no formal ties between Saga Education and The Bridge. HDT programs were developed as a way to address Bloom's "2 sigma problem" (Bloom 1984). Bloom describes experimental results in which students subjected to small group tutoring interventions outperformed a control group in a regular classroom by 2 standard deviations (i.e. 2 sigma's). Bloom concluded that small group instruction was highly effective, but also very costly. He raised the question of whether (and how) it would be possible to devise a mode of instruction that could be as effective as 1-on-1 tutoring, but that would be cheaper.

The Bridge HDT is also intensive, but it limits costs by offering small group tutoring for (only) 50 minutes (one class period) each day, and by relying on paraprofessional tutors. Nickow, Oreopoulos, and Quan (2020) refer to paraprofessionals in the context of tutoring interventions as paid professionals, but who are not certified teachers. In the beginning of the school year selected students were allocated (nonrandomly) to one tutor and one other student. The formation of these groups was done in coordination with teachers and based on perceived fit, which depended in part on baseline math performance. These groups would remain together for the duration of the program. The stability of this environment is meant to foster a relationship of trust between students and tutors, in which students feel comfortable making mistakes and trying out new things. The tutors deliver personalized instruction using principles associated with mastery learning. Also, in this context, successes are observed, acknowledged and explicitly celebrated. The Bridge HDT therefore has a strong focus on the socioemotional aspects of learning (see Kosse, Deckers, Pinger, Schildberg-Hörisch, and Falk (2020) who provide evidence on the influence of the social environment). Socioemotional, or noncognitive skills, like confidence and motivation, are important factors in explaining success in school and in life more generally (Jackson 2018).

The HDT program we evaluate worked with 5 full-time tutors and 1 full-time so-called *site director* or program manager. The site director plays a pivotal role in the process.

He or she trains and prepares tutors before the start of program and continues to support and mentor the tutors on a daily basis once the program starts. The site director also maintains communication with teachers and the school administration. Tutors, alongside their tutoring duties, also build and maintain relationships with parents through weekly phone calls.

The HDT program was implemented during regular school hours. Therefore, treated students would miss one instructional hour of their other courses each day. To avoid students falling behind in a particular domain, these courses were rotated. It was agreed that students would not miss any regular math classes. The HDT intervention therefore also increased the time students spent on math instruction and practice. A typical The Bridge HDT program would last a full school year, but due to budgetary constraints The Bridge had to experiment with a half year program, implemented from September 2017 to January 2018.

3 The experimental design

3.1 Experimental sample and randomization

All students in their first year of secondary school from 7 (of 8 total) classes were eligible for participation. Parents of the target group were asked to sign a consent form allowing students to participate in the project. $98/112 = 88\%$ of parents consented, indicating that parents and students were generally enthusiastic about the program.[11]

The 98 students who consented were randomly assigned (in September 2017) to treatment or control through a block randomization procedure. We ranked students within each class based on a baseline math achievement test score. Based on this ranking we grouped students in pairs, from the lowest-performing pair of students to the highest-performing,

[11] We do not find that baseline performance was predictive of giving consent.

within each class. The computer randomly assigned one student of each pair to treatment and one to control. This block randomization procedure is more efficient than simple randomization if the group identifiers are good predictors of the outcome variables (Bruhn and McKenzie 2009). In Appendix A we present baseline characteristics of the experimental sample.

3.2 Outcome variables

Our main outcome variables are scores on two standardized math tests, a verbal math test and a nonverbal math test. These tests are developed by test developer ICE in the Netherlands.[12] The tests are known in the Netherlands as TOA tests. The former (verbal) TOA test is widely used in the Netherlands to assess school performance of secondary school students, including in our school.[13] The latter (nonverbal) math test is not routinely used by schools but was added as an additional source because our sample included students who recently immigrated to the Netherlands (and who might speak very little Dutch) and students who were multiple years behind (average) grade level. We expected that the language used in the verbal math test could be an obstacle for some of the students in our sample. The inclusion of the nonverbal test would purge results of possible biases arising from lacking comprehension of textual instructions, and instead only measure math skills.

We also included the TOA reading comprehension test as one of the outcomes to verify any potential unintended negative side-effects of the intervention on non-math related achievement. On the other hand, the HDT program could also benefit students in other domains (such as reading comprehension) through improved motivation and self-confidence. We did not have a clear prior conceptualization about the sign of effect (if any) of HDT program on the reading comprehension test.

[12] `https://www.bureau-ice.nl/contact/`

[13] See e.g. `https://www.bureau-ice.nl/voortgezet-onderwijs/`.

To measure skills in the socioemotional domain, we have also collected some indicators of socioemotional characteristics or skills. We are in the process of analyzing this data for the next draft of this paper.

3.3 Statistical model

We estimate treatment effects based on the following regression model:

$$Y = a + b \times T + \sum_{k=1}^{K} \gamma_k BLOCK_k + \varepsilon \quad (1)$$

where T indicates students who are randomly assigned to receive treatment. The parameter b is the parameter of interest and measures the treatment effect. Y is the standardized achievement score. We standardize by subtracting the mean and dividing by the standard deviation of the control group.[14] $BLOCK_k$ indicates the K different sampling blocks. In other words, we estimate the treatment effects based on a block fixed effects regression model. All models are estimated with OLS.

3.4 Testing for random attrition

For some students, we do not observe a full set of outcome test scores. In Table 1 we test whether attrition is differentially affected by treatment assignment. We find relatively small attrition rates across treatment and control groups and no statistically significant difference between them. These results do not suggest a threat to a common interpretation of our findings.

[14]Therefore, our estimated treatment effects are expressed in control group standard deviation units. This concept is also known as Glass' Δ.

Table 1: Tests on random attrition

	(1) Math v. ML	(2) Math v. EL	(1) Math n.v. ML	(2) Math n.v. EL	(1) Reading ML	(2) Reading EL
Treatment group	0.00	0.04	0.00	0.02	0.00	0.06
	(0.05)	(0.05)	(0.05)	(0.05)	(0.06)	(0.05)
	[1.00]	[0.42]	[1.00]	[0.71]	[1.00]	[0.26]
	<0.06>	<0.08>	<0.06>	<0.08>	<0.10>	<0.08>

Notes: The table reports parameter estimates of a block fixed effects regression of a dummy variable that is 1 for a missing test score, on a treatment dummy. *,**,*** indicate statistical significance at the 10, 5 and 1% level. Robust standard deviations reported in parentheses. p-values reported in brackets. Control group means are presented in <>. Math v. refers to the verbal math test. Math n.v. refers to the verbal math test. ML refers to midline (at the end of the first semester). EL refers to endline (at the end of the school year).

4 Results

In Table 2 we present the estimated treatment effects.[15] The estimates for both math tests are sizable and highly significant (with t statistics of about 3 for tests based on the verbal math test and about 6 for tests based on the nonverbal math test). We do not find any negative (side) effect on reading comprehension. If anything, the effects on reading comprehension seem positive. However, the parameter is not statistically significant from zero and somewhat imprecisely estimated.

Table 2: Treatment effects measured after 1 semester (directly after the end of the HDT program) [column (1)] and at the end of the 2017/18 school year [column (2)].

	(1) Mid-year (end of intervention)	(2) End of school year
(Regular) verbal math test	0.44***	0.28**
	(0.15)	(0.12)
	[92]	[88]
Nonverbal math test	0.72***	0.51***
	(0.12)	(0.16)
	[92]	[89]
(Regular) reading comprehension test	0.17	0.33
	(0.19)	(0.23)
	[88]	[87]

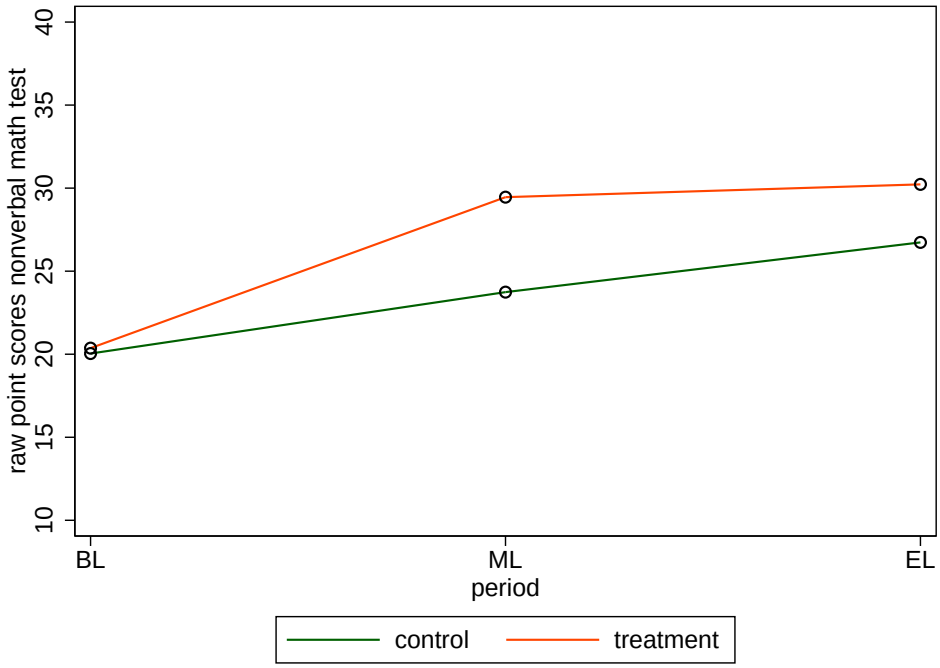
Notes: *,**,*** indicate statistical significance at the 10, 5 and 1% level. Robust standard deviations reported in parentheses. Sample size used in estimation reported in brackets.

These treatment effects are large compared to other reported treatment effects in educa-

[15] Shortly after the project started, one of the selected students no longer wanted to participate. We do not account for this in Table 2 and therefore these effects must be (formally) interpreted as intention to treat effects. As we did not further observe any dropout explicitly linked to the program, we conclude that the intent to treat effects are close estimates of the average treatment effects of the HDT program on treated students.

tion research (see Kraft (2020) for context on the interpretation of effect sizes in education research). The HDT program had a substantial impact on math performance. The estimated treatment effects based on the nonverbal math test are particularly large. Figure 1 below provides additional context to the interpretation of this result. Treated students gained 9 raw score points in the first semester, while control students only gained 6 points on average in an entire school year. The difference between this is statistically significant at the 1% level. This result clearly highlights the relevance of The Bridge HDT in this context.

Figure 1: Graphical representation of treatment effects by period, on nonverbal math scores



Notes: BL refers to baseline (the beginning of the school year), ML refers to midline (after one semester, and after the HDT intervention had ended) and EL refers to endline (the end of the school year)

An exploratory analysis in Table 3 shows that the difference in the magnitude of the treatment effects measured with the two different math tests is mostly driven by students (about 30% of our data) from the lowest pre-vocational track, i.e *praktijkonderwijs* or

'practical education'. This group typically scores between 0.6 and 0.9 standard deviations below the rest of our sample on average.[16] Table 3 also shows that treated students in *praktijkonderwijs* essentially close the gap with the control group of the rest of the sample when the assessment was based on the nonverbal math test[17], while this gap remains large when we look at the verbal math test[18]. The (relative) complexity of the language used in the verbal math test might prevent students from being able to show the mathematics they have learned. This exploratory finding might have broader implications in the Netherlands and elsewhere. For example, most tests used in Dutch primary education (as a basis for secondary school track assignment) are heavily reliant on math word problems.

[16]Control group mean estimates for the verbal math and the nonverbal math test are reported in Table 3. After the first semester (column [1]), control group students in *praktijkonderwijs* are 0.85 (full) control group standard deviations behind on the nonverbal math test and 0.57 (full) control group standard deviations behind on the verbal math test.

[17]Treated students in *praktijkonderwijs* score $-0.40 + 0.56 = 0.16$ on the nonverbal math test. Control students from the rest of the sample score 0.17.

[18]Treated students in *praktijkonderwijs* score $-0.61 + 0.01 = -0.60$ on the verbal math test, while the control group students from the rest of the sample score 0.24. The difference between the two groups is 0.84 standard deviations

Table 3: Treatment effects on the regular verbal math test [Panel A] and on the nonverbal math test [Panel B], where we investigated heterogeneous effects by track type (the practical education track or other pre-vocational)

	(1) Mid-year (end of intervention)	(2) End of school year
A: Verbal math test		
Treatment (in other pre-vocational tracks)	0.62***	0.42***
	(0.18)	(0.15)
	[0.00]	[0.01]
	<0.24>	<0.34>
Treatment (in practical education track)	0.01	0.05
	(0.20)	(0.18)
	[0.97]	[0.80]
	<-0.61>	<-0.64>
B: Nonverbal math test		
Treatment (in other pre-vocational tracks)	0.80***	0.67***
	(0.14)	(0.22)
	[0.00]	[0.01]
	<0.17>	<0.22>
Treatment (in practical education track)	0.56***	0.24
	(0.17)	(0.18)
	[0.01]	[0.20]
	<-0.40>	<-0.54>

Notes: *,**,*** indicate statistical significance at the 10, 5 and 1% level. Robust standard deviations reported in parentheses. p-values reported in brackets. Control mean of subsamples reported in <>. Subsamples are defined by students in the practical education track and students in other pre-vocational tracks.

5 Conclusions

In this paper we present results of an experimental evaluation of a high dosage math tutoring program implemented by The Bridge Learning Interventions in a secondary school for pre-vocational education in the low-income *Nieuw-West* area in Amsterdam. We document large treatment effects of 0.44 and 0.72 control group standard deviations on a verbal and a nonverbal standardized math test respectively, after 16 weeks of personalized 2-on-1 tutoring.[19] Our findings demonstrate considerable unrealized potential among students assigned to the lowest tracks of the Dutch secondary education system. The HDT program can appreciably reduce the gap between what these low-achieving students need to thrive, and the educational opportunities that are typically afforded to them.

Despite that the HDT program is meant to be delivered over the course of an entire school year, the students in our sample made strong gains after just 16 weeks of personalized tutoring. Furthermore, as we have shown, most of these gains persisted half a school year after the program ended. This generates questions about the potential effects a full year of HDT in this context. In the coming years, we will continue to evaluate HDT interventions at both the primary and secondary educational levels in the Netherlands. We are also in the process of evaluating different HDT models (e.g. an entire year rather than half a year of tutoring and three hours of tutoring per week instead of five). We hope to clarify the degree to which the results presented here are generalizable and whether they can be achieved with more cost-effective delivery models.

[19] In this paper we discuss effects on math and reading test scores. As part of this project, however, we have also collected data on socioemotional characteristics through surveys and games. We are in the process of analyzing this data for a next version of this paper.

A Baseline characteristics

Table 4: Baseline descriptives, with treatment and control group comparison

	(1) Treatment	(2) Control	(3) Difference	(4) p-value
Age	12.37	12.23	0.15	0.24
Fraction in lowest pre-vocational track^{+}	0.33	0.33	0.00	
Raw baseline math test score	21.20	21.35	-0.14	0.71

Notes: The table provides baseline descriptive statistics. p-values in column (4) are based on robust block fixed effects regression. $^{+}$For the fraction enrolled in the lowest pre-vocational track, the practical education track (or *Praktijkonderwijs* in Dutch), the difference between treatment and control is zero by construction.

References

Berends, M. (1995): "Educational Stratification and Students Social Bonding to School," *Journal of Sociology*, 16(3).

Bloom, B. S. (1984): "The 2 sigma problem: The search for methods of group instruction as effective as one-on-one tutoring," *Educational Observer*, 13(6).

Bram Spruyt, F. V. D., and D. Kavadias (2015): "Educational tracking and sense of futility: a matter of stigma consciousness?," *Oxford Review of Education*, 41(6).

Bruhn, M., and D. McKenzie (2009): "In pursuit of balance: Randomization in practice in development field experiments," *American economic journal: applied economics*, 1(4), 200–232.

Carlana, M., and E. La Ferrara (2021): "Apart but Connected: Online Tutoring and Student Outcomes during the COVID-19 Pandemic," *HKS Faculty Research Working Paper Series*, 21(001).

Dronkers, J., and R. Korthals (2016): "Tracking in the Netherlands–Ability selection or social reproduction?," in *Models of Secondary Education and Social Inequality*. Edward Elgar Publishing.

Duflo, E., P. Dupas, and M. Kremer (2011): "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 101(5).

Garlick, R. (2018): "Academic Peer Effects with Different Group Assignment Policies: Residential Tracking versus Random Assignment," *American Economic Journal: Applied Economics*, 10(3).

Guryan, J., J. Ludwig, M. P. Bhatt, P. J. Cook, J. M. Davis, K. Dodge, G. Farkas, R. G. Fryer Jr., S. Mayer, H. Pollack, and L. Steinberg (2021): "Not Too Late: Improving Academic Outcomes Among Adolescents," *NBER working paper*, (28531).

Hanushek, E. A., and L. Woessman (2006): "Does Educational Tracking Affect Performance and Inequality? Differences inDifferences Evidence Across Countries," *Economic Journal*, 116(510).

Jackson, C. K. (2018): "What Do Test Scores Miss? The Importance of Teacher Effects on NonTest Score Outcomes," *Journal of Political Economy*, 126(5).

Kosse, F., T. Deckers, P. Pinger, H. Schildberg-Hörisch, and A. Falk (2020): "The Formation of Prosociality: Causal Evidence on the Role of Social Environment," *Journal of Political Economy*, 128(2).

Kraft, M. A. (2015): "How to Make Additional Time Matter: Integrating Individualized Tutorials into an Extended Day," *Education Finance and Policy*, 10(1).

——— (2020): "Interpreting Effect Sizes of Education Interventions," *Educational Researcher*, 49(4), 241–253.

Nickow, A., P. Oreopoulos, and V. Quan (2020): "The Impressive Effects of Tutoring on PreK-12 Learning: A systematic Review and Meta-Analysis of the Experimental Evidence," *NBER working paper series*, (27476).

Pellegrini, M., C. Lake, A. Neitzel, and R. E. Slavin (2021): "Effective Programs in Elementary Mathematics: A Meta-Analysis," *AERA Open*, 7(1).

Sund, K. (2013): "Detracking Swedish compulsory schools: any losers, any winners?," *Empirical Economics*, 44(2), 899–920.

Van Houtte, M. (2016): "Lower-Track Students Sense of Academic Futility: Selection or Effect?," *Journal of Sociology*, 52(4).