



HAL
open science

Epidemiological and cognitive evaluations in mathematics and language in the whole population of school-age children in France

Pauline Martinot

► **To cite this version:**

Pauline Martinot. Epidemiological and cognitive evaluations in mathematics and language in the whole population of school-age children in France. Education. Université Paris Cité, 2023. English. NNT : 2023UNIP7259 . tel-04715278

HAL Id: tel-04715278

<https://theses.hal.science/tel-04715278v1>

Submitted on 30 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Paris Cité

Ecole doctorale FIRE – ED 474

Laboratoire INSERM-CEA Cognitive Neuroimaging unit
CEA/SAC/JOLIOT/Neurospin

Epidemiological and cognitive evaluations in mathematics and language in the whole population of school-age children in France

Par **Pauline MARTINOT**

Thèse de doctorat de Neurosciences et Troubles neuroaux

Dirigée par **Ghislaine DEHAENE-LAMBERTZ**

Soutenue publiquement le 20 décembre 2023

Devant un jury composé de :

Ghislaine DEHAENE-LAMBERTZ, DR HDR, Université Paris Saclay,
directrice de thèse.

Marcela PEÑA, PR HDR, Université Catholique Pontificale du Chili,
rapporteuse.

Liliane SPRENGER-CHAROLLES, PR HDR Émérite, Université Aix-
Marseille, rapporteuse.

Ignacio ATAL, Chercheur, Université Paris Cité, examinateur.

Stanislas DEHAENE, PR HDR, Université Paris-Saclay, examinateur.

Caroline HURON, HDR, Université Paris Cité, examinatrice.

Raphael PORCHER, PUPH, Université Paris Cité, membre invité.

Résumé

Titre : Évaluations épidémiologiques et cognitives en mathématiques et en langage dans l'ensemble de la population des enfants d'âge scolaire en France

Résumé : Récemment, la France a mis en place des évaluations nationales visant à mesurer précisément les acquisitions et les difficultés d'apprentissage auxquelles les enfants sont confrontés tout au long de leur première année d'école jusqu'à leur deuxième année, en se basant sur l'évaluation cognitive des compétences en mathématiques et en langage. Chaque année, environ 750 000 enfants ont effectué 46 exercices pour évaluer leurs performances, couvrant un total de 2,9 millions d'enfants entre 2018 et 2022. En analysant cet ensemble riche de données sur l'ensemble de la population, cette thèse visait à fournir une meilleure compréhension des conditions qui favorisent ou entravent l'acquisition de l'apprentissage académique chez les enfants. Pour ce faire, nous avons mené une série d'études en utilisant des données longitudinales provenant de quatre cohortes françaises représentatives de la population, évaluant l'influence relative d'une large gamme de facteurs individuels, de classe et d'établissement sur différents aspects de la réussite scolaire au primaire. Tout d'abord, nous avons décrit les données obtenues dans le programme national et identifié les prédicteurs des compétences en lecture et en compréhension de la lecture. De plus, nous avons exploré le pouvoir prédictif des caractéristiques de l'enfant et des facteurs environnementaux, aux niveaux individuel, de classe et d'établissement. Grâce à l'expérience naturelle de la Covid-19 (c'est-à-dire, l'absence d'exposition à l'école pendant une période spécifique), nous avons pu estimer l'impact de l'exposition à l'école en comparant une cohorte à l'autre. Plus particulièrement, nous avons pu identifier les besoins d'apprentissage parmi les différentes catégories socio-économiques des écoles. Enfin, nous avons centré notre dernière analyse sur les différences de genre en langue et en mathématiques, en estimant l'influence de différents facteurs sur les résultats des enfants. Notamment, nous avons pu identifier que l'écart entre les genres en mathématiques est déclenché par l'école et non par l'âge. Dans l'ensemble, nous avons discuté des preuves récentes en matière de compréhension de la lecture et des compétences en mathématiques, à un niveau

populationnel. Ces approches scientifiques peuvent conduire à la conception de programmes d'apprentissage ciblés, tant pour les apprenants normaux que pour les apprenants à risque de développer des difficultés, ainsi que pour les apprenants rencontrant des difficultés d'apprentissage en langage et/ou en mathématiques. Tout au long de cette thèse, nous présentons des exemples de la manière dont les données massives et les analyses basées sur les sciences cognitives peuvent aider les apprenants et informer le système éducatif national. En parallèle de chaque approche, nous discutons des limites de l'approche et proposons des solutions pour les surmonter.

Mots-clefs : Sciences cognitives, développement de l'apprentissage, disparités de genre, lecture, mathématiques, école primaire

Abstract

Title: Epidemiological and cognitive evaluations in mathematics and language in the whole population of school-age children in France

Abstract: Recently, France implemented national evaluations to precisely measure the learning acquisitions and difficulties children are facing all along their first grade until their second grade, based on cognitive assessment of mathematics and language skills. Every year, about 750 000 children completed 46 exercises to assess their performances covering a total of 2.9 million children between 2018 and 2022. Analyzing these rich set of data on the complete population, this PhD aimed at providing a better understanding of the conditions that promote or hinder the acquisition of academic learning in children. To do so, we conducted a series of studies using longitudinal data from four whole-population-French cohorts assessing the relative influence of a wide range of individual-, class- and school-level factors on various aspects of academic success in primary school. Firstly, data obtained in the national program were described and predictors of learning skills in reading abilities and in reading comprehension were identified. In addition, predictive power of both child characteristics and environmental factors, at the individual, class, and school levels, were explored. Thanks to the natural experiment of Covid (i.e., absence of school exposure for a specific duration), we were able to estimate the impact of school exposure when comparing one cohort to the other. More particularly, we were able to identify learning needs among different SES school categories. Finally, we focused our last analysis on gender differences in language and mathematics, estimating the influence of different factors on children's results. Notably, we were able to identify that the gender gap in math is triggered by school and not by age. Overall, we discussed recent evidence in reading comprehension and math learning abilities, at a populational level. These scientific approaches can lead to design targeted learning programmes for both normal learners, learners at risk of developing difficulties, as well as for learners facing learning difficulties in language and/ or math. All along this dissertation, we show examples of how massive data and analyses based on cognitive science may help the learners and inform the national

education system. Alongside each approach, we discuss the limitations of the approach and propose solutions to overcome them.

Keywords: Cognitive sciences, learning development, gender gap, reading, mathematics, primary school

Table of contents

Résumé	2
Abstract	4
Acknowledgements	8
Abbreviations list	11
Chapter 1. Introduction	18
A. Literacy and numeracy, two fundamental skills in the first and second years of elementary school	18
B. French students' levels in reading and math: which level do they have and how to assess it?	25
C. The French national programme Evalaide	28
Chapter 2. Data management and descriptive analyses	31
I) Introduction	31
A. Cohort design	32
B. Study design and data collection	34
C. Children' characteristics	36
D. Selection of cognitive tests	39
E. Internal reliability	44
F. Data ethics	44
II) Methods and Data management	46
A. Data preprocessing and cleaning	46
B. Missing data management and imputation	48
C. Sensitivity analysis	51
D. Creation of new data and composite covariates	52
E. Statistical analyses	53
III) Results of Data description	54
A. Study population's description	54
B. Description of children regarding their age category when entering in first grade	59
C. Description of children regarding their school category when entering in first grade	63
D. Correlation matrices between all cognitive tests	66
IV) Conclusions	70
V) Supplementary materials	71
A. Tests' contents	71
Chapter 3. Data sciences in education and learning how to read: predictors and learning patterns	104
I) Introduction	104
A. French students decline in language levels in the last two decades.	104
B. Reading and Reading comprehension: required skills and identified predictors.	105
C. Reading and reading comprehension difficulties: different predictors.	107
D. Aims of this work and research hypothesis	108
II) Material and methods	110
A. Materials	110
B. Methods	110
III) Results	115
A. Descriptive analyses of correlation matrices	116
B. Descriptive analysis using Principal component analysis – PCA	123

C.	Predicting reading and reading comprehension at T3 with multilevel regression linear models	131
D.	Comparison between cohorts and estimation of the COVID-19 impact on reading acquisition and estimation of progress.	145
E.	Children presenting with difficulties in Reading comprehension at T3	150
IV)	Discussion	163
A.	Main results and discussion	163
B.	Conclusion	168
V)	Supplementary materials	169
Chapter 4. Schooling triggers a gender gap in math: evidence from three million children.		
I)	Introduction	183
II)	Materials and methods	186
A.	Materials	186
B.	Methods	186
III)	Results	197
A.	Rapid emergence of a math gender gap in first grade	197
B.	A new element: age is not a modulator of the gender gap, school is.	204
C.	Matching experiments: a massive gender gap remains after matching processes	206
D.	Language gender gaps present different dynamics than math	208
E.	Reproducibility of the results from 2018 to 2022	209
F.	Covid-19: A natural experiment of a lower exposure to school	209
IV)	Discussion	210
V)	Supplementary materials	216
Chapter 5. Perspectives and Conclusion		
Bibliography		238
Bibliography		250
Data and materials availability and transfer agreements (MTAs)		284
Figures list		285
Tables list		289
Résumé long en Français		293

Acknowledgements

This body of work has been made possible thanks to the amazing scientists working at the UNICOG-Neurospin lab. I am forever grateful to have had the chance of carrying out this project under the lightly guiding touch of both **Ghislaine Lambertz-Dehaene** and **Stanislas Dehaene**, whose extraordinary intelligence, strong scientific integrity, immense commitment to our society, open-mindedness, and hard work have been daily drivers to me and will remain models engraved in my memory for life. I am forever grateful to both of you for the quality of our scientific discussion and criticism for the sake of science, and I am particularly grateful for all your patience and trust during these four past years of work together, especially for allowing to step into additional ambitious projects all along my thesis. It was such a great motivation to test how to implement science in different domains, such as decision makers and politics, NGOs, companies, and working hand-in-hand with teachers and children in their classroom as well.

I am fondly grateful to Pr. Marcela Peña and to Dr. Liliane Sprenger-Charolles who accepted to review this entire work, and I am very grateful to Dr Ignacio Atal, Dr. Caroline Huron, and to Pr. Raphael Porcher for accepting to become my jury members this year.

I am forever grateful to Pr. Béchir Jaraya, whom I met during a hackathon about health and sciences in 2018 and who introduced me to Ghislaine and encourage me to pursue research works in France. I am thankful to my dear colleagues, especially Severine Desmidt with whom I shared an intense first part of work on fMRI at NeuroSpin, studying a whole new protocol at the lab and learning together how to code functional MRI images; I am thankful to all the baby lab team for their stimulating and very nice moments shared together: Marie Palu for her intense motivation boosts, her renown and daily enthusiasm, and in all the amazing projects she takes on; François Leroy for the great coffee and scientific talks we shared on babies and language, as well as his precious PhD well-being advices; Chanel Valera for her incredible kindness and relevant advices and feedbacks on my work, I particularly appreciated our shared strength during the hard-covid period, thank you for all your support Chanel; Lucas Benjamin for his positive attitude and anti-stress recommendations, and multiple coffees about society, research in France, politics and childhood, I enjoyed all of our meetings and deep conversations in interpreting our mutual results, thank you for all the precious time you shared with me Lucas; Milad Eckramnia for our debates about hypothesis in brain research, our shared passwords for the talks of language specialists (#noamchomsky class in the US) and our common passion about sailing in Brittany; Elena Kulagina for all of her kind words and working moments together; and Marie Lubineau for our common work at the national scientific council for education, for our common work when preparing our mutual projects in coding at the DEPP, a great thanks for all your kindness and excellent feedbacks on my work Marie. A special thanks goes to Véronique, Laurence, Gaëlle : you were such a great help in welcoming children with me and preparing them for their adventures in neuropsychological tests and in the fMRI for several trials; A special thanks goes to Valérie, Chantal and Yann, for their co-working experience in allowing the fMRI machine to function well during all the trials, for our discussions together on how to improve the health system, and for

their permanent pleasant mood. A special thanks goes to Bernadette Martins for initiating and including me to the ethical group in research (CER - Polethis) for two years. A very special thanks to Vanna Santoro, our amazing lab manager who is always there to listen to us and contributes intensely to our well-being at work, as well as to the NeuroSpin students first support, Christos-Nikolaos Zacharopoulos. Christos, I will be forever impressed by and thankful for your listening qualities, your availability during hard times I was going through – especially the covid loneliness time and the high level of stress I underwent through while helping as a medical professional.

This lab experience would not have been the same without the amazing doctorates and post-doctorates that belong to it, my dear friends whom I've not seen enough these past years and whom I need to plan our brainstorming and sailing/ surfing weekends with: Alexis Thual, Tiffany Bounmy, Lorenzo Ciccione, Audrey Mazancieux, Mathias Sablé-Meyer, Fosca Al Roumi and Yvan Nedelec. Thank you, guys, for your amazing positiveness, fun times shared together, climbing and beer sessions and deep conversations about life, our impact in science and on our society. I truly believe you are all game changers and I measure how lucky I am to all have met you in my life, you amaze me by your intelligence and open-mindedness, and you guided me through this big adventure of a PhD at the crossroad of cognitive sciences, data sciences and politics.

More particularly, this three-year project had been the start of a special workshop and friendship with the amazing Benedicte Colnet, whom I am proud to have intensively collaborated with, and shared so many scientific discussions on health, education, causal inference, methods, massive data, and on how to constantly improve our common work and make it available in open source. We worked days and nights and during weekends for such a long time. Data management and modelization are demanding and I've learned so much by your side. I am extremely impressed by both your modesty and your intelligence, your curiosity, and your dedication to greater causes: serving societies and humanity. I am forever grateful that our paths have crossed, and I am already looking forward to our common projects.

This work has been checked, encouraged, and critiqued by my supportive thesis advisory committee including Pascal Bressoux, Ignacio Atal and Jean-Luc Berthier. Their insights and encouragement have been helpful in improving the protocol of my research and lifting my confidence to try again when the results have seemed far from encouraging. It is also the fruit of conversations from so many amazing researchers and education specialists that have shared their experiences with me: Lilliane Sprenger-Charolles, Johannes Ziegler, Elisabeth Spelke (thank you for your amazing lectures and help on the domain of gender gap and learning math), Jérôme Deauvieu, Paul Gioia, Emmanuel Sanders, Frank Ramus and Hugo Peyre, thank you all for the deep talks about gender gaps in learning acquisition and our multiple moments about methodology for these works; Julie Josse, a special thanks for your precious advices and relevant feedbacks on all the data management and the modelization in our multiple works and for allowing to follow your stimulating courses about causal inference at Ecole polytechnique; Pr. Raphael Porcher, for his professionalism, for being demanding and constantly striving for a high quality of scientific work done, as well as our long discussions on our lifestyles and society. Thank you for being inspiring,

for being demanding in the quality of my work, and for your patience towards my parallel engagements in public health; Astrid Chevance, for inspiring me through your engagement for our society and succeeding into creating bridges between science and our society. A very special thanks for all your kind advice and support all along the projects I am implementing; Barbara Heude, Sandrine Lioret, Blandine De Lauzon-Guillain, Marion Lecorguillé and Marie-Aline Charles, from the Villejuif EDEN and ELFE cohorts, a special thank for our common work on childhood, in improving pregnant women, children's health and well-being and for all our great discussions on improving both research and health systems in France. I've been more than lucky to have worked with all of you, and I'm forever grateful for your support and continuous encouragements in all the projects I was implementing; Francois Taddei, I am forever grateful I got engaged with the LPI 4 years ago and became a humble change maker thanks to your continuous support and advices), Brigitte Moltrecht (at the department of schooling at the ministry of education); Christine Lequette (an inspirational medical doctor at school at the Grenoble academia); Catherine Grenier (an inspirational medical doctor at school in French Guyana); Christophe Gomes (whose energy and work were absolutely amazing, as well as his team of the NGO "Agir pour l'école", and all the amazing teachers I met at Calais in 2021); Pascale Berthier-Buteau and Flora Baret (you are both an inspiration, both passionate primary school teachers who I was lucky to collaborate with during 3 years, and whose students are the luckiest); A special thanks to Marc Gurgand, Esther Duflo, Adrien Pawlik, Quentin Daviot, Lou Aisenberg of the J-PAL team for their precious feedbacks on this massive data work and its possible applications to education – I was more than delighted and inspired by our exchanges and scientific talks; A special thanks to Thierry Rocher, Sandra Andrieu, Axelle Charpentier and Alexis Lermite, who made the access to these incredible data possible at the DEPP; Jean-Luc Berthier and Laurence Bonfigli-Berthier, for their incredible dedication to education, learning based on cognitive sciences and permanent happiness and positive attitude, I still need more tips from you about how to be as open-minded and determined into my future works as you both are. I measure how lucky I am to have met both of you during this thesis, thank you for your continuous support and relevant advice; Penny and Mark Dressler and Julia Tramelli, three incredible teachers in the US, where it all began: a desire to impact society, to improve the country's education system based on international examples, and a will to excel in multiple domains to serve my country.

Éléments de l'acknowledgements retirés

Éléments de l'acknowledgements retirés

Éléments de l'acknowledgements retirés

Éléments de l'acknowledgements retirés

Les éléments de l'acknowledgements retirés concernent des remerciements personnels.

Abbreviations list

- ANOVA: Analysis of variance
- COVID: Corona virus Disease
- CSEN: Conseil Scientifique de l'Education Nationale, *scientific council of national education in France*
- DEPP: *Direction de l'évaluation, de la prospective et de la performance, statistical department of evaluation, prospective and performance*
- EvalAide: Programme « évaluer pour mieux aider », *assess to better help*
- E.U.: European Union
- HPE: higher priority education public schools
- NA: a missing value, not available
- OECD: Organization for Economic Cooperation and Development
- PA: phonological awareness, an individual's awareness of the sound structure of language
- PCA: Principal component analysis
- PE: priority education public schools
- PIRLS: Progress in International Reading Literacy Study
- PISA: Programme for International Student Assessment
- SD: Standard deviation
- SE: Standard error
- SES: Socioeconomical Status score
- SOM: Supplementary (Online) Material
- T1: Beginning of first grade in primary school
- T2: Mid-year of first grade in primary school, 4 months after T1
- T3: Beginning of second grade in primary school, 12 months after T1
- TIMSS: Trends in International Mathematics and Science Study
- UK: United Kingdom

Particular abbreviations for PCA figures, as shown on Figure 18.

- OWord: Oral comprehension of words

- OSent: Oral comprehension of sentences
- OText: Oral comprehension of texts
- LSound: Letter-sound association
- LRecog: Letter recognition
- Phon: Phoneme handling
- Syll: Syllable handling
- LKnow: Letter knowledge
- ReadM: Reading words
- ReadT: Reading texts
- WWord: Writing words
- WSyll: Writing syllables
- Lett (PCA at T2): Letter-sound association
- ReadCS: Reading comprehension of sentences
- ReadCT: Reading comprehension of Texts

Chapter 1. Introduction

A. Literacy and numeracy, two fundamental skills in the first and second years of elementary school

Learning is viewed as the process of transforming received information into actionable knowledge, enabling individuals to engage with and contribute to their societies. The dynamic interplay of learning needs and temporal factors is noteworthy: over a century ago, the ability to read was not a prerequisite for social inclusion or participation. However, in contemporary times, children aged 5 to 7 are expected to demonstrate competence in reading, writing and arithmetic. These achievements represent milestones that prehistoric and medieval human populations could scarcely fathom, despite the relatively stable nature of the human brain throughout our species' evolutionary history. As children transition from the formative years of kindergarten to the critical phase of primary school, their capacity to comprehend spoken language, decode written text and comprehend arithmetic concepts significantly influences their learning trajectories. A deficiency in these pivotal skills can have enduring repercussions on their academic progress and their social integration and, a dysfunctional or deficient environment (i.e., such as an inadequate or insufficient stimulation transmission from adults) impedes their progress in language and mathematics.

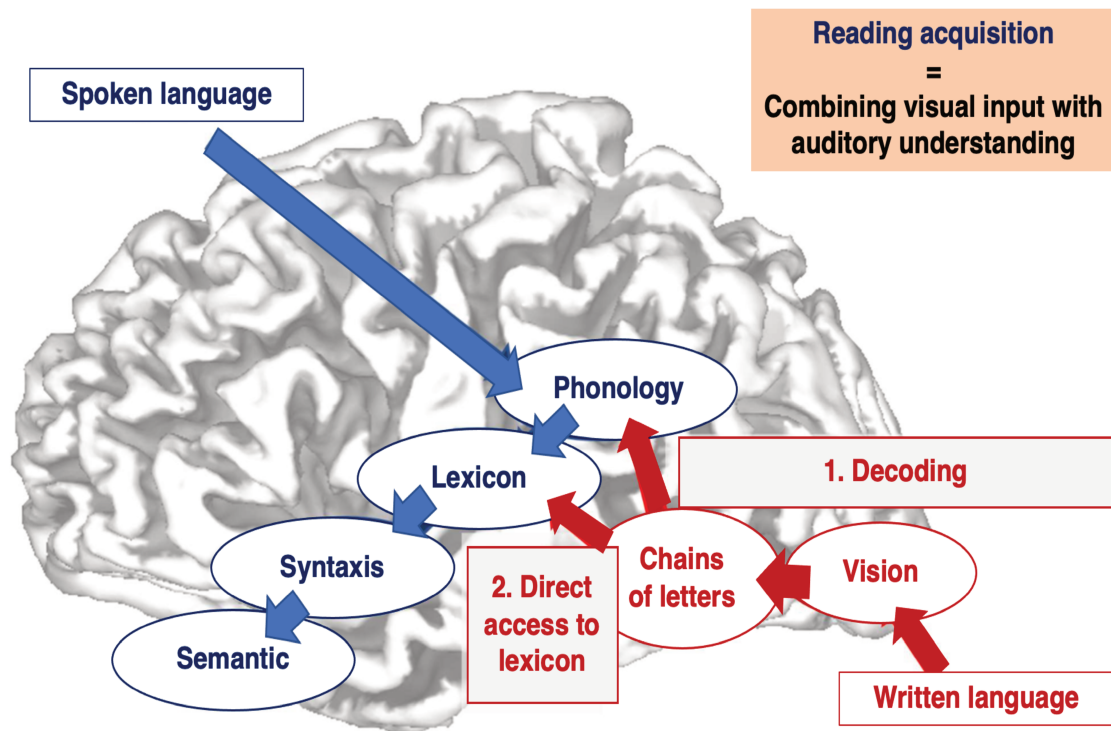
1) Oral and written language development processes, from birth to the first years of life

The newborn brain is genetically determined and programmed to interact, understand, and learn. From birth to the first years of life, his linguistic cognitive abilities are profoundly influenced by the cultural and environmental transmissions inherent in his upbringing: the baby treats and analyzes multiple information coming from its numerous interactions with its surrounded environment. The baby acts like a “statistician”: he extracts data and information from his environment, analyzes if any regularity exists and understands the causal effect of a phenomenon (Dehaene-Lambertz, 2017; Dehaene-Lambertz & Spelke, 2015). Any new non-predicted situation

will lead to a surprise, a reconsideration of the predictions, eventually to a correction of the internal models and a reinforcement of some neuronal connections vs. a weakening of others, leading to new performances and abilities (Dehaene S., 2021). Reading comprehension mastery is a determinant of both a good level of integration in the society, and of a good health (i.e., less pathology, and a higher well-being), and is the basis for academic success and social integration (Carroll et al., 2005; Christle & Yell, 2008; McArthur et al., 2016; McArthur & Castles, 2017). For instance, neglecting reading abilities in primary school can start a general learning decline with lifelong effects, while good readers reap innumerable benefits (Castles et al., 2018). Children entering at school while experiencing delays in oral language comprehension are exposed to a higher risk of developing reading acquisition difficulties, and rarely make up for their shortcomings (McArthur & Castles, 2017; Nachshon & Horowitz-Kraus, 2019).

When it comes to the acquisition of reading, spoken and written languages use different codes to arrive at the same meaning: an oral form (i.e., the prosody and phonology of speech) and a visual form (i.e., the alphabet). In terms of learning, these codes are eminently different. **Spoken language** is probably the result of the biological evolution of the human brain, it is acquired spontaneously by all typical human children (i.e., without the need for prior specialized instruction) as soon as a child is immersed in a sufficiently rich linguistic environment and is part of the human genetic endowment. Language acquisition starts at birth and at 3 years the main milestones of language acquisition have been reached (i.e. a child speaks and understands complex sentences) even if language continues to complexify during the following years (Pinto, S. & Sato, M (Eds), 2016). By contrast, **written language** is a recent cultural invention that varies greatly from one culture to another. Its invention is remarkable because it exploits the possibilities of the human brain, but it needs to be taught explicitly (Bressoux, 2012). Learning to read usually begins at around age 6 during the first year of school. The goal of reading is to enter in the previously developed language pathway through the eyes instead of the ears and is necessary to set up a correspondence between a spoken and a visual unit. These units are different across writing systems but alphabetic systems, such as French's, associate graphemes to phonemes.

Figure 1. Reading acquisition brain mechanisms (from Dehaene-Lambertz G. et al., *PLoS Bio*, 2018; Monzalvo et al., *Neuroimage*, 2012). Learning to read is a new path (vs. oral language pathway in blue) to access meaning through vision (in red).



Graphemes (e.g., <i><i><ou>, <on></i></i>) are the smallest units of written language, and **phonemes** (e.g., /i/, /o/, /u/, /ô/) are the smallest sound units of spoken language that distinguish two words (e.g., bat vs. rat, where /b/ and /r/ are phonemes in English). Even in alphabetic systems, the speed of grapheme-phoneme automation depends on the degree of regularity in the relationships between graphemes and phonemes: some languages are said to be transparent (i.e., each phoneme corresponds to one grapheme such as in Spanish or Italian (e.g., “filosofia” in Italian)) whereas French is an intermediate language (i.e., on the one hand, French language possesses complex graphemes such as <an> or <ou> due to a higher number of phonemes relative to the number of latine letters, and on the other hand, there are many ways to write the same phoneme in French (e.g., the phoneme /o/ can be written o, au, eau, aut, etc...)). By contrast, reading French is more regular compared to English which is irregular in both directions, from sound to letter and from letter to sound. Thus, writing and reading rules have a cost in learning and the most transparent languages are learned faster and with less difficulties reported in children: Students learn to read more quickly and effectively in Spanish than in French and in French than in English (Moll et al., 2014; Ziegler,

2018). Depending on the complexity of the writing system, its orthography and the effectiveness of the teaching strategy, the written code can be acquired in just a few months, and is then, in the vast majority of cases, grafted onto spoken language: typical children usually achieve fluent reading between 2nd and 4th grades (Pinto, S. & Sato, M (Eds), 2016; Vaessen et al., 2010; Ziegler et al., 2010). Overall, learning to read means replacing speech (in blue) with a new visual input (in red) (see **Figure 1**).

Two stages are distinguished in the process of learning to read: the acquisition of the code of letter-sound correspondences (graphemes-phonemes), which is slow and sequential, where the child gradually retrieves the sound image of the word; and the automation of procedures for identifying written words (Grainger et al., 2016; Lonigan et al., 2018a). The first procedure is almost always used at the beginning of learning, while the other is gradually established. In the initial stage, a beginning reader reads both invented words (e.g., "lople") and regular words (e.g., "table") equally well, but makes numerous phonological errors when reading irregular words (e.g., "sept" read as "septembre") (Sprenger-Charolles et al., 2003). A little later (typically by the end of the first grade), they read regular words better than invented words, but still struggle significantly with reading irregular words. It is not until the end of the third grade that the majority of students are able to read irregular words as accurately and quickly as regular words (Sprenger-Charolles et al., 2003). In alphabetic writing, the development of precise and rapid word identification skills requires early, intensive and systematic instruction in grapheme-phoneme correspondences (Desrochers, 2018).

Having in mind that the primary objective of reading is comprehension, as the reader internalizes the intricacies of grapheme-phoneme correspondence and automatized the identification of words, the reader can then allocate his cognitive resources, working memory, and attention toward the inherent cognitive processes essential for understanding the text they read. While it is relatively easy to assess the mastery of word identification procedures and related skills, the problem becomes considerably more complex when it comes to comprehension. Expert readers, during reading, engage in the simultaneous construction of a cognitive representation, known as a "situation model," as they identify words, giving the comprehension process a "sense of obviousness." This sense is often an illusion stemming from the fact that the reader

has automated the procedures involved in word identification and, at least partially, certain processes involved in comprehension (Kolinsky et al., 2018).

2) Acquisition of math concepts: the number sense, space and symbolic

When it comes to learning mathematics, non-symbolic skills, founded on the perception of the approximate number of a set of objects (i.e., the number sense) and also on the understanding of the most basic concepts of geometry (such as parallelism and distance) (i.e., space), are present from birth. As in language with the distinction of an innate development of oral language comprehension versus the acquisition of the writing code invented by humans, some skills in math are innate, others are the result of intentional teaching acquired in school: a combination of both is necessary to handle math concepts on a daily use.

Representing the size of a collection of objects, and comparing several objects, are skills that are spontaneously developed at a very early age in all cultures. More particularly, babies possess a “tool box” to perceive and appreciate numerosity, which is called “subitizing”, the ability to distinguish very small quantities from 1 to 3 (Dehaene, 1999). At this point, activities related to quantification do not require the use of numbers (either orally or in writing): the very young child knows how to quantify approximately whether a collection of candies is "a little, a medium, or a lot". This is called the "number sense" and corresponds to innate numerical skills in the developing child (Bellon et al., 2019; Dehaene, 1999; Gennari et al., 2023; Izard et al., 2008). In addition to subitizing, 6-months-old infants also possess an approximate sense of quantities that enables them to distinguish between two collections of objects with a quantity ratio of 1/2 (e.g., 10 vs. 20 or 4 vs. 8). This approximate sense is shared with animals and is called the “Weber fraction”, it refines with age and education and serves as the foundation for the development of exact calculation (Halberda & Feigenson, 2008; Odic et al., 2013). Even before school entry, the availability and accuracy of these numerical skills are predictors of later performance in mathematics (Gilmore et al., 2007, 2010; Gimbert et al., 2019; Lyons et al., 2014).

Similarly, the distance effect is an expected consequence of approximate calculation since the Weber fraction is obviously smaller between close numbers than distant ones, and we initially have a logarithmic representation of numbers, which yields this second property: there is more perceptual difference in our mental representation of numbers between 1 and 2 than between 101 and 102. The development of the priming distance effect (i.e., situation in which discriminating between two numbers that are far apart is easier than discriminating between two numbers that are close), which represents our notion of “magnitude”, has been identified as early as in grade 1 (Reynvoet et al., 2009). Furthermore, numbers are inherently on a number line, meaning they have a spatial representation, but this representation is not linear (i.e., the difference of 1 is the same between 1 and 2 as between 101 and 102) and is rather logarithmic (Dotan & Dehaene, 2016; Hamdan & Gunderson, 2017).

On the other hand, the symbolic representation of the number allows a precise, exact quantification. To have meaning and to be used for good, these numerical symbols must be linked to the number sense. The acquisition of symbols (words, numbers) is an essential element of progress in elementary school mathematics, and it is crucial for a child to automate the transition from symbols to corresponding quantities. It is this link which gives the meaning (i.e., semantic) to the sign. It is very gradually, typically between the ages of 2.5 and 4, that the child begins to connect words to their quantities. The acquisition of symbols (i.e., words, numbers) is an essential part of progress in elementary school mathematics. The child needs to automate the transition from symbols to the corresponding quantities: It is between 4 and 5 years of age that children begin to associate their comprehension of numbers together with their non-symbolic representations of small and large numerosities (Spelke, 2005). This operation of converting symbols into quantities is gradually automated between first and third grade (Girelli et al., 2000). Indeed, this rapid, automatic and unconscious comprehension becomes, in the first years of school, a new predictor of success in mathematics (Geary, 2011). The advanced-in-math children have a deep and fluid understanding of numbers: for example, they can count with agility based on a rapid decomposition of numbers into subsets (e.g., $8 = 4$ groups of 2) (Starkey & McCandliss, 2014). The child relies on this knowledge to solve problems of daily life, posed in verbal

or concrete form - a skill that the TIMSS 2015 and 2019 survey both identified as one of the major sources of French students' delay in mathematics (PIRLS and TIMSS, 2015a). One of the key elements for developing an agile sense of number is the understanding that numbers can be represented as a numerical line, oriented from left to right, on which additions of integers correspond to rightward shifts and subtractions to leftward shifts.

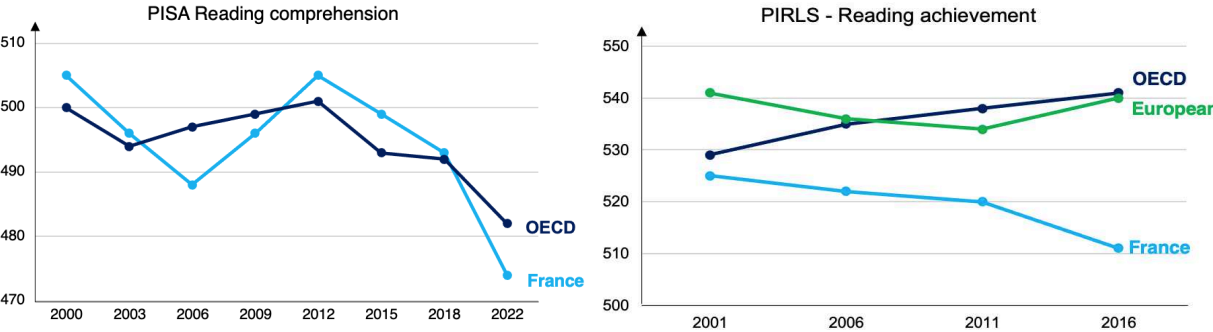
The representation of number in space, using a number line, gradually becomes linear with the learning of exact numbers, first between 1 and 10 and then across all numbers (Booth & Siegler, 2008; Dehaene et al., 2008; Siegler & Opfer, 2003). Uneducated children and adults tend to think that larger numbers are "closer" than smaller numbers (9 seems closer to 10 than 2 does to 1). The idea that all numbers are evenly distributed on the number line (i.e., that there is the same distance of 1 between all consecutive numbers) is an important turning point in learning arithmetic (PIRLS and TIMSS, 2015a). It can be improved by mathematical games (Dillon et al., 2017), in particular board games where one moves in space according to the number drawn on the dice (Siegler & Ramani, 2008). Recent meta-analyses reported modest, but significant, relationships between early approximate number discrimination ability (number line) and later math achievement (Schneider et al., 2017, 2018). However, the direction of causality reflected in this correlation remains debated (Lyons et al., 2014). The current consensus is that it is the introduction of symbols for numbers, and the understanding of how these symbols relate to the concepts of set, cardinal number, ordinal number, and position on the number line, which are the most important factors in children's later mathematical development. Regarding potential gender effects in math, no sex differences have been reported at any of these transition points, not even in studies with substantial sample sizes. Secondary, mathematical abilities also develop similarly in boys and girls (Spelke, 2005).

B. French students' levels in reading and math: which level do they have and how to assess it?

With the aim to assess students' proficiency in mathematics and language consistently over time, utilizing the same assessment methods, three key programs have been established to evaluate and enhance the effectiveness of education in various countries (i.e., among European Union (E.U.) and OECD): (1) Programme for International Student Assessment (PISA), Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading Literacy Study (PIRLS). France participated to these three international assessment projects: the first one, PISA, comprised ~ 85 OECD countries and examined ~ 600,000 15-year-old children at the international scale, including ~ 8,000 students in France every 3 years; TIMSS measured students' performance in mathematics both at the end of the fourth and of the eighth grade (i.e., 9-year-old and 13-year-old children in average) and the last one, PIRLS, focused its assessment on 8th graders in reading comprehension (i.e., 10-year-old children).

The level of French young students in both language and math has been mainly decreasing in PISA since 2000 (except for years 2009 and 2012 (see the light blue line on **Figure 2** and **Figure 3**)), as well as in PIRLS international assessments (see **Figure 2**) (i.e. France's average was significantly below both OECD and EU average while other countries have improved (i.e., England, Portugal) or stabilized their levels (i.e., Ireland, Romania)) and as well as in TIMSS (i.e., a significant decrease of level for France between 1995 and 2019, and in 2019, an average math level inferior to all European Union countries and OECD countries in math, except for Chili's) (OECD, 2018b; PIRLS and TIMSS, 2016a, 2011; *PISA 2012 Results*, 2012). PISA and TIMSS reports regularly highlighted the relatively high percentage of French children with reading and mathematical difficulties, more so among lower socioeconomical scores (SES) children (Mullis et al., 2012; OECD, 2019b, 2019c, 2019d; PIRLS and TIMSS, 2015b).

Figure 2. PISA (left) and PIRLS (right) results since 2000-2001 comparing France and OECD/European countries.

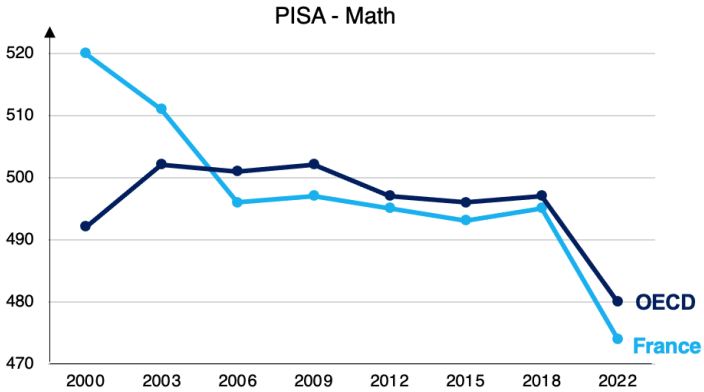


In addition, an alarming decrease of level in math has been measured for two decades in France in the international PISA and TIMSS results in math. Between 2000 and 2022, French student achievements in math declined significantly in PISA assessments (OECD, 2018b; *PISA 2012 Results*, 2012), with a math average of 520 points in 2000 and of 474 in 2022 (see **Figure 3**) and TIMSS math scores being below the OECD averages since 2015 (i.e., In 2015, France: 488 vs. 527 points for OECD; In 2018, France 485 vs. 529 points for OECD, see **Table 1**).

Table 1. TIMSS results in Math from 2015 to 2019

	2015	2019
France	488	485
OECD	527	529

Figure 3. Math PISA results since 2000



Also, in 2019, France brought only 2% of its students to the advanced level in mathematics in TIMSS Math assessments (vs. 11% in OECD countries). The gap between the scores obtained by students in France and those in other OECD countries was more pronounced among socially disadvantaged students (i.e., score 53 points below the OECD average for lower SES children in France and only 4% vs. 19% of lower SES children in France reached the high level of the TIMSS math assessment compared to OECD countries respectively). The self-confidence and motivation of students with regard to mathematics deteriorated sharply between grade 4 and junior year of high school (PIRLS and TIMSS, 2015a). Even if these international surveys identify learning difficulties for French students, they come too late to optimize the effectiveness of interventions to prevent learning difficulties at school (i.e., they are conducted between age 9 and 15).

To act on the decrease of level of their children and following these international comparisons several countries implemented local to national measurement methods to (1) assess the level and identify specific difficulties of each child, (2) to be able to intervene the earliest and correct his learning trajectory and (3) to assess direct and indirect effects of environmental and individual elements (i.e., effects and effect sizes of class size, age, gender, social category) on the children' learning improvements (Department of education, 2020; OECD, 2019a; Thomas et al., 2022) as recommended by the OECD report of 2013 (OECD, 2013). As an example, in 2012, the United Kingdom introduced a national and mandatory decoding assessment called "Phonics screening check" for all first graders (Duff et al., 2015). Every child underwent reading aloud 20 frequent words and 20 non-existing words, alone with the teacher. This assessing step aimed at identifying specific children-at-risk and providing them with individualized exercises. These identified children underwent a second assessment a year later to estimate their progresses and to supplement with more exercises if needed. Although it is not possible to attribute England's success to the introduction of the phonics check alone, it should be noted that the level of English students on PIRLS test has also increased from year to year (PIRLS and TIMSS, 2016a).

C. The French national programme Evalaide

In France, until recently, there were no national standards nor references set by the central administration, but rather objectives and orientations transmitted to the education system, which served as guidelines for the actions to be carried out in the teaching units and all along the school year, teachers provided their students with their own evaluations (Broccolichi & Sinthon, 2011).

For reaching higher level in reading comprehension and math, intervening as early as in primary school matters as this early period of life is a phase of strong vulnerability, but also of sensitivity to an optimal environment, allowing the full development of the child's potential in language and math (Barnett, 2011; Bianco et al., 2010, 2012; Ehri et al., 2001; Noble et al., 2006; Nores & Barnett, 2010).

With the aim to target this sensitivity window, and inspired by other countries, the country-wide French national evaluation program *EvalAide* (“*évaluer pour mieux aider*”: assess to better help) was implemented for the first time in 2018. It was proposed by the Conseil Scientifique de l'Éducation Nationale (CSEN) with two goals 1) to enable teachers to assess achievements in the basic cognitive skills needed to learn reading and mathematics 2) to provide specific help in skills that are insufficiently developed. This battery of 46 language and math tests was designed by scientists and educators to provide French teachers with a detailed picture of the needs, achievements, and progress of every child in their classroom, thus supporting focused pedagogical interventions and the setting of national standards. Every year, all French children underwent longitudinal tests at the beginning of first grade (T1), after 4 months of school (T2), and at the beginning of second grade (T3). Math tests included digit identification, counting, number comparison, number-line knowledge, problem solving, calculation, and geometry, while language tests cover letter knowledge, letter-sound correspondences, phonological awareness, reading aloud, vocabulary, oral comprehension and reading comprehension (see details in Chapter 2).

Having this powerful tool with such massive data (i.e., four consecutive cohorts of 5-to-7-year-old first graders from 2018 to 2022 for a total of ~2.8 million children), we wondered if analyzing these individual benchmarks would allow to understand and affine the French children' learning needs and to detect children-at-risk of developing learning difficulties in order to be able to act earlier and tackle learning difficulties. In addition, as mentioned above, France remained one of the highest OECD countries with a large SES gap in learning abilities (OECD, 2018b) and we wondered if Evalaide would help adapt and specify the learning needs of the lower SES populations and would help to measure its efficiency in reducing the SES gap after intervention.

Compared to existing systems in other countries, three innovative aspects characterize Evalaide. The first is the scope of the skills measured. It was chosen not to stop at a single performance measure, such as decoding in reading, but rather to include a multitude of skills that would be expected to predict the success or failure of students in reading and in math. This allowed teachers to get a very accurate picture of the skills that needed to be reinforced and gave them informed hypotheses about where the difficulties were coming from. Secondly, it was decided to start this scheme at the beginning of the year and not at the end of the year as in the UK to give teachers time to set up pedagogical interventions throughout the year. It is not a question of verifying what has been learned at the end of the year, but rather of alerting and allowing teachers to tackle children' potential difficulties so that they can help the student well before he or she finds himself or herself in failure. Finally, the third innovative aspect is its "longitudinal" nature: students were evaluated at three times (i.e., at the beginning of 1st grade, in the middle of 1st grade and at the beginning of 2nd grade) which allowed teachers to evaluate a student's progress during the year, and therefore the effectiveness of their teaching strategy put in place, and to revise it if necessary. From the beginning of 1st grade, the first evaluations gave indications on the level and needs of the students, which already allowed teachers to adapt their teaching. The mid-1st grade evaluations (towards the end of January) provided a "progress report" on the students' progress. They enabled teachers to identify those who, at the end of the first term of 1st grade, had difficulty learning to read and to identify the nature of their difficulties.

As we begin this work, we aimed at understanding why some students do not fully acquire reading comprehension or fail at developing proper math abilities. We aimed at precisely identifying predictors and factors associated with learning improvement and the models of reading comprehension abilities using both description and analysis of Evalaide data from 2018 to 2022. We expected to affine children's language and math abilities dynamics over time during the first year of primary school and to detect more easily some children' premises of difficulties. In addition, we aimed at characterizing the effects of elements in the schooling environment associated to a better learning acquirement in math and language. In addition, these massive data allowed us to measure some effects of school exposure on the development of skills in math and language with the particularity of the covid year (i.e., associated with an important absence of school). This thesis embarks on an exploration of the multifaceted impact of first grade learning acquisitions, shedding light on how related and predictive the language and math items were to the later levels in second grade and on the factors that potentially hinder a child's progress in primary school.

Chapter 2. Data management and descriptive analyses

This section is based on the following scientific article:

- Martinot P., Colnet B., Huguet P., Spelke E., Bressoux P., Dehaene-Lambertz G., Dehaene S. (*submitted to Nature*) “Schooling induces a gender gap in math: evidence from three million children”.

I) Introduction

As mentioned earlier, PISA, TIMSS and PIRLS surveys regularly highlight the relatively high percentage of French children with reading and mathematical difficulties (Mullis et al., 2012; OECD, 2019b, 2019c, 2019d; PIRLS and TIMSS, 2015b). However, these international surveys, conducted at age 15 and age 9 respectively, came too late to optimize the effectiveness of interventions to prevent learning difficulties at school.

The EvalAide programme (“Évaluer pour mieux aider”, national assessments in first and second grade) was designed to provide information and identification elements of pupils at risk of developing difficulties in reading and mathematics, as early as possible, with two goals 1) to enable teachers to assess achievements in the basic cognitive skills needed to learn reading and mathematics 2) to provide specific help in skills that are insufficiently developed.

Comparable schemes were implemented in several countries with excellent results, such as Sweden with its “individual improvement projects” (known as “IUPs” - Individuell Utvecklingsplan) for reinforcement sessions in specific learning domains for struggling children and those at risk of developing difficulties (H. Smith, 2010); Finland with the ALLU tests (Ala-asteen Lukutesti [Reading test for Primary School] (Lindeman J., 1998; Psyridou et al., 2020)), followed by personalized support for pupils' progress; the UK with the “phonics check” in first grade, which enabled its young people to go from 58% success in word reading by the end of first grade in 2012 to 82% in 2018 (Department of education, 2020); or Singapore with its early detection programme around the fundamentals in first, second and third grades and the associated daily

program of individualized help in small groups (OECD, 2011). As early as kindergarten, research suggests that equipping teachers with more precise assessment tools facilitates children's progress (Raudenbush et al., 2020).

Since 2018 and led by a similar aim, the program "Evalaide" was implemented for all first graders and second graders in France. A total of 2.9 million children underwent 44 to 46 cognitive tests (i.e., depending on the year, see below) assessing language and math skills at three time points. The following chapter presents the data management, cleaning processes and sanitizing checks applied to Evalaide data as well as the description of general data to explore how informative and in which matter these massive data can be useful in classrooms. We provide all of our coding scripts, precise variables dictionaries, descriptions, and workflow, to make these data accessible to all researchers.

A. Cohort design

1) Purpose of the Evalaide program

In collaboration with the cognitive scientists, member of CSEN, the Department of Evaluation, Prospective and Performance of the French national education ministry (DEPP, *Direction de l'évaluation, de la prospective et de la performance*), developed the EvalAide program with the purpose of providing every teacher in 1st and 2nd grade with a detailed picture of the needs, achievement, and progress of each child in their class, in both math and language.

A bulletin explaining the purpose of the tests and proposing exercises in the event of failure was produced in conjunction with the tests. At the beginning of 1st grade, the tests were selected to assess the basic skills for reading and mathematics and detect children who were lagging behind. In the middle of 1st grade and at the beginning of 2nd grade, the tests were carried out to monitor the children's progress, examining whether they were responding correctly to the pedagogical intervention, for teachers to adapt their pedagogical strategies in case progress was deemed insufficient. Parents were also informed of the results by mail. Teachers were encouraged to hand

over the results during a one-to-one meeting with the parents, fostering parent-teacher collaboration.

The entire Evalaide test battery was implemented with both aims of (1) assessing every child’s progress and individual learning needs, and (2) facilitating the implementation of targeted pedagogical interventions if needed. Its main goal was to proactively address and prevent the development of challenges in children' language and mathematics learnings. Nevertheless, as a secondary goal, the data also enabled fine-grained statistical monitoring of children's school performance in France, identifying the effectiveness of specific elements in the school or classroom environment that would benefit the most to children’s progress.

2) Population

We analyzed four consecutive longitudinal French national assessment cohorts, targeting all children entering 1st grade respectively in 2018, 2019, 2020 and 2021. The total number of 1st-grade classes tested, the number of classes and the number of schools were presented in **Table 2**. Following French law, most children entered in first grade in September of the year of their sixth birthday (see the description of “Age in first grade” below). First grade represents a major change associated with the beginning of both the math and reading curriculum in France.

Table 2. Total number of first graders tested in 2018, 2019, 2020 and 2021.

Cohorts in Evalaide	2018	2019	2020	2021	Total
N: Initial number of children in the database	610,905	711,452	743,734	804,989	2,871,080
N classes	43,970	51,599	54,073	54,224	203,866
N schools	27,043	30,578	31,515	31,772	120,908

The increasing class size numbers were the consequence of a political decision to reduce class size for priority education and higher priority education public schools, a decision which was progressively implemented during those four years (see **Table 2**). In France, there were four types of schools defined by the DEPP and the ministry of

education, using a combination of school status (private or public) and four additional characteristics: the proportion of disadvantaged socio-professional categories in the geographic area surrounding the school; the proportion of students benefitting from social aid and scholarships in the living area surrounding the school; the proportion of pupils living in a sensitive urban area within the school; and the proportion of pupils attending the school who repeated a school year before their sixth grade. Following this gradient, public schools were categorized in three tiers: (1) Regular public schools, (2) priority education (PE) public schools and (3) higher priority education (HPE) public schools. Private schools were considered as a fourth category, for a total of 4 categories. Note that being categorized as PE or HPE meant that the school was entitled to special educational benefits. In both categories, starting in 2018, more teachers have been assigned to reduced class size. The goal was to halve class size in those school districts. More precisely, in 2017-2018, 2200 HPE classes were halved; in 2018, 3200 first grade PE classes and 1500 second grade HPE classes were halved; in 2019, 3900 second grade PE and HPE classes were halved; and finally, in both 2020 and 2021, all kindergarten, first and second grade classes in PE and HPE were halved. Outside the priority education system, a maximum of 24 children per class was mandatory for all regular public schools and kindergartens.

B. Study design and data collection

Assessments were implemented at three specific times: beginning of 1st grade (between the 3rd and 4th week of September), hereafter called T1; middle of 1st grade (between the 3rd and 4th week of January; T2) and beginning of 2nd grade (between the 3rd and 4th week of September; T3) (see **Figure 4** below). Each test assessed a specific skill whether in oral language, in reading, in mathematics or in problem solving. Tests were administered to the whole class, and children responded by circling the answers or writing in an individual notebook. The only exception to this procedure was the 1-minute reading aloud test, which was administered individually. Testing sessions lasted around 35 min for language tests and 25 min for math tests at T1, 35 min for language tests and 25 min for math tests at T2, and up to 35 min for language tests and 30 min for math tests at T3.

In the days following testing, teachers and schools were responsible for entering each individual response into a dedicated computerized system, then data were copied and anonymized at regional level and sent to the national level where it was stored in accordance with the European General Data Protection Regulation (GDPR).

Prior to the national programme, pilot studies were conducted by the DEPP in January and May 2018 eight months before the launch of the first cohort to finalize the tests design. About 150 private and public schools (excluding priority education and high priority education schools) took part in these pilot studies, which included about 5 000 first graders and 300 teachers, educators and inspectors who gave feedback on the tests (see **Table 3**). Those surveys were used to select the tests which were first implemented in September 2018 for the whole population of first graders in France.

Table 3. Characteristics of the pilot studies population in 2018 (n = 9797).

School type categories	Number of schools, n	Number of first graders, n	Number of second graders, n	Total number of first and second graders, n
Regular public schools	120	3902	3876	7778
Private schools	24	1058	961	2019
Total for public and private schools	144	4960	4837	9797

In subsequent years, feedback from teachers, educators, inspectors, and scientists from the CSEN was gathered to improve the tests explaining a few changes between the four cohorts. 2 tests were withdrawn from 2018 (recognizing letters among symbols at T1, and reading nonexistent words at T2), 7 tests were slightly changed in either their ergonomics or their number of items, and 2 tests were added in 2019, 2020 and 2021 (geometry at T1 and reading comprehension of sentences at T2). The tests remained similar between 2020 and 2021. For the analyses concerning a specific cohort, all 44 common tests were used despite the minimal variants between cohorts. For between-cohort comparisons, only the 37 identical tests were considered.

C. Children' characteristics

The variables used for the four cohorts (2018, 2019, 2020 and 2021) were detailed below (**Table 4**).

Child gender. Gender was registered in a binary manner as male or female and reported by the teacher. As recommended in the literature, we used the term “gender” instead of “sex” all along this manuscript, as the gender was declared by an external person. In regression analyses, boys were attributed a value of 0.5 and girls a value of -0.5.

Child age at T1 (months). The birth month and year at T1 were recorded by the teacher. As most children entered in first grade in September of the year of their sixth birthday, the “typical age in first grade” was defined as being between 69 months in September ($= 6 \times 12 - 3$) and 80 months of age in September ($= 6 \times 12 + 8$), both included. A few children with higher abilities in kindergarten were allowed to enter in first grade “one year ahead”, thus with a younger age ranging from 57 to 68 months included. Conversely, children with learning difficulties at school were encouraged to repeat their first grade (although this is quite rare), or some children entered school late for other reasons such as immigration, specific needs or because they were considered “immature”; these children were therefore one year older than their peers, with an age ranging from 81 months to 92 months included. The variable “Age at T2” corresponded to “Age at T1” plus 4 months, and “Age at T3” corresponded to “Age at T1” plus 12 months.

Child age category. Using the child age at T1, we defined a 3-level categorical variable by subdividing the children into 3 groups based on their age: all children aged 57-68 months were categorized as “advanced”; all children aged between 69 and 80 months included were categorized as “typical age”; and all children aged 81-92 months were categorized as “late”.

Class size. Class size was defined as the number of children per class at T1. Some classes were declared with fewer than 5 children per class. This is a rare but possible situation, found mainly in rural areas where the weak number of children forces schools to gather all children belonging to primary school into a single class, called a “multi-level class”, with an age range from 6 to 11 years old. Unfortunately, our database did not include information about whether a class was multi-level or not. As we were interested in the gender gap, which could only be meaningfully computed within a given class if that class comprised a sufficient number of children of either gender, when analyzing class-level variables, we only selected classes ranging from 6 to 27 children per class, in line with the referenced STAR experiment (Angrist & Lavy, 1999).

Heterogeneity in math or language per class at T1. Initial class heterogeneity was calculated as the standard deviation of children standardized and Gaussianized math or language scores per class at T1.

Proportion of boys per class. For every selected class, the proportion of boys was built as the number of boys divided by the total number of children and, ranged from 0 to 1.

Gender of the first of class in math or language tests. Separately for language and math, we identified the first of class and registered his or her gender. Boys were attributed a value of 0.5 and girls a value of -0.5. When several children were tied at the top of the class, their genders were averaged. For instance, if 2 boys and 1 girl were tied at the top of the class in math, the value for this variable was 0.1667. A positive score implied that a majority of boys were first of class in math in this class. The variable was numerical and ranged between -0.5 and 0.5.

Table 4. Description of the collected variables at the individual level, at the class level and at the school level.

Covariate	Level	Description	Type	Unit
Gender	Individual	Child gender declared by the teacher at T1	Binary	Male or female
Age at T1	Individual	Age in month when beginning first grade	Continuous	Month
Age categories at T1	Individual	Age in month when entering first grade, categorized in 3 groups	Categorical	Advanced, typical, late
Class size	Class	Number of children per class	Continuous	Number of children per class
Initial level in a class in math or language tests	Class	Mean level in math or language at T1 per class	Continuous	0 to 100; Percent of success in math or language
Heterogeneity of level at T1 in a class in math or language tests	Class	Standard deviation of the mean level in math or language at T1 per class	Continuous	0 to 1
Math or language tests at T2 and T3 per class	Class	Mean level in math or language at T2 and T3 per class	Continuous	0 to 100; Percent of success in math or language or percentile rank in math or language
Boys' proportion per class	Class	Number of boys per class over the total number of children per class	Continuous	0 to 1
Gender of the First in class in math or language tests	Class	Gender average of the top of the class in math or languages (0.5 for boys and -0.5 for girls)	Continuous	-0.5 to 0.5
Gender gap in math or language tests	Class	Subtraction of boys' and girls' class averages in mathematics or language tests	Continuous	Percent of success in math or language
Social categories of school	School	Four categories of school categories on the whole French territory	Categorical	Private, regular public, priority education public, higher priority education public schools
Socioeconomical status score	School	Averaged index of school's socioeconomical status attributed to each child	Continuous	~ 50 to ~ 150

School category. As described above, this variable was defined by the DEPP and the ministry of education using a combination of school status (private or public) and of four additional characteristics (i.e., the proportion of disadvantaged socio-professional categories in the geographic area surrounding the school; the proportion of students benefitting from social aid and scholarships in the living area surrounding the school; the proportion of pupils living in a sensitive urban area within the school; and the proportion of pupils attending the school who repeated a school year before their sixth grade), for a total of four categories: Private schools, regular public schools, priority education (PE) public schools and higher priority education (HPE) public schools.

School socioeconomic score (SES). This score reflects the socioeconomic score (SES) of the environment surrounding children. We did not have the individual SES score, but a proxy computed by the DEPP. It consisted of a combination of the following data: parents' diploma level, material conditions level, family composition, cultural capital, cultural ambition, parental implication levels and cultural practices registered for children in 6th grade. Then, a retrospective projection of the children socioeconomic characteristics was implemented, through a post-hoc attribution of the SES score to the primary school that a given 6th grader had attended. Finally, the school SES score was computed as the mean of all SES scores of children who attended the same primary school (Rocher, 2016). Ultimately, the score was a numerical variable, defined by the DEPP, going from ~ 50 to ~ 150 and representing school socioeconomic status, 50 being the lowest, and 150 the most advantageous.

D. Selection of cognitive tests

Math skills.

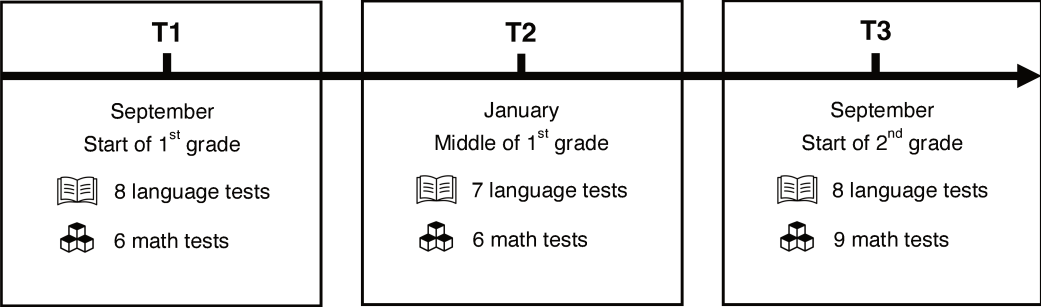
Even before kindergarten, children already have the intuition that two collections can be combined to form a third, which is the addition of the first two. Initially, the calculation is slow and sequential, it must be automated by regular exercises. With practice, the student acquires a panoply of arithmetic strategies adapted to each problem (counting, finding the result in memory, using tens, symmetry, etc.). The elementary operations

themselves must be mastered, such as knowing how to add or subtract two numbers between 0 and 9, either by counting or in the form of a memorized table. The execution of such complex procedures calls massively on executive attention, i.e., on all the systems for supervising mental operations (cognitive control, choice of strategies, inhibition of distractions and undesirable strategies, capture and correction of errors) which are mainly under the aegis of the prefrontal cortex. Therefore, it requires a great deal of attention and concentration, and is particularly sensitive to distraction. A total of 6, 6 and 9 math assessments were presented to children, respectively at T1, T2 and T3 as showed in the **Figure 4**. All the math assessments were detailed in the supplementary materials.

Reading skills.

The automation of written word identification procedures, which allows to dedicate memory and attention resources to reading comprehension, is a progressive process that translates from a serial reading mode (the slow and laborious letter-to-letter decoding of the beginner) to a so-called parallel mode, in which the letters contained in a word are processed simultaneously. An assessment of reading level must, therefore, include tests to not only verify the accuracy of reading but also its speed. Furthermore, reading performance is predicted by phonological awareness (PA) (i.e., the ability to break down speech explicitly into its constituent phonemes), vocabulary and letter knowledge both in preschools and in early primary schools (Massonnié et al., 2019). Studies on reading comprehension identified word decoding, oral language, vocabulary and listening comprehension as the main predictors of reading comprehension abilities. A total of 8, 7 and 8 language assessments were presented to children, respectively at T1, T2 and T3 as showed in the **Figure 1**. All the language assessments were detailed in the supplementary materials.

Figure 4. EVALAIDE study design



Details of each assessment and their corresponding cognitive functions were shortly described on **Table 5** and detailed in the supplementary materials.

Table 5. Summary of the tests performed at each time period.

Domain	Skill area	skills evaluated	Time of evaluation		
<i>Math skills</i>					
	Number reading	Converting an Arabic numeral into a spoken number word	T1		T3
	Number writing	Converting a spoken number word into an Arabic numeral	T1	T2	T3
	Enumerating	Counting an organized or disorganized collection and identifying the Arabic numeral corresponding to that quantity.	T1		(T3)
	Number comparison	Selecting the larger of two Arabic numerals	T1	T2	
	Problem solving	Understanding an orally presented arithmetic problem, choosing the correct operation, and finding the exact result.	T1	T2	T3
	Number line	Finding the number corresponding to a given position on the number line.	T1	T2	T3
	Addition and subtraction	Solving simple written addition and subtraction problems.		T2	T3
	Mental calculation	Mentally calculating additions of two spoken numbers.			T3
	Geometry	Recognizing and using concepts of alignment, right angle, length, and symmetry.	(T1)		T3
<i>Language skills</i>					
	Oral comprehension of words	Understanding a word read aloud by the teacher and finding the corresponding picture.	T1		T3
	Oral comprehension of sentences	Understanding a sentence read aloud by the teacher and finding the corresponding picture.	T1	T2	T3
	Oral comprehension of texts	Understanding a text read aloud by the teacher and finding the corresponding picture.	T1		

Phoneme handling (or phoneme awareness)	Identifying the word that begins or ends with the same phoneme as a target word	T1	T2	
Syllable handling	Identifying words that begin or end with the same syllable as a target word. At T2 and T3, writing a spoken syllable under dictation (mu, ti, na, lur, sar, ol, moi, che,...).	T1	T2	T3
Letter-sound association	Identifying the initial phoneme of a spoken monosyllabic word and associating it with the corresponding letter.	T1	T2	
Letter recognition	Recognizing the different writings of a letter read aloud by the teacher.	T1		
Visuo-attentional abilities	Comparing two consonant strings, in a limited time (2 minutes).	T1		
One-minute word reading	Reading correctly as many words as possible from a list of words, in a limited time (1 minute).		T2	T3
One-minute text reading	Reading correctly as many words as possible in a text, in a limited time (1 minute).		T2	T3
Writing words under dictation	Writing the correct spelling of a spoken word (silent letters at the end of the word accepted).		T2	T3
Reading comprehension of sentences	Understanding a sentence and circling the corresponding picture.		(T2)	T3
Reading comprehension of texts	Understanding a short text and answering questions read by the teacher.			T3

Note: A period indicated between parenthesis means that the assessment was added after 2018.

E. Internal reliability

No direct evidence was measured on test-retest reliability – it was already a challenge to implement national tests that took ~2 hours of testing in all French children, and the administration that implemented them considered impossible to add a re-testing period, even for a fraction of children. However, we had several indications that test-retest reliability was high enough, which we showed with math tests: Firstly, we measured the internal consistency of the results (e.g., the coefficient alpha and omega of Cronbach). Using the results per sub-scores (e.g., number line, problem solving, subtractions), we were able to calculate an approximation of the internal consistency of the results from T1 to T2 to T3, using the alpha and omega coefficient of Cronbach. In the literature, an $\alpha > 0.7$ represents a strong internal consistency. In our results, the internal consistency coefficients were always very strong: Math at T1 in 2018 (= alpha 0.70; Omega = 0.80); Math at T2 in 2018 (= alpha 0.80; Omega = 0.82); Math at T3 in 2018 (= alpha 0.79; Omega = 0.85), and similar alpha and omega coefficients were found in language, and for both domains in 2019, 2020 and 2021 (results not shown). Secondly, we used the measurement of correlation coefficients between T1, T2 and T3 to estimate reliability: we calculated a correlation matrix for all the tests. The correlation coefficients across time (from T1 to T2 and T3) gave a lower bound of the reliability of the test scores within a given subject (with additional variability resulting from the fact that the tests were not identical, and that the student might have progressed). All the correlation coefficients were very positive among math scores at T1, T2 and T3 and in language at T1, T2 and T3 (see **Figure S1** in supplementary materials).

F. Data ethics

1) European General Data Protection Regulation

For each class, teachers were responsible for registering all their students' results on the provided software. Then data were copied and anonymized at the regional level, then sent to the national level where they were stored following the European General Data Protection Regulation (GDPR). Parents were informed about the national assessments made by the ministry of Education's statistical department and could

specify their refusal to allow their children's data to be used. The data were subject to various quality controls such as deletion of duplicates, comparisons with former data sets and a control of correct and valid values for each variable by the national statistical institution of the ministry of education (DEPP). All Personal ID-numbers were checked for errors.

2) Data and materials availability and transfer agreements (MTAs): Open science

For confidentiality reasons, the raw data were not shared in public but were accessible through a secured data convention established with the DEPP. For reproducibility, code and models that were used to generate results, text, figures, and tables both in the main text and in the supplementary information were available on the following GitHub repository: PauMdlm/Education and were openly shared with both the department of statistics at the ministry of national education in France, and the scientists willing to explore these data at Ecole Nationale Supérieure (ENS), Paris School of Economics, J-PAL and IDEE Lab.

II) Methods and Data management

A. Data preprocessing and cleaning

For some covariates, outliers were identified, such as age inconsistency or a child apparently absent of school on the day of the exam and having either zeros or missing values in math and/or language on the whole assessment period. **Table 6** summarized the **outlier's management steps**.

Age outliers. When children had identified aberrant birth dates (e.g., a child registered as born in 2018, and thus supposedly entering 1st grade at the age of 2), their age was replaced by a missing value (not available, NA). A total of 169, 310, 261 and 446 children had aberrant birthdates respectively in 2018, 2019, 2020 and 2021. Ages outside of 51-98 months were replaced by NA.

Missing values on an entire session. A child who was absent from school on the day of the assessment was assigned either zeros or missing values in math and/or language for a given assessment period. When math or language tests contained missing values or zeros on a whole session while having plausible results elsewhere, only the scores for this specific session were replaced by NA's, while keeping the other two tests' sessions as valid. All students with at least one valid test session (T1, T2 or T3) were kept in our analysis. A total of 75, 101, 128 and 1222 children were excluded because all sessions were missing, respectively in 2018, 2019, 2020 and 2021.

Class size. When class size contained aberrant values (more than 28 children per class), the class size was replaced by NA. This situation corresponded to 0.005% of the dataset, in 2018, 2019, 2020 and 2021.

Missing values on gender. As the outcome used on chapter 4 was the gender gap between children, classes for which gender information was not available were removed from our analysis. A total of 60, 41, 135 and 0 children were removed respectively in 2018, 2019, 2020 and 2021.

Missing values and outliers on four tests: One-minute words and texts reading abilities at T2 and T3. Regarding these four variables, a large proportion of outliers was noted (~ 20 to 25%), resulting from an abnormality in the data collection process: For some children that read all the words and texts in less than a minute, some teachers transmitted their results with a cross product (e.g., if they read 30 words in 30 seconds, then the teachers indicated : “60” out of 30 read words possible). In some schools, teachers were asked to apply a cross product, and some academies added their own cross products to these results as well. Three fourth of the other schools and academies did not apply any cross products on the final results in reading. This situation led to up to ~ a quarter of students that had results overpassing the maximum score expected in these exercises (e.g., 55 words read in a minute / 30 words possible), sometimes presenting with extreme aberrant outliers such as a score of 5547 words read in a min. If we were to take off all the outliers, we would have included an important bias, excluding probable a large part of very good students at the one-minute reading task. Therefore, we decided to apply two modifications on these variables. Firstly, we created a variable with a higher maximum score, going up to 100, 195, 93 and 136 (respectively for reading words at T2, reading texts at T2, reading words at T3 and reading texts at T3), thresholds under which 97% of all students’ results were included. All results above these thresholds were considered as “excessive outlier” and were replaced by NA. Secondly, we decided to keep the four variables with their official maximum (e.g., 30, 29, 60 and 102 respectively), and transformed all the results above these thresholds as NA, therefore inducing a large bias. Both these data modifications were mentioned and described below (see **Table 10**). When computing the composite variables in language, and as results were similar between two types of composite variable in language, we kept the first type of variable including the largest results (i.e., 97% of students’ results).

B. Missing data management and imputation

Missing data imputation.

Among the 2,871,080 children followed up from 2018 to 2022, 122,922, 140,580, 129,153 and 236,898 children (respectively in 2018, 2019, 2020 and 2021), had at least one missing value on the different variables before outlier management. **Table 6** detailed the missing values and their proportions.

All these missing values were assumed as missing completely at random (i.e., a group called MCAR). Some imputation techniques such as removing uncomplete observations or imputing by the mean could have led to an elevated bias both in the analyses and in the conclusion. Therefore, we conducted an imputation by Chained Equations (ICE) on all the missing values to impute missing data using the *mice* package in the R software (Buuren & Groothuis-Oudshoorn, 2011).

Table 6. Overview of missing values among the four different whole-population cohorts and their data management.

Cohort	2018	2019	2020	2021
Initial number of children in the database	610,905	711,452	743,734	804,989
Overall proportion of missing values (%)	1.27	1.18	1.06	4.30
Overall proportion of children with at least 1 missing value (%)	20.12 N = 122,922	19.76 N = 140,590	17.37 N = 129,157	25.58 N = 205,992
Step 1: Number of remaining children after removal of those who were absent from all 3 sessions (T1, T2, T3)	610,830 N classes: 43,970 N classes > 27 : 157 N classes < 6 : 2676 N schools: 27,043	711,351 N classes: 51,599 N classes > 27 = 106 N classes < 6 = 2880 N schools: 30,578	743,606 N classes: 54,073 N classes > 27 = 93 N classes < 6 = 2920 N schools: 31,515	803,767 N classes: 54,224 N classes > 27 = 264 N classes < 6 = 2341 N schools: 31,772
Age	Young (51-68): 3,719 Typical (69-80): 592,779 Late (81-98): 14,219	Young (51-68): 4,426 Typical (69-80): 689,833 Late (81-98): 16,848	Young (51-68): 4,213 Typical (69-80): 720,670 Late (81-98): 18,534	Young (51-68): 5,240 Typical (69-80): 773,423 Late (81-98): 25,880
Step 2: Number of remaining children after removing classes where gender was missing and imputation of missing data	610,785 N classes with gender missing : 3 N students : 45	711,316 N classes with gender missing : 3 N students: 35	743,476 N classes with gender missing : 7 N students: 56	803,767 N classes with gender missing: 0 N students: 0
Gender	Boys: 310,644 Girls: 300,141	Boys: 361,745 Girls: 349,571	Boys: 376,760 Girls: 366,716	Boys: 410,700 Girls: 393,067

<p>Step 3: Selecting children in classes containing 6 to 27 children per class (removing extreme size classes)</p> <p>As this stage we gaussianized the test scores and computed the following class-level variables :</p> <ul style="list-style-type: none"> - boys proportion per class - gender of first of class - Heterogeneity of level in math and language per class 	<p>586,936</p> <p>N classes : 39,573</p>	<p>686,138</p> <p>N classes: 46,671</p>	<p>717,326</p> <p>N classes : 49,010</p>	<p>749,402</p> <p>N classes : 49,703</p>
<p>Step 4: Selecting children of typical age (69 to 80 months) at T1</p> <p>Subject-level regression models of the gender gap were performed on these data</p>	<p>569,771</p>	<p>665,632</p>	<p>695,449</p>	<p>722,230</p>
<p>Step 5 : Selecting classes with at least 30% of boys and 30% of girls</p> <p>Class-level evaluations of the gender gap were performed on these data.</p>	<p>526,556</p> <p>N = 43,215 children belonging to classes with an excess of boys or an excess of girls</p>	<p>614,264</p> <p>N = 51,368 children belonging to classes with an excess of boys or an excess of girls</p>	<p>642,870</p> <p>N = 52,579 children belonging to classes with an excess of boys or an excess of girls</p>	<p>671,793</p> <p>N = 50,437 children belonging to classes with an excess of boys or an excess of girls</p>

C. Sensitivity analysis

All the analyses presented in this paper were performed on the imputed data set, which therefore contained no missing data. To assess whether this decision impacted our conclusions, we performed a sensitivity analysis by comparing the average differences among variables between (1) the imputed dataset and the non-imputed dataset and (2) the imputed dataset and the non-imputed dataset in which we removed all the missing values (see **Table 7**). These analyses confirmed that no systematic difference existed between participants with missing data and those with complete data, especially regarding the results on gender gap. There were, on the other hand, significant differences between the imputed population and the non-imputed population withdrawn from all missing values – demonstrating that analyzing the latter would have exposed to a high risk of bias and non-representativeness (see **Table 7**).

Table 7. Sensitivity analysis presented for year 2018, comparing (1) the selected population (i.e., imputed, using *mice*) to the non-imputed population (i.e., model 1) and comparing (2) the imputed population to the non-imputed population withdrawn from all missing values (i.e., model 2).

	Imputed population	Model 1 : Imputed vs. Original population (Non-imputed population)	P	Model 2: Imputed vs. Original population (Non-imputed population) in which missing values have been removed	P
N	586,936	586,936	-	465,934	-
Age at T1, mean (SD)	74.64 (3.84)	74.64 (3.84)	0.993	74.60 (3.79)	< 0.0001
Class size, mean (SD)	17.21 (5.85)	17.21 (5.85)	1.000	15.54 (5.46)	< 0.0001
Gender - Boys, n (%)	298,642 (50.9)	298,642 (50.9)	1.000	236,448 (50.7)	< 0.0001
Gender - Girls, n (%)	288,307 (49.1)	288,307 (49.1)	-	229,486 (49.3)	-
SES score, mean (SD)	102.36 (17.79)	102.35 (18.21)	0.749	102.99 (18.21)	< 0.0001
Number of children in private schools (%)	63,304 (10.8)	63,304 (11.4)	1.000	56,664 (12.2)	< 0.0001
Number of children in public schools (%)	426,649 (72.7)	426,649 (72.7)	-	331,764 (71.2)	-
Number of children in priority education PE (%)	58,416 (10.0)	58,416 (10.0)	-	47,467 (10.2)	-
Number of children in high-priority education HPE (%)	38,580 (6.6)	38,580 (6.6)	-	30,039 (6.4)	-
Math at T1, mean (SE)	72.77 (13.47)	72.81 (13.56)	0.067	73.29 (13.17)	< 0.0001
Math at T2, mean (SE)	76.43 (17.80)	76.15 (18.25)	< 0.0001	77.73 (17.16)	< 0.0001

Math at T3, mean (SE)	68.75 (18.12)	68.73 (18.17)	0.548	69.79 (17.63)	< 0.0001
Problem solving at T1, mean (SE)	63.34 (30.12)	63.47 (30.09)	0.014	64.27 (29.81)	< 0.0001
Problem solving at T2, mean (SE)	68.60 (28.18)	68.63 (28.21)	0.535	70.92 (27.10)	< 0.0001
Problem solving at T3, mean (SE)	67.93 (27.27)	67.95 (27.26)	0.661	69.48 (26.57)	< 0.0001
Number line at T1, mean (SE)	51.01 (30.61)	51.13 (30.61)	0.026	51.81 (30.52)	< 0.0001
Number line at T2, mean (SE)	54.17 (24.74)	54.20 (24.77)	0.495	55.89 (24.21)	< 0.0001
Number line at T3, mean (SE)	47.11 (24.21)	47.19 (24.21)	0.062	47.94 (24.14)	< 0.0001
Language at T1, mean (SE)	72.69 (15.69)	72.65 (15.80)	0.124	73.51 (15.29)	< 0.0001
Language at T2, mean (SE)	64.62 (13.91)	64.44 (14.61)	< 0.0001	60.18 (10.89)	< 0.0001
Language at T3, mean (SE)	71.29 (16.08)	71.28 (16.36)	0.777	61.82 (11.52)	< 0.0001

D. Creation of new data and composite covariates

Scoring

When comparing all three periods of time, and as tests evolved in nature and difficulty from T1 to T2 and T3, scores could not be directly compared between periods. We therefore decided to normalize all scores into percentage of success (ranging from 0 to 100). For the analyses and several figures, these scores went through a gaussianization process (i.e., variables were centered with a mean of 0 and a standard deviation of 1 and distributions were gaussianized) using the function *gaussianize* in the package *LambertW* in the R software. As we obtained z-score after gaussianization for all continuous variables, we were able to monitor a child's progress between T1, T2 and T3, relative to other children. We used standardized data (z-score after Gaussianization) in all of our regression models.

In addition, when needed to compare specific subgroups (e.g., boys' and girls' results in chapter 4), using the previous standardized data, we implemented a Cohen's d transformation, allowing to measure the variables' effect sizes. This was made for all multilevel regression models and for some figures when mentioned in the legend. For this transformation, we used the function *cohen_d* in the package *rstatix* in the R software.

Another way we used to present results was in percentile ranks, using the R function *rank* with the option 'ties.method' = average in the package *base*. The 'ties.method'

average implied that when two or more children were tied in their results, they were assigned the mean rank of their score (e.g., if two children had the same score of 1, instead of arbitrarily rank them as 1 and 2, they were assigned rank 1.5). Percentile ranks ranged from 0 to 100, 0 being the worst and 100 being the best rank. Using percentile ranks presented two advantages: Firstly, it allowed us to compare math and language tests between T1, T2, T3 even though the tests were different from one period to the next. Secondly, it allowed to assess a measure of progress between T1 to T3 in math and language.

Covariates created at the individual level.

Individual math level at T1. Math at T1 was computed as the mean performance in every normalized math test at the beginning of first grade. For multilevel models and above-mentioned figures, these variables were then gaussianized. The same processes were applied to the following variables: math at T2, math at T3, language at T1, language at T2 and language at T3.

Covariates created at the class level.

Class's mean of math level at T1. For each class, math at T1 was computed as the mean performance in every normalized math test at the beginning of first grade. The same processes were applied to the following variables: math at T2, math at T3, language at T1, language at T2 and language at T3.

E. Statistical analyses

Whenever quantitative variables were compared, Student's t tests were used for comparing two groups and the *CreateTableOne* function of the package *tableone* allowed to measure more groups if needed (using a unidirectional ANOVA), whereas when categorical variables were compared, Chi2 tests were implemented, both using the packages *tidyverse*, *dplyr* and *tableone* in *R software*.

III) Results of Data description

A. Study population's description

After data management and data imputation, we implemented a descriptive analysis of the characteristics of the studied population (see **Table 8** and **Table 9**) and description of the cognitive assessments (see **Table 10**).

Table 8. Description of characteristics of the population continuous variables in 2018 (n = 586,936)

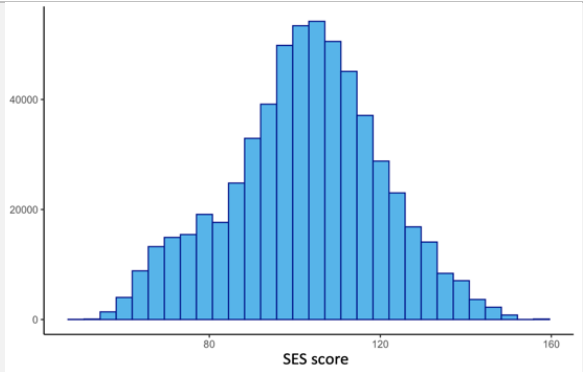
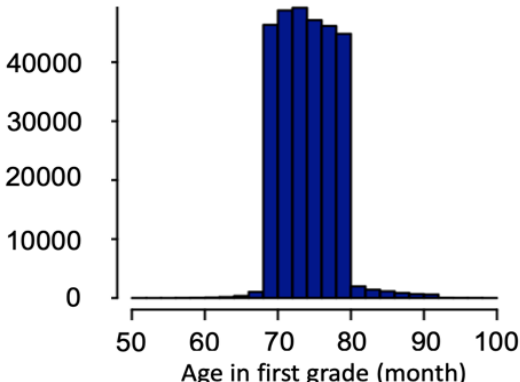
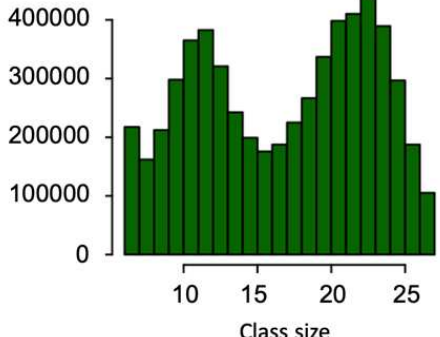
Variables	Mean \pm SD	Median	Distribution
Social Economical Status Range [48.7; 157.6]	102.36 (\pm 17.79)	103.13	
Age in 1st grade (month) Range [51; 99]	74.64 (\pm 3.84)	74	
Class size Number of children per class Range [6; 27]	17.21 (\pm 5.85)	18	

Table 9. Description of characteristics of the population categorial variables in 2018
(n = 586,936)

Variables	Categories	n	%	Distribution per school category
Gender	Male	298,633	50.9	
	Female	288,303	49.1	
Type of school	Private school	63,304	10.80	
	Regular public school	426,643	72.70	
	Priority education public schools (PE)	58,413	10.00	
	Higher priority education public schools (HPE)	38,576	6.60	
				<p><i>*School categories are « added » and not “superposed”. Public schools are the majority of type of school by far</i></p>
Type of School and SES score	Priv sup: Private with SES >= Private median	31,404	5.40	
	Priv inf: Private with SES < Private median	31,900	5.40	
	Pub sup+: Public with SES >= 4 th quartile	107,261	18.30	

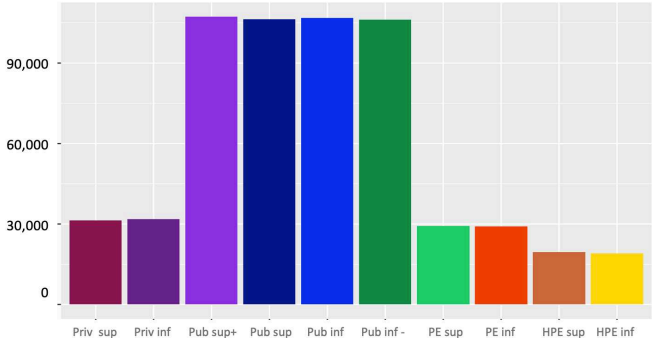
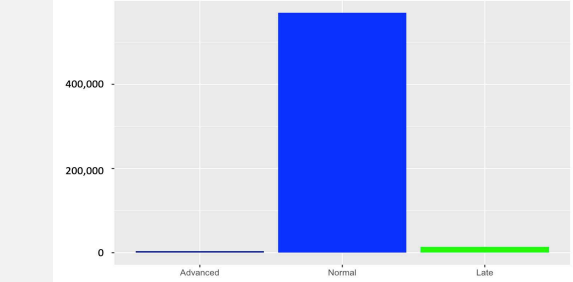
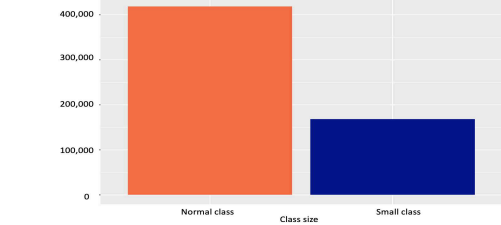
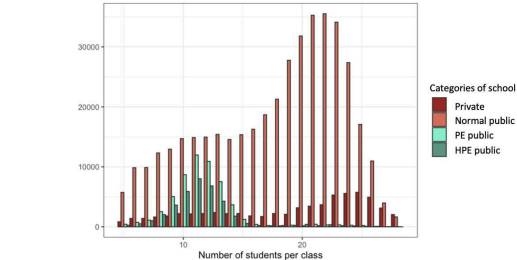
	Pub sup: Public with high SES \geq 3 rd quartile	106,405	18.10	
	Pub inf: Public with SES \geq 2 nd quartile	106,791	18.20	
	Pub inf-: Public with SES $<$ 2 nd quartile	106,186	18.10	
	PE sup: Priority education with SES \geq PE median	29,289	5.00	
	PE inf: Priority education with SES $<$ PE median	29,124	5.00	
	HPE sup: Higher Priority education (HPE) with SES \geq HPE median	19,519	3.30	
	HPE inf: Higher Priority education (HPE) with SES $<$ HPE median	19,057	3.20	
Age in first grade	Advanced ($<$ 69 months)	3,519	0.60	
	Typical (69 to 80 months)	569,755	97.1	
	Late ($>$ 80 months)	13,662	2.30	
Class sizes	Small class ($<$ 13/ class)	168,566	28.70	
	Typical class (\geq 13 / class)	418,370	71.30	
				

Table 10. Descriptive analysis of cognitive tests in 2018 (n = 586,936).

Assessments	Raw results, mean (Standard error)	Percentage of assessment success (Range 0-100) mean (Standard error)
N students	586,936	586,936
<i>Language mean</i>		
T1 Oral Comprehension of Words, range 0-15	11.89 (2.71)	79.27 (18.09)
T1 Oral Comprehension of Sentences, 0-14	12.13 (2.17)	86.67 (15.50)
T1 Oral Comprehension of Texts, 0-18	13.27 (3.64)	73.71 (20.23)
T1 Phoneme handling, 0-15	8.90 (3.77)	59.30 (25.12)
T1 Syllable handling, 0-15	11.49 (3.11)	76.62 (20.76)
T1 Letter-sound association, 0-10	7.44 (2.55)	74.43 (25.54)
T1 Decoding, letter writings recognition, 0-7	4.77 (1.84)	68.16 (26.25)
T1 Comparing letters, visuo-attentional abilities, 0-24	15.23 (6.69)	63.45 (27.89)
T2 Oral Comprehension of Sentences, 0-14	12.22 (1.89)	87.30 (13.49)
T2 Reading a list of 30 words in 1 minute, 0-100	24.97 (17.36)	24.93 (17.34)
T2 Reading 29 words of a text in 1 minute, 0-195	29.83 (25.74)	15.28 (13.19)
T2 Reading a list of 30 words in 1 minute, 0-30	19.59 (7.54)	65.28 (25.13)
T2 Reading 29 words of a text in 1 minute, 0-29	19.17 (8.07)	66.10 (27.83)
T2 Writing Syllables, 0-10	7.95 (2.37)	79.49 (23.71)
T2 Writing Words, 0-8	5.91 (2.22)	73.89 (27.79)
T2 Phoneme handling, 0-12	9.31 (2.76)	77.55 (22.96)
T2 Decoding, Letter recognition, 0-10	9.39 (1.28)	93.93 (12.80)
T3 Oral Comprehension of Words, 0-15	13.41 (2.01)	89.38 (13.43)
T3 Oral Comprehension of Sentences, 0-15	13.79 (1.65)	91.95 (10.98)
T3 Writing Syllables, 0-12	9.65 (2.87)	80.45 (23.91)
T3 Writing Words, 0-12	8.62 (3.11)	71.87 (25.93)
T3 Understanding reading a sentence, 0-10	7.78 (2.45)	77.84 (24.49)
T3 Understanding reading a text, 0-8	5.97 (2.03)	74.61 (25.39)
T3 Reading a list of 60 words in 1 min, 0-93	43.49 (19.06)	46.76 (20.50)

T3 Reading 102 words of a text in 1 min, 0-136	50.98 (31.21)	37.48 (22.95)
T3 Reading a list of 60 words in 1 min, 0-60	40.49 (15.16)	67.49 (25.27)
T3 Reading 102 words of a text in 1 min, 0-102	49.04 (27.94)	48.08 (27.39)
Math mean		
T1 Writing numbers, 0-11	10.28 (1.65)	93.42 (14.96)
T1 Reading numbers, 0-10	9.70 (0.97)	96.96 (9.67)
T1 Problem solving, 0-6	3.80 (1.81)	63.34 (30.12)
T1 Enumerate quantities, 0-8	7.49 (1.15)	93.58 (14.41)
T1 Associate number to quantity, 0-60	22.99 (14.34)	51.01 (30.61)
T1 Number line, 0-6	3.06 (1.84)	38.31 (23.90)
T2 Comparing numbers, 0-40	36.58 (7.41)	91.45 (18.51)
T2 Number line, 0-10	5.42 (2.47)	54.18 (24.73)
T2 Additioning, 0-7	5.74 (1.68)	81.98 (23.97)
T2 Subtrationing, 0-7	5.09 (2.46)	72.72 (35.17)
T2 Writing numbers, 0-10	8.97 (1.87)	89.72 (18.65)
T2 Problem solving, 0-5	3.43 (1.41)	68.62 (28.17)
T3 Geometry, 0-8	5.80 (1.64)	72.56 (20.51)
T3 Number line, 0-15	7.07 (3.63)	47.11 (24.21)
T3 Additioning, 0-7	4.39 (2.20)	62.69 (31.50)
T3 Subtrationing, 0-8	3.78 (2.64)	47.23 (33.00)
T3 Mental calculus, 0-10	8.43 (2.04)	84.32 (20.38)
T3 Writing numbers, 0-10	8.29 (2.44)	82.93 (24.43)
T3 Reading numbers, 0-10	8.52 (2.03)	85.24 (20.35)
T3 Associate number to quantity, 0-16	9.69 (3.93)	60.56 (24.58)
T3 Problem solving, 0-6	4.08 (1.64)	67.93 (27.27)
Composite variables		
Language at T1	-	72.70 (15.68)
Language at T2	-	64.64 (13.89)
Language at T3	-	71.29 (16.08)
Math at T1	-	72.77 (13.45)
Math at T2	-	76.45 (17.78)
Math at T3	-	67.84 (17.81)

Results found in **Table 8**, **Table 9** and **Table 10** were replicated in 2019, 2020 and 2021 and showed in supplementary materials (see **Table S1**).

B. Description of children regarding their age category when entering in first grade

In France, students normally enter first grade the year they turn 6. However, some were born at the beginning of the year, and others at the end of the year. Thus, among students who are neither early nor late, there was a difference of up to 12 months between students in the same class. In 2018, most children (97.06%) entered first grade the year of their 6th birthday, with a mean of 74.64 (± 3.84) month-old, a smaller part (2.36%) presented learning difficulties (either in math or language) justifying entering in first grade with a year of delay (on the year of their 7th to 8th birthday). An even smaller part (0.61%) of children attended their first grade on the year of their 5th birthday due to advanced skills in preschool (**Table 8** and **Table 9**).

Table 11. Description of gender and school categories for advanced and late children when beginning first grade (T1) in 2018

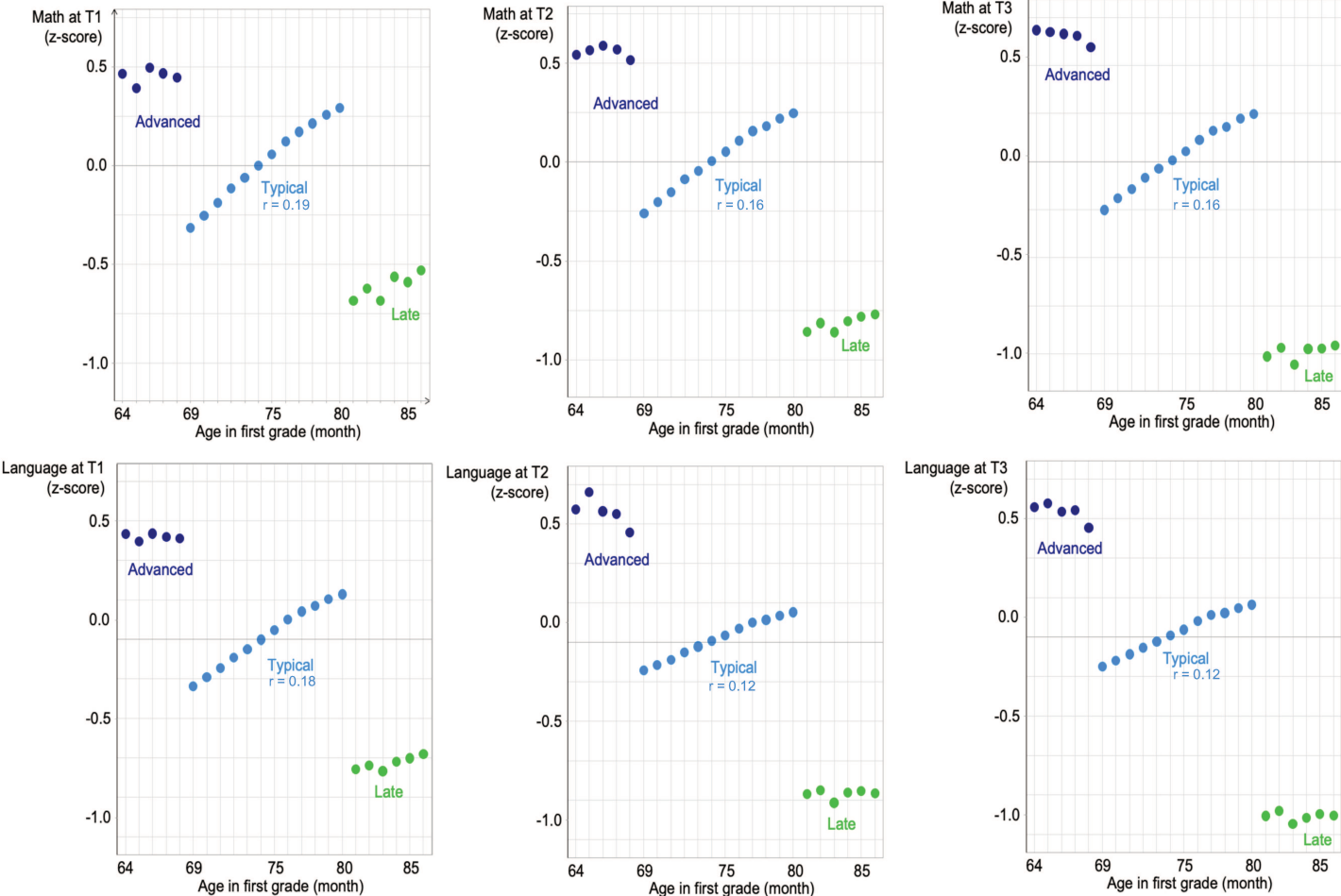
<i>Variable</i>	<i>Category</i>	T1			
		Private schools	Regular public schools	PE schools	HPE schools
School categories, % (n)		10.8 (63,304)	72.7 (426,643)	10.0 (58,416)	6.6 (38,580)
Gender					
	Boys, %	51.1	50.9	51.0	50.1
Age					
	Advanced, %	1.0	0.6	0.5	0.5
	Boys, %	45.3	45.3	43.5	39.9
	Late, %	2.0	2.1	3.5	3.1
	Boys, %	62.1	62.3	62.3	59.7

As presented in **Table 9** for 2018, even if overall boys (50.9%) outnumbered girls (49.14%), more girls were one year ahead at school in first grade (55-60%) compared to boys (**Table 11**). This more important proportion of advanced-in-age girls could be explained by the selection on oral language abilities for dropping a school year

between preschool and school, where girls present with a slight advance in oral language skills compared to boys (Etchell et al., 2018). On the opposite, the proportion of boys was higher in the category with a year of delay at school in the first grade, compared to girls (i.e., 60% of children who were one year behind were boys) (**Table 10**). These results were replicated in 2019, 2020 and 2021 (see **Table S2** in supplementary material).

Among children who entered in first grade on the year of their 6th birthday, a linear relationship indicated that elder children presented with an advantage in both language and mathematic skills compared to younger children and this relation was linear regarding the month of age (see **Table S3** and **Table S4**). In addition, advanced-in-age children constantly presented with higher language and math skills compared to typical-age children (see **Figure 5**). On the other end, late-in-age children constantly presented with worst results in all domains compared to typical-age children. All tests were described per age category and presented in supplementary materials (see **Table S3**). As an illustration, **Figure 5** showed language and math scores at T1, T2 and T3 across the entire population of 586,936 students entering first grade in 2018, by age at first grade entry (in months). The linear relationship observation for typical age children was not seen for the children in advance (represented in **dark blue**) neither for those already one year late (represented in **green**) as shown in **Figure S2** in supplementary material, comprising children's age ranging from 50 to 99 months. These trends were replicated at T2 and T3 in both math and language (**Figure 5**). As we will consider only typical-age children in the following work, we presented the following results with a scale focused on typical-age children and presented only 5 to 6 months of age among both advanced-in-age children and late-in-age children to facilitate the data visualization (see **Figure 5**).

Figure 5. Effect of age in Math and Language, reproduced in 2018 (main figure), 2019, 2020 and 2021. For children within the typical age range, the yearly data was precise enough to detect a strictly monotonic effect of age in months at each time period (T1, T2 and T3). For children with accelerate or delayed schooling, a large and increasing learning gap is detected. Bars, indicating one standard error, are often too small to be visible. Math and language level at T1, T2 and T3 as a function of age when beginning first grade (in month). Math and language at T2 and T3 followed similar dynamics and were presented in the supplementary material. Ages on the x-axis were cut 6 months below and above the typical age limits for a clearer data visualization of math/language in function of age.



Among younger students who have benefited from a waiver (the group in **dark blue**), results were higher than the average, regardless of the students' precise age. Here, our national education system seemed to be too conservative: it only allowed a few students to enter first grade before their 6th birthday, whose cognitive development was clearly very advanced (it reached or even exceeded that of their peers who were one year older), and moreover they maintained their lead in second grade (see **Figure 5**). On the other hand, among older students (the group in **green**), results were below average, and again, regardless of the students' age. This category probably included a wide variety of situations, the nature of which we did not have access: repeating grades, allophone students, handicaps, etc. These students were clearly behind, not only compared to students of comparable age, but also in relation to students of one year younger. All means in math and language cognitive tests were significantly different and tests in **Table S3** between the three different categories of age (see **Table S3** in supplementary materials) and (see **Table S4**). In addition, a linear effect was found for typical-in-age children and not for advance-in-age nor late-in-age children (see **Table S5**) where the age effect in math was diminishing for every additional month of age, and the age effect was significantly less important at the beginning of second grade compared to the beginning of first grade (**Table S4**). On the opposite, language presented with an effect of age that rose up with time and with additional months of age, the age effect was significantly more important in second grade compared to first grade (**Table S4**).

In detail and as shown in **Table S3**, the younger children (age < 69-month-old, labeled as "advanced") presented with significant higher performances at all periods (T1, T2, T3) and in both domains (math and language) compared to the other age categories (i.e., 1.64 to 2.33 points more in math and 2.6 to 3.78 points more in language when comparing with the typical age population of students). Whereas elder children (age > 80-month-old, labeled as "late") presented with the lowest results for all periods and domains (i.e., up to 29.28 points less in math and 30.61 points less in language when comparing with the 1-year-ahead population of students, 19.37 points less in math and 20.81 points in language from the typical-aged-students). Also, there were no tasks where the youngest typical-age children's average overpassed the level of the oldest

typical-age children's average from T1 to T3, a linear relationship between age in month and math and language results was shown in **Table S4**. Whereas children being a year in advance when entering in first grade presented better results in both language and math compared to the oldest typical-aged children, and children being a year late when entering first grade presented with lower results both in language and math. There were no tasks where the oldest typical-age children overpassed the level of youngest one-year-in-advance children. And on the other hand, even with an additional year of age, children that were one year older (and late when entering first grade) did not manage to catch up a correct oral language level nor math level compared to typical age children (see **Table S3** in supplementary material). Owing to the large samples, all comparisons were highly significant (all $p < 0.0001$, **Table S3** in supplementary material). Similar results were found in 2019, 2020 and 2021 (see **Figure S3** in supplementary materials). These three-age-groups specific differences were significant and presented in **Table S5** in supplementary materials). As these two age-extreme categories presented with specificities, we decided not to explore them in this work, and focused on the typical-age children when beginning first grade.

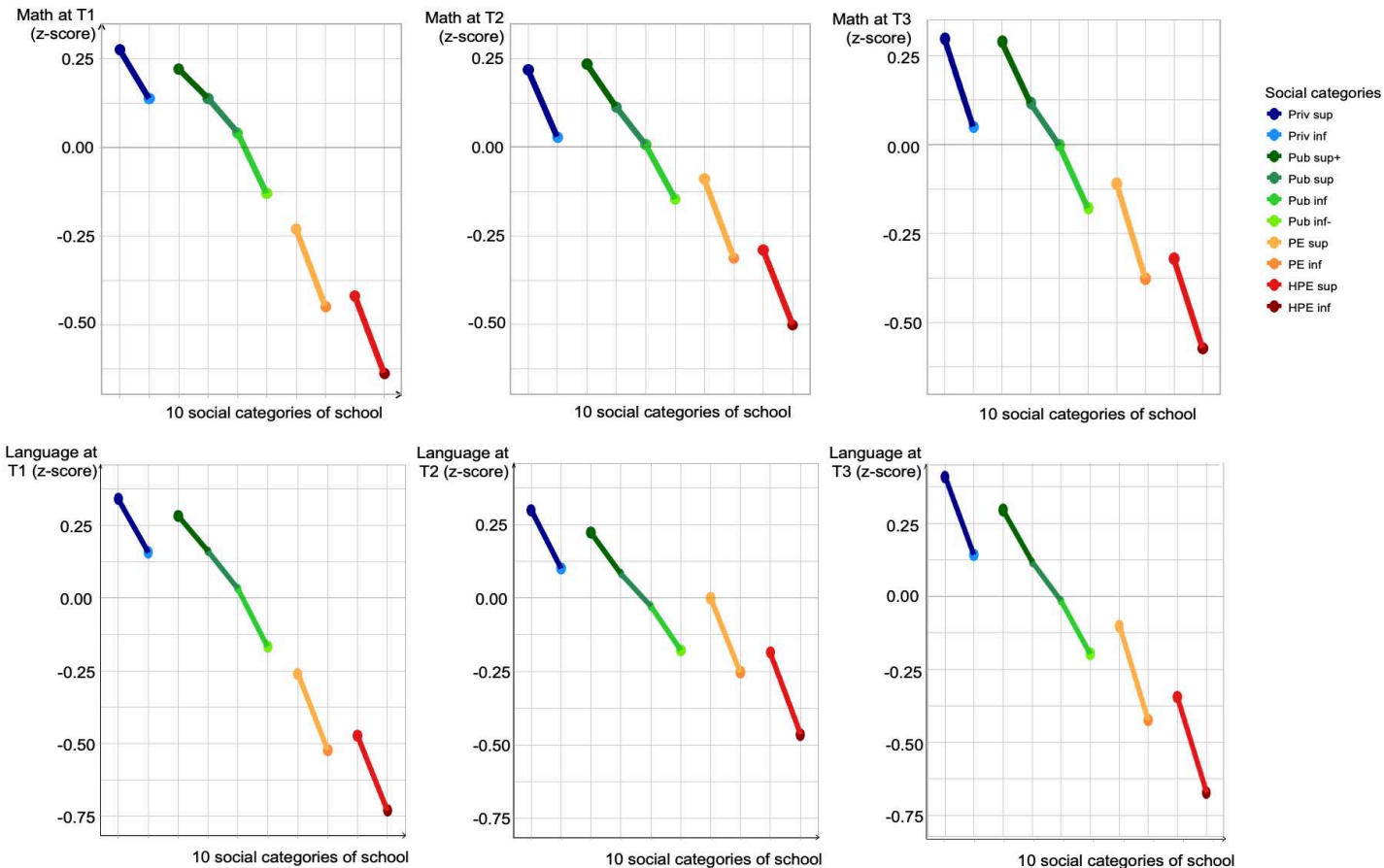
C. Description of children regarding their school category when entering in first grade

After the age categories, we wondered what the SES and school category effects were on the math and language results of children at T1, T2 and T3, as well as the effects on gender inside each school category. The SES score (i.e., a continuous variable ranging from ~ 50 to ~ 150 points according to the year, with 50 indicating the lowest SES score) was cross tabulated with the school categories (i.e., the variable of 4 school categories (i.e., private, regular public, PE and HPE public schools)), creating the following 10 sub-categories: the split was done on the median of the SES score for private schools, PE schools, and HPE schools. As a large number of children attended regular public schools, this category was divided into 4 subgroups on the basis of SES score quartiles (**Figure 6**).

Regarding the categories of schools, most children went to regular public schools in France (72.70%), 10.80% went to private schools, and 16.60% went to priority or higher priority education public schools. These proportions remained similar among the four cohorts (**Table S2**). We noted that twice as many children were one year ahead (1%) in private schools, compared to the other categories (0.5%). Conversely, PE and HPE public schools included a higher proportion of children with delay (3.11 to 3.55% compared to 1.98% in private schools and 2.15% in public schools) (**Table 11**). Similar proportions were found in 2019, 2020 and 2021 (see **Table S2** in SOM).

Math and language results at T1, T2 and T3 showed a significant gradient of achievement with always higher scores for private schools and regular public schools with higher SES scores, while PE and HPE and regular public schools with lower SES scores presented with lower results both in math and language (see **Table S6** and **Table S7** in SOM). Interestingly at T2, after 4 months of schooling, all the level gaps were brought closer to the average and SES gaps were reduced. However, the SES gap widened again 12 months after the beginning of first grade (T3), notably after the 12-month of summertime vacation (i.e., called the summertime vacation effect by others (Hammerstein et al., 2021; Shinwell & Defeyter, 2017)), with levels in math and language that worsened for children belonging to PE and HPE public schools and showing a larger advantage for children with a higher SES score compared to lower SES scores (see **Figure 6**). Similar results were found in 2019, 2020 and 2021 (see **Figure S4** in supplementary materials).

Figure 6. Effect of school categories and SES score in Math and language, reproduced in 2018 (main figure), 2019, 2020 and 2021. For each of four school categories (private schools, or regular, priority education [PE], and higher-priority education [HPE] public schools), a median split or quartile split (for regular public schools only) was implemented based on the school average socio-economic status (SES). Each school category comprised 2 points, on the left of the x-axis would stand the highest SES score and, on the right, the lowest SES score, for a total of 10 school subcategories: 2 median-split for private schools (in blue), 4 quarter-split for regular public schools (in green), and 2 median-splits for PE (in orange) and for HPE public (in red) schools. Disparities at the start of 1st grade remained present at subsequent time points, apart from PE and HPE schools whose gap decreased during schooling (T2) and increased again after the summer break (T3).



Regarding the type of language exercise at T1, a larger SES difference was found for “oral language comprehension” compared to decoding and visuo-attentional abilities: In oral comprehension level at T1, there were about 20 points (over a 100 points) of difference between children’ results in private schools and in HPE schools. For other tasks, the difference was around 10 points between private and HPE schools (see **Table S6** and **Table S7** in supplementary materials). Regarding the time-limited exercises (i.e., comparing letters at T1, reading words and texts at T2 and T3), the gradient was similar with higher scores for children attending private schools compared to regular public schools and a difference of 10 to 12 points (over 100 points) between private and HPE schools. In other words, time-limited exercises favored children with a higher SES score compared to children with a lower SES score. Regarding difficult exercises (i.e., involving language oral comprehension and mathematical reasoning for problem solving at T1, T2, and T3), there were 15 to almost 20 points (over 100 points) of difference in favor of private schools when compared to HPE public schools. At T3, understanding a self-read sentence or a self-read text – which is the aim of reading in first and second grade - presented with up to 20 points of difference between children’ results in private schools compared to HPE public schools. Regarding new exercises (i.e., number line at T1, T2 and T3), there were 15 to almost 20 points (over 100 points) of difference in favor of private schools when compared to HPE schools. Therefore, difficult and new exercises were better performed by higher SES children (see **Table S6** in supplementary materials). These differences were confirmed by implementing several models with interaction in **Table S7** and **Table S8** in SOM.

D. Correlation matrices between all cognitive tests

After describing the effect of age in first grade and school categories on math and language test scores, we wondered how all cognitive tests were related to each other and varied over time, and how class variables were associated with them.

We used correlations as the first step in describing the simple relationships between each test. Using correlation matrices, we intended to answer the following questions: (1) How correlated are the class level variables (i.e., SES score, class size, boy

proportion, heterogeneity of level in math at T1, heterogeneity of language at T1, class mean in math at T1, class mean in language at T1)? (2) Is the heterogeneity of level in a class correlated to the SES score? It is hypothesized that the higher the classes' SES scores are, the more homogeneous and of a higher level the classes are, compared to lower SES where a higher heterogeneity of level and a lower level is expected. (3) How correlated are all the cognitive tests at T1?

 **Tips for results interpretation**

Correlation is a statistical measure that expresses how strongly two variables are linearly related, meaning that they evolve together at a constant rate.

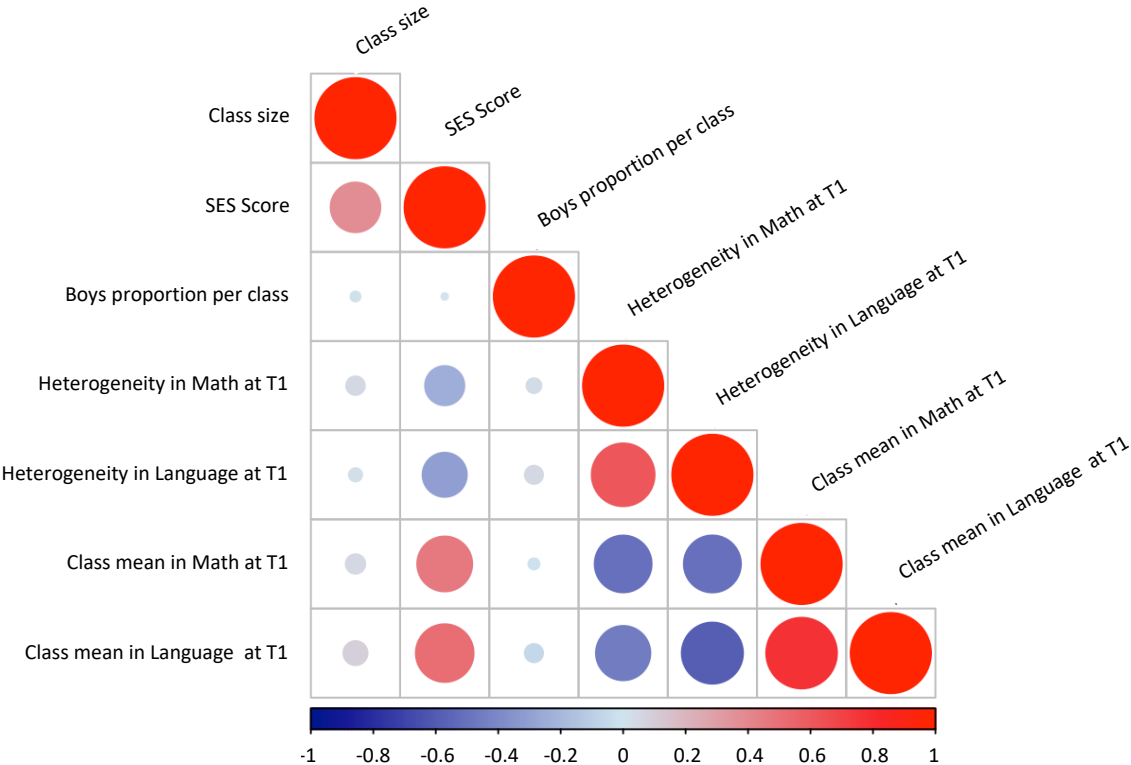
- It is a common tool for describing simple relationships without making statements about cause nor effect. The sample correlation coefficient, r , quantifies the strength of the relationship. It ranges from -1 to +1.
- The closer r is to zero, the weaker the linear relationship.
- Positive r values indicate a positive correlation, where the values of the two variables tend to increase together.
- Negative r values indicate a negative correlation, where the values of one variable tend to increase when the values of the other variable decrease.
- Non-significant correlations are leaved blank in the correlation matrix.

Correlation matrix were performed using the R package *corr*. Significance (p-values) were obtained with the *rcorr* function in the R package *Hmisc*.

Firstly, as expected, larger class sizes were associated with a higher SES, following the political decision to reduce class sizes to a maximum of 13 children per class for priority education and higher priority education schools, a decision which was progressively implemented during those four years (**Figure 7**).

We noted that classes with higher SES scores were negatively associated with heterogeneity in math and language performances at T1, indicating that classes with higher SES scores were more homogeneous in terms of performance level. In addition, the class mean in math and language at T1 was positively associated with SES scores, indicating that higher SES classes were more homogeneous classes with higher mean performances in both math and language at T1 (**Figure 7**).

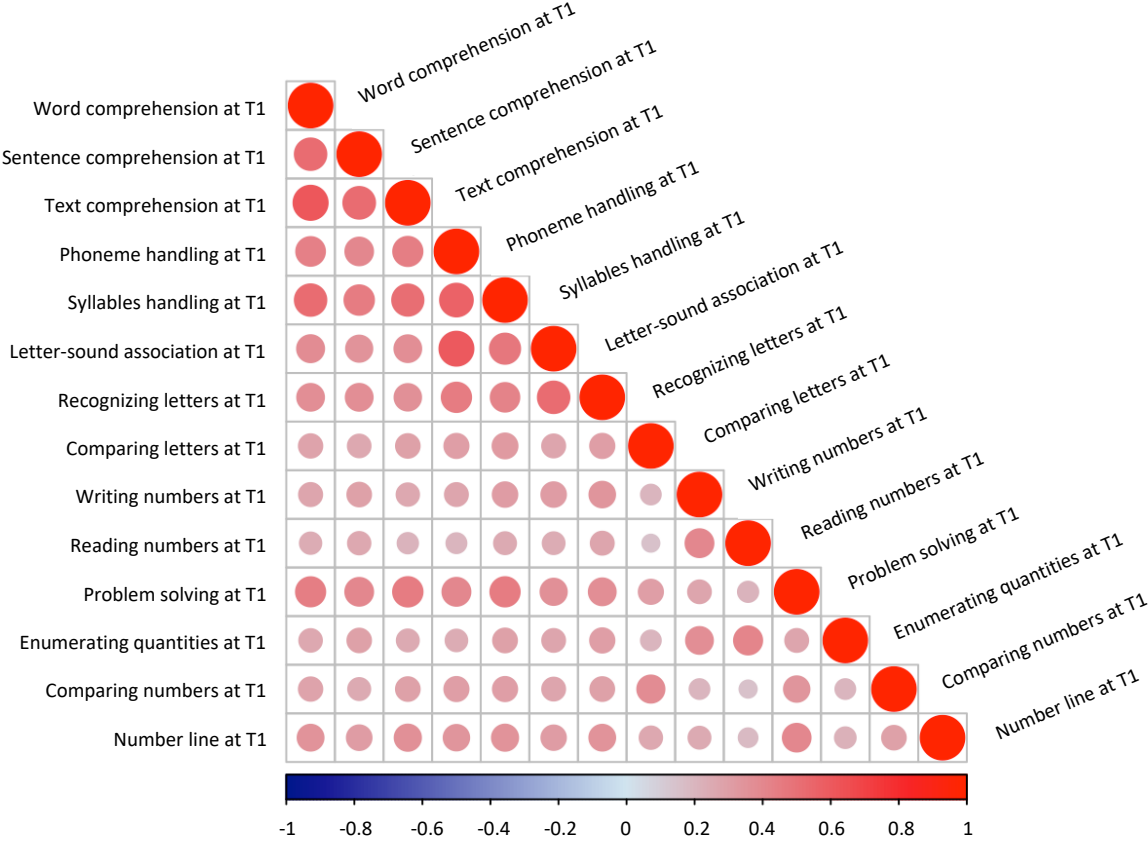
Figure 7. Correlation matrix between all class and school level variables at T1 in 2018, on the total population.



In other words, there were larger performances disparities in the same class and lower mean in both language and math domains among classes with lower SES scores.

In addition, when the class average was high in mathematics, the class average in language was also higher ($r = 0.73, p < 0.0001$). Individual tests correlation matrices at T1 indicated that all tests were positively correlated and confirmed that some specific tests had a stronger interaction - especially when assessing the same cognitive domain. Problem solving had a particular profile being strongly correlated with both tests in language (i.e., the problem comprehension) and math domains (**Figure 8**).

Figure 8. Correlation matrix for cognitive tests both in math and language at T1



IV) Conclusions

Summary of the novel elements identified in this chapter concerning learning trajectories.

Thanks to Evalaide, a programme carried out on the entire generation of children over four years (2018-2021) in France, we observed that age difference had a very strong impact on children's results both in language and mathematics tests, regardless of the school category. Because children in advance or delayed might correspond to specific biological, environmental, and cultural backgrounds, we decided not to explore them in this thesis work, and we focused on the children with the typical age in first grade in France. Regardless of the school category and age in months, normally aged children performed better for every task, both in language and math, in T1, T2, and T3 when they belonged to private schools, compared to public schools, with a performance gradient as follows: private > regular public > PE public > HPE public schools. Time-limited exercises as difficult and new exercises favored children with a higher SES.

Differences between social categories of students were reduced after 4 months of schooling, whereas social inequalities deepened between T2 and T3, where a break of 2.5 months of summer vacations happened. It was within the priority education (PE) and the higher priority education (HPE) public schools that children made the most progress in language and math from T1 to T2. However, it was also among these subgroups that they made less progress in language and math from T2 to T3.

Regarding gender, more girls were one year ahead at school in the first grade (55-60%) compared to boys. The proportion of boys was higher in the category with a year of delay at school in the first grade, compared to girls (60% of children who were one year behind were boys). In private schools, twice as many children were one year ahead (1%) compared to other categories of schools (0.5%). There were larger level disparities (i.e., class heterogeneity of level) in the same class in both language and math domains among classes with lower SES scores. Classes with higher SES had more homogeneous performances per class and higher-class averages in math and language at T1.

V) Supplementary materials

A. Tests' contents

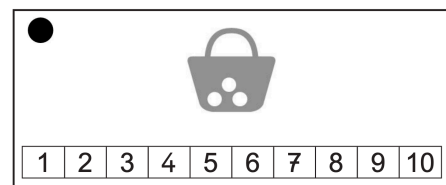
Note that the tests' descriptions and sources below were presented and detailed based on the information found on the document "2. Evalaide, évaluer pour mieux aider", presented on the "conseil scientifique de l'éducation nationale" website.

1) Math tests

Number reading (T1, T3). In this test, the teacher stated a number orally, and the student had to choose and circle the corresponding Arabic number among 6 possibilities. 10 different numbers were assessed at T1 and were the following "3 – 5 – 8 – 2 – 7 – 10 – 6 – 4 – 9 – 0" and "29 – 67 – 90 – 64 – 76 – 54 – 98 – 73 – 83 – 89" at T3.

Number writing (T1, T2, T3). The teacher stated a number orally and the student had to write it down. The numbers went from 0 up to 10 at T1, up to 31 at T2, and up to 100 at T3. Children were exposed to 11, 10 and 10 read-aloud-numbers, respectively at T1, T2, T3. At T1, the numbers were the following "3 – 5 – 1 – 4 – 2 – 6 – 9 – 0 – 8 – 10 – 7".

Enumerating a concrete set (T1, T3). Children viewed a collection of eggs in a basket and had to select the corresponding Arabic numeral on a number line with 10 cells marked and graduated from 1 to 10. 8 items composed this exercise at T1. At T3, children had to count figures in dominos and associate them with their Arabic numeral. 16 items composed this exercise at T3.



Number comparison (T1, T2). Students had to cross out the larger of two Arabic numerals, presented side by side. The assessment at T1, adapted from the *Belgian Symp test* (Brankaer et al., 2017), included 60 pairs of numbers between 0 and 9, half of the pairs being distant by one unit, the other by 3 to 4 units. There was a time limit of 1 minute, after which the test was stopped. This assessment was replicated at T2 to assess student's progress, with 40 items, and with a time limit of 2 minutes.

1	3
6	3
4	5
9	7

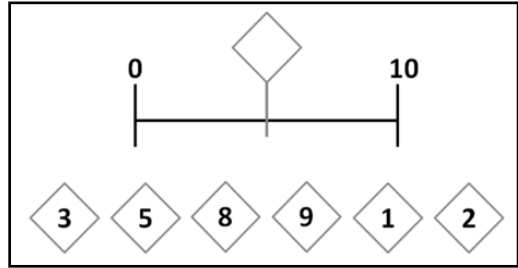
Problem solving (T1, T2, T3). Problem solving simultaneously involved language comprehension and arithmetic skills. In this task, students heard **6 oral arithmetic problem read** by the teacher at T1 (5 **oral arithmetic problem read** at T2 and 6 problems at T3), and also had the possibility to read the corresponding written sentences (at T3) – for instance “Lucie had one marble, and now she has seven. How many marbles did she win?”. The child had to find the correct answer among 6 choices.

French version : 6 poules veulent aller couvrir 1 œuf chacune. Il y a seulement 3 œufs. Combien d'œufs doit-on ajouter pour que chaque poule couve un œuf ?

English version : 6 hens want to go and hatch 1 egg each. There are only 3 eggs. How many eggs must be added for each hen to incubate one egg?

The numbers involved respected the range of numbers introduced in the national curriculum: numbers below ten at T1 and T2, and 2-digit numbers at T3. All the statements were read by the teacher, and children had one minute and thirty seconds to respond to each of them.

Number line (T1, T2, T3). On each trial, the child saw an ungraduated horizontal line marked at both ends with some reference numbers (e.g., 0 at left and 10 at right). One location was marked with a vertical bar and a diamond shape. The children had to figure out which number corresponded to this bar, and to select it among 6 possible choices that were proposed, in randomized order, below the line.



An example appears at right:

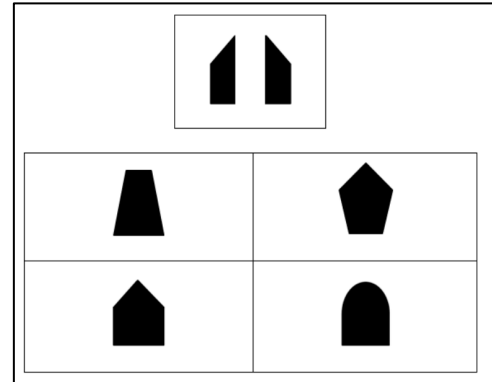
At T1, all 6 target lines ranged from 0 to 10. At T2 (10 items) and T3 (15 items), the endpoint labels could vary and could include 2-digit numbers, and therefore exposed students to problems of different levels of difficulty. For example, one item involved finding the middle of two close numbers (12 and 14), another the middle of two numbers that were further apart (2 and 6), and a third required finding how to proceed when the segment was not in the middle (e.g., 17 when the line goes from 10 to 20). This exercise was modified in 2021, where children previously had to place a number on a numerical line, and from 2021, children had to place a number on a graduated numerical line.

Addition (T2, T3). Addition problems were presented in written form in Arabic numbers to the children (e.g., “ $2 + 3 =$ ”). They had to select the correct answer among 6 choices. 7 problems were presented at T2 and T3 in 2018, and respectively 10 and 8 problems at T2 and T3, both in 2019, 2020 and 2021.

Subtraction (T2, T3). Subtraction problems were similarly presented. There were 7 (at T2) and 8 (at T3) problems in 2018, and respectively 10 and 7 problems at T2 and T3 for both 2019, 2020 and 2021.

Mental calculation (T3). Students were asked to perform arithmetic calculations (additions and subtractions) without the support of a written medium. 10 elementary spoken arithmetic problems were presented to the child (e.g., “ $10 - 2$ ”), who had to select the correct answer among 6 choices.

Geometry (T1, T3). The geometry tests were only introduced at T3 in 2018, and at both T1 and T3 in 2019, 2020 and 2021. The tests were adapted from a prior intruder test (Dehaene et al., 2006). In each of 16 boards (plus a practice one), children had to identify the intruder item among 4 possible choices (i.e., the shape that deviated from the others in a certain geometric property). The different boards evaluated the concepts of straight line, parallelism, mirror image, right angle, distance, circle, alignment, and spacing.



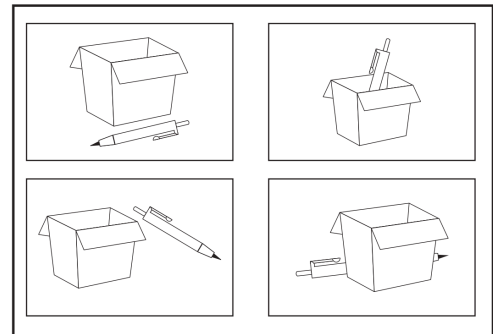
2) Language tests

Oral comprehension of words (T1, T3). In a word-to-picture matching task, children had to circle, among 4 pictures, the one that matched the word read aloud by the teacher. Words were either nouns or verbs. Two of the distractor images had either a pronunciation close to that of the target word – i.e., phonological distractors (e.g., “pédale” and “pétale”), or a semantic relationship with the target word (i.e., having the same function or category). Therefore, this test allowed to assess both the amount of vocabulary a child knew and the nature of his mistakes. The test consisted of 15 words at T1 and at T3, that were the following: ”Hiver – rire – clou – coudre – voile – cacher – pédale – scier - s’éveiller – courir – briser – tronc – quille – coude – orage ” (English version : “Winter - laugh - nail - sew - sail - hide - pedal - saw - wake up - run - break - trunk - keel - elbow - storm”). The exercise and the words were the same at T1 and at T3 which allowed a comparison between a 12-month period for every child.

Oral comprehension of sentences (T1, T2, T3). This test was adapted from the French syntactical and semantical comprehension E.CO.SSE. test (Lecocq, 1996), itself inspired by the *Test for Reception of Grammar (TROG)* in English (Bishop, 2003). Children had to circle, among 4 pictures, the one that matched the sentence read aloud by the teacher. The level of difficulty was increasing while pursuing the exercise, starting with simple sentences (e.g., subject-verb-complement) going towards more

complex sentences. Different kinds of syntactical structures were used, containing for instance spatial prepositions, active or passive sentences. A total of 15, 14 and 16 sentences were tested at T1, T2 and T3 respectively.

E.g.: “The pen is in the box” followed by the associated 4 items (here presented on the right).



At T1, the sentences were the following: “Le stylo est dans le carton ; Le garçon court ; La tasse est

grande ; Le chien n’est pas dans la niche ; La fille ne court pas ; La dame les regarde ; Le garçon la porte ; L’étoile est dans le cercle ; Le bol est derrière la tasse ; La tasse est devant le crayon ; La dame marche ; Le vélo suit la moto ; Le garçon les regarde ; La dame le porte ; Le crayon est sur le cahier. » (English version «The pen is in the box; The boy is running; The cup is big; The dog is not in the kennel; The girl is not running; The lady is looking at them; The boy is carrying her; The star is in the circle; The bowl is behind the cup; The cup is in front of the pencil; The lady is walking; The bike is following the motorcycle; The boy is looking at them; The lady is carrying him; The pencil is on the book. »).

At T2, the sentences were the following: “ Le chien n’est pas dans la niche ; La fille ne court pas ; Le vélo est suivi par la moto ; La dame les regarde ; Le garçon la porte ; Le chat est grand mais pas noir ; Le vélo suit la moto ; L’étoile est dans le cercle ; Le bol est derrière la tasse ; La tasse est devant le carton ; La voiture est suivie par la moto ; Le garçon est poussé par la fille ; Le garçon les regarde ; Le crayon est sur le cahier. »

At T3, the sentences were the following: “Le stylo est dans le carton ; Le chien n’est pas dans la niche ; La fille ne court pas ; Le vélo est suivi par la moto ; La dame les regarde ; Le garçon la porte ; Le chat est grand mais pas noir ; La dame est debout mais pas le garçon ; L’étoile est dans le cercle ; Le bol est derrière la tasse ; La tasse est devant le carton ; La voiture est suivie par la moto ; Le garçon est poussé par la fille ; Le garçon les regarde ; La dame le porte ; Le crayon est sur le cahier. »

Oral comprehension of texts (T1). To assess the comprehension of spoken small texts, 18 questions about 4 small texts at T1 (in 2018) and 11 questions about 3 Texts

(Texts 1, 3 and 4) at T1 (in 2019, 2020 and 2021) were read aloud by the teacher and were of increasing length and complexity. For every story, children had to circle the corresponding image among 4 items.

French version

Text 1 : *Ce matin, papa prépare Mathieu pour aller à l'école. Il lui enfle ses bottes, lui dit de bien garder son manteau et sa capuche pendant la récréation. « Tu feras attention de ne pas trop te mouiller quand tu sortiras de la classe et je te demande de ne pas sauter dans les flaques pour ne pas éclabousser tes camarades. »*

- *Où se passe l'histoire ? Dans une maison, dans une forêt, dans la rue, à la piscine.*
- *Quel temps fait-il dans cette histoire ? Il y a de la neige, du soleil, de la pluie, du vent.*
- *Entoure ce que le papa interdit de faire à Mathieu. Il lui interdit de jouer aux billes, de grimper sur la barrière, de sauter dans les flaques, d'éclabousser ses camarades avec son vélo.*

Text 2 : *Hector a décidé qu'il était maintenant assez grand et qu'il pouvait se débrouiller tout seul dans la vie. Ce matin, il quitte le terrier de sa famille, court dans les champs vers une forêt et se met à la recherche de son plat préféré : des carottes. Tout à coup, il entend un aboiement et voit arriver vers lui un drôle d'animal à quatre pattes avec un collier autour de son cou et qui renifle le sol sans s'arrêter... Hector se cache et aperçoit derrière le drôle d'animal un homme qui tient dans ses mains un long tube bizarre. L'homme crie au drôle d'animal qui renifle toujours le sol : « Cherche, cherche !! Trouve une piste !... » Hector a très peur et préfère revenir très vite dans le terrier familial : il détale, mais à peine a-t-il fait quelques mètres qu'il entend, grâce à ses longues oreilles, un bruit terrible : PAN !! PAN !! Il a senti quelque chose passer très près de lui ! Il aperçoit enfin le terrier de ses parents, s'y précipite et ne bouge plus... Ouf, le drôle d'animal et l'homme passent en courant à côté du terrier et continuent leur chemin sans*

s'arrêter. Finalement Hector ne se sent plus si grand et décide de rester encore quelque temps dans sa famille !!!

- *Qui est Hector ? Un petit chat, un lapin, un oiseau, une souris.*
- *Quel est cet animal qui court partout en reniflant le sol ? Un renard, un lapin, un chien, un taureau.*
- *Quel est le plat préféré d'Hector ? La soupe, les carottes, les tartines de pain, les pommes*
- *Quel est cet homme qui poursuit Hector ? Un boxeur, un indien, un chasseur, un cow-boy.*
- *Dans : « Hector a très peur et préfère revenir très vite dans le terrier de ses parents : il détale », que veut dire « il détale » ? Il dort, il court, il mange, il est assis à l'affût.*
- *Quel est le long tube que porte l'homme ? Un fusil, un bâton, une épée, un arc avec une flèche*
- *Quelle est l'image qui correspond le plus à l'histoire ? Un lapin mangeant des carottes avec ses parents dans la forêt, un chasseur qui tire sur un oiseau, un chasseur avec son chien en arrière-plan et un lapin tapi dans son terrier, un chien qui course un sanglier.*

Text 3 : *Pour préparer la tarte, déroulez la pâte au fond du moule puis ajoutez de la compote par-dessus. Étalez celle-ci avec le dos d'une cuillère. Disposez ensuite les pommes sur cette préparation. Enfournez et laissez cuire pendant 25 minutes.*

- *avec quoi étale-t-on la compote ? Une fourchette, un couteau, une cuillère, une louche.*
- *Entourez l'image qui correspond à « Disposez ensuite les pommes sur cette préparation ».*
- *Entourez l'image qui correspond à enfourner.*
- *Entourez l'image qui correspond au plat qui sort du four d'après le texte que j'ai lu.*

Text 4 : *La nuit est en train de tomber. Arthur va voir son chien pour lui donner des croquettes. Il lui demande de bien monter la garde pour protéger les moutons du cruel animal qui rôde.*

- *Qui demande de monter la garde ? Une femme, un homme, un mouton, un chien.*
- *Qui doit monter la garde ? Un mouton, un homme, un cheval, un chien.*
- *Qui doit être protégé ? Un troupeau de moutons, une poule, des oies, un homme.*
- *Quel pourrait être l'animal qui rôde ? Un chaton, un loup, un poussin, un cheval.*

English version

Text 1: *This morning, Dad gets Matthew ready for school. He puts on his boots and tells him to keep his coat and hood on during playtime. "He says, "Be careful not to get too wet when you leave the classroom and I ask you not to jump in the puddles so as not to splash your friends.*

- *Where does the story take place? In a house, in a forest, in the street, at the swimming pool.*
- *What is the weather like in this story? There is snow, sun, rain and wind.*
- *Circle what Daddy doesn't allow Mathieu to do. He forbids him to play with marbles, to climb the fence, to jump in the puddles, to splash his friends with his bike.*

Text 2: *Hector has decided that he is now big enough and can make his own way in life. This morning he leaves his family's den, runs across the fields to a forest and goes in search of his favourite food: carrots.*

favourite food: carrots. Suddenly, he hears a barking sound and sees a strange animal on all fours with a collar around its neck coming towards him, sniffing the ground without stopping... Hector hides and sees behind the strange animal a man holding a long, strange tube in his hands. The man shouts to the strange animal

that is still sniffing the ground: "Search, search! Find a trail!..." Hector is very afraid and prefers to return very quickly to the family burrow: he runs away, but hardly has he gone a few metres when he hears, thanks to his long ears, a terrible noise: PAN! PAN!!! He felt something passing very close to him! He finally sees his parents' burrow, rushes into it and doesn't move... Phew, the strange animal and the man run past the burrow and continue their way without stopping. Finally Hector doesn't feel so big anymore and decides to stay with his family for a while!

- *Who is Hector? A little cat, a rabbit, a bird, a mouse.*
- *What is that animal running around sniffing the ground? A fox, a rabbit, a dog, a bull.*
- *What is Hector's favourite food? Soup, carrots, bread, apples*
- *Who is the man chasing Hector? A boxer, an Indian, a hunter, a cowboy.*
- *In: "Hector is very afraid and prefers to return very quickly to his parents' den: he runs away", what does "he runs away" mean? He sleeps, he runs, he eats, he sits in wait.*
- *What is the long tube that the man carries? A gun, a stick, a sword, a bow with an arrow*
- *Which picture most closely matches the story? A rabbit eating carrots with its parents in the forest, a hunter shooting a bird, a hunter with his dog in the background and a rabbit lurking in its burrow, a dog chasing a boar.*

Text 3: *To prepare the tart, roll out the pastry to the bottom of the tin and then add the compote on top. Spread it with the back of a spoon. Place the apples on top of this mixture. Place in the oven and bake for 25 minutes.*

- *What do you use to spread the compote? A fork, a knife, a spoon, a ladle.*
- *Circle the picture that corresponds to "Then arrange the apples on this mixture".*
- *Circle the picture that matches "Put in the oven".*
- *Circle the picture that corresponds to the dish coming out of the oven according to the text I read.*

Text 4: *It's getting dark. Arthur goes to see his dog to give him some food. He asks the dog to keep watch to protect the sheep from the cruel animal that is prowling around.*

- *Who asks to stand guard? A woman, a man, a sheep, a dog.*
- *Who should stand guard? A sheep, a man, a horse, a dog.*
- *Who is to be protected? A flock of sheep, a hen, geese, a man.*
- *What animal might be lurking around? A kitten, a wolf, a chick, a horse.*

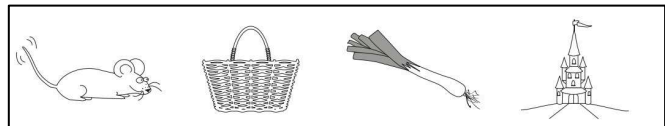
Phoneme handling (T1, T2). Children underwent two types of tests at T1: the first one was a series of 8 spoken words, read aloud by the teacher. For each word, children had to select, among 4 images, another word beginning with the same phoneme as the spoken word. The words were the following “Fille, cheval, valise, poule, tulipe, biscotte, médaille, lapin” (English version : “Girl, horse, suitcase, chicken, tulip, rusk, medal, rabbit”) in 2018, and “Fille, cheval, valise, poule, tulipe, car, biscotte, dent” in 2019, 2020 and 2021. Each image allowed to relieve the child’s working memory and to organize a collective testing (vs. individual testing). The second series was of 7 spoken words. For each word, children had to select, among 4 items, another word that would end with the same phoneme as the spoken word. The words were the following: “Bille, maison, pirate, bateau, verrue, message, petit”.

At T2, a series of 6 different words (“cheval, médaille, lapin, biscotte, tulipe, poulpe”) were presented to children, with the goal of identifying another word, among 4 items, that would begin with the same phoneme than the spoken word. A second series of 6 different words (“maison, pirate, bateau, verrue, message, petit”) were presented to children, with the goal of identifying another word, among 4 items, that would finish with the same phoneme than the spoken word.

The student who frequently chooses a word with an initial consonant that is phonologically similar to the one at the beginning of the target word is likely to experience difficulties in phonemic discrimination, unlike the student who consistently

selects the correct item (which is indicative of both good segmentation and phonemic discrimination abilities).

Syllable handling (T1, T2, T3). Children underwent two tests at T1: the first one was a series of 10 spoken words, read aloud by the teacher. Children had to identify, among 4 images, another word beginning with the same syllable than the spoken word. Each image allowed to relieve the working memory and to organize a collective testing (vs. individual testing).



E.g., “Between - souris, panier, poireau, chateau – which word starts with the same syllable as “cha-peau””.

The following words were “Couleur, binocle, monture, vacances, râteau, tangram, bouton, chamois, chanteur, pirogue” (English version: “Color, binocular, frame, vacation, rake, tangram, button, chamois, singer, dugout”).

The second was 5 series of 4 spoken words with the aim of identifying, among the 4 items, the word which end would contain a different syllable than others. The series were the following: “(1) chateau, hérisson, glaçon, ourson; (2) moto, couteau, bateau, voiture; (3) poussin, taureau, zéro, bureau; (4) saucisson, hérisson, otarie, paillason ; (5) canari, cinéma, écurie, otarie ».

At T2, children were asked to write the following 10 simple and complex syllables, read aloud by the teacher, using a correct phonetical writing: “mu, ti, na, lur, sar, ol, moi, che, tra, pli”.

At T3, children were asked to write the following 12 simple and complex syllables, read aloud by the teacher, using a correct phonetical writing: “vu, moi, che, tra, pli, clou, pal, bol, miam, dual, plaf, vroum”.

Letter-sound association (T1, T2).

Children had to isolate the initial phoneme of a dictated word phoneme and had to circle its corresponding first letter, among 5 possible choices. The assessment consisted of 10 items which were the following: “fil, sol, vol, pile, tard, bulle, dos, mal,

lune, rose” in 2018, 2019 and 2020 except at T1 in 2019, 2020 and 2021 which were the following: “fil, sol, vol, poule, tard, bon, dent, mal, lune, rose”.

Recognizing letters is a new learning process that requires modifying our visual behavior. Unlike a chair, which remains a chair regardless of its orientation, this is not the case for letters (cf. p-q and u-n). This task requires two abilities related to reading. For instance, when asked to identify the letters corresponding to the oral word "poule" (chicken), the options provided were <p>, , <t>, <q>. This exercise demanded proficiency in two key reading-related skills. Firstly, the child needed to distinguish the initial consonant of a syllable from the subsequent vowel, a skill known as phonemic analysis capacity. Secondly, the child had to apply their understanding of the connections between phonemes and graphemes (i.e., the association between a letter and its sound). In the given example, children who failed to select the correct answer (<p>) might (1) exhibit challenges in phonology when they confused phonemes like /p/ and /t/, (2) encounter visual hurdles when they mixed up mirror-image letters like <p> and <q> (Dehaene et al., 2010), or (3) face both issues when they struggled to differentiate between closely related letters such as <p> and (i.e., letters which are closed phonetically and visually).

Letter recognition (T1). Children had to circle, among 18 items, the 3 instances of a spoken letter which was read aloud by the teacher. The targets varied in font and case. The assessment consisted in a series of 7 such items.

Visuo-attentional abilities (T1). To assess letters relative position to each other, children had to identify and circle the similar duo, trio or quatuor of letters, between 24 couples of letters.

One-minute reading aloud.

Reading fluency refers to the number of words read correctly aloud within a specified time frame, typically one minute. The significance of assessments of this nature is manifold. Firstly, they ensure a consistent administration duration for all participants, facilitating comparisons. Additionally, the brief duration helps prevent the onset of

fatigue. Most importantly, fluency serves as an indicator of the level of automation in word identification processes.

In the one-minute word reading task, a list is presented comprising common words, most of which exhibit regular grapheme-phoneme correspondences (e.g., friend, table) and feature simple syllabic structures. Towards the end of the list, a few words with graphemes whose pronunciation depends on context (e.g., the two <g>'s in garage) are included, along with some irregular words (e.g., “sept” in French, pronounced <set>).

Challenges observed during the one-minute word reading task can stem from various sources. Firstly, there may be a decoding issue, which can range from severe (if numerous regular words are read incorrectly or very slowly) to mild (when only words containing context-dependent graphemes are read less accurately). Secondly, a deficit in the lexical procedure may be identified, primarily affecting irregular common words rather than regular ones. It is crucial to assess the mastery level of skills associated with word reading and based on the results, offer specific assistance to children in need.

One-minute word reading (T2, T3). To evaluate word reading fluency, children were asked to read aloud as many words as they could, within one minute. The items in each list were presented in increasing order of difficulty. A maximum of 30 words were presented at T2 (“à, où, la, au, tu, un, il, été, un, mur, ni, sur, qui, (...) avril, roi, faire »), and 60 words at T3 («ta, bol, lune, gare, lire (...) sept, visage, soixante, trésor, lourd, femme, garage, hibou »).

One-minute text reading (T2, T3). To evaluate text reading fluency, children were asked to read aloud, in less than 1 minute, as many words as they could within a text of 29 words (T2) or 102 words (T3).

--

Text at T2 “ Le renard court dans la forêt. Il arrive à la ferme. Va-t-il voler une poule ? Lola l’a vu. Elle crie et le chasse. « Bravo Lola ! », dit la poule.

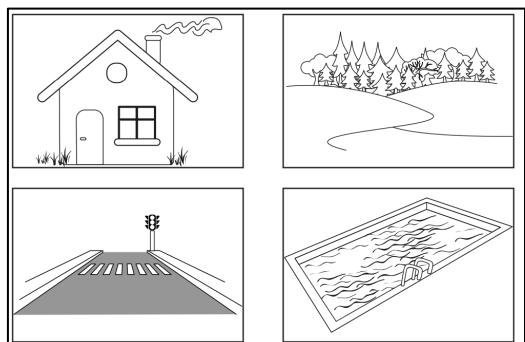
Text at T3 « Madame et Monsieur Petit vivent dans une grande maison entourée d’un jardin avec leur chien, Médor. La porte du jardin reste toujours fermée pour que Médor ne s’échappe pas. Médor aime se coucher en regardant le ciel. Au début de l’hiver, Madame et monsieur Petit décident de le mettre dans la cuisine, bien au chaud. Comme il préfère s’endormir en regardant les étoiles, Médor aboie très fort et très longtemps au début de la nuit. Madame et Monsieur Petit n’arrivent plus à dormir. Au bout d’une semaine, ils décident remettre Médor dans le jardin, mais avec une niche et une couverture. ».

Writing words to dictation (T2, T3). To assess writing abilities, children were asked to write 8 dictated simple and regular words at T2 (“moto, midi, uni, samedi, tour, lavabo, mardi, riche”), and 12 regular words at T3 (“libre, mardi, barbe, riche, toile, jeudi, avril, larme, tarte, tache, poudre, lundi”).

Reading comprehension of sentences (T3).

Children had to circle, among 4 pictures, the one corresponding to the sentence they read. 10 sentences of increasing length and complexity were presented at T3.

Reading comprehension of texts (T3). The test consisted of reading 2 texts of increasing length and complexity and then answer to questions about these texts. For each text, children had to read the text by themselves, and then, on one hand, the teacher would state 4 questions about the text orally where children would have to circle, among 4 pictures, the answer corresponding to the text they had read; on the other hand, children would have to read the other 4



questions by themselves and find the correct answer. Questions asked orally by the teacher were the following:

French version : "(1) Ce texte est : un documentaire, une recette, un menu, un album ; (2) Ce texte permet de préparer : une compote de pommes, des crêpes, une tarte aux pommes, une tarte aux poires ; (3) Que doit-on étaler ? De la compote, des pommes, du sucre, des poires ; (4) Comment fait-on cuire le plat ? Dans une poêle, dans une casserole, dans un four, au barbecue ».

English version : "(1) This text is: a documentary, a recipe, a menu, a scrapbook; (2) This text makes: applesauce, pancakes, apple pie, pear pie; (3) What should be spread? Applesauce, apples, sugar, pears; (4) How is the dish cooked? In a frying pan, in a saucepan, in the oven, on the barbecue".

Note that in future multilevel models presented on chapter 3, we qualified "letter recognition" and "letter knowledge" as "**decoding skills**" and "phoneme handling", "syllable handling" and "letter-sound association" as "**meta phonology**".

Table S1. Description of children' characteristics and tests in 2018, 2019, 2020 and 2021.

	2018	2019	2020	2021
n	586,936	686,138	717,326	749,402
	Mean (sd)	Mean (sd)	Mean (sd)	Mean (sd)
Age in first grade (month)	74.64 (3.84)	74.62 (3.85)	74.67 (3.86)	74.71 (3.95)
Class size	17.21 (5.85)	17.00 (5.76)	16.91 (5.71)	17.42 (5.84)
SES score	102.36 (17.79)	102.57 (17.84)	102.61 (17.88)	103.33 (18.46)
Gender – Boys, n (%)	298,633 (50.9)	348,933 (50.9)	363,594 (50.7)	382,858 (51.1)
School categories, private, n (%)	63,304 (10.8)	80163 (11.7)	87379 (12.2)	90011 (12.0)
School categories, regular public, n (%)	426,643 (72.7)	488039 (71.1)	505306 (70.4)	526207 (70.2)
School categories, PE, n (%)	58,413 (10.0)	72393 (10.6)	76517 (10.7)	80674 (10.8)
School categories, HPE, n (%)	38,576 (6.6)	45543 (6.6)	48124 (6.7)	52510 (7.0)
<i>Language mean</i>				
T1 Oral Comprehension of Words, range 0-15	11.89 (2.71)	11.63 (2.89)	11.52 (2.99)	11.57 (3.08)
T1 Oral Comprehension of Sentences, 0-14	12.13 (2.17)	12.27 (2.08)	12.16 (2.21)	12.10 (2.31)
T1 Oral Comprehension of Texts, 0-18	13.27 (3.64)	8.24 (2.26)	8.21 (2.35)	8.21 (2.39)
T1 Phoneme handling, 0-15	8.90 (3.77)	9.17 (3.47)	9.01 (3.55)	9.33 (3.65)
T1 Syllable handling, 0-15	11.49 (3.11)	11.74 (3.35)	11.48 (3.53)	11.59 (3.53)
T1 Letter-sound association, 0-10	7.44 (2.55)	7.79 (2.50)	7.59 (2.63)	7.71 (2.66)
T1 Decoding, letter writings recognition, 0-7	4.77 (1.84)	4.68 (1.83)	4.57 (1.90)	4.82 (1.86)
T1 Comparing letters, visuo-attentional abilities, 0-24	15.23 (6.69)	10.06 (2.72)	9.93 (2.82)	9.91 (2.83)
T2 Oral Comprehension of Sentences, 0-14	12.22 (1.89)	12.22 (1.91)	12.18 (1.95)	12.23 (2.00)

T2 Reading a list of 30 words in 1 minute, 0-100	24.97 (17.36)	25.61 (17.57)	26.57 (18.06)	30.02 (21.21)
T2 Reading 29 words of a text in 1 minute, 0-195	29.83 (25.74)	29.47 (24.86)	30.70 (25.70)	35.85 (30.34)
T2 Reading a list of 30 words in 1 minute, 0-30	19.59 (7.54)	19.93 (7.42)	20.39 (7.33)	15.87 (8.80)
T2 Reading 29 words of a text in 1 minute, 0-29	19.17 (8.07)	19.78 (8.10)	20.42 (8.02)	11.10 (9.65)
T2 Writing Syllables, 0-10	7.95 (2.37)	8.08 (2.35)	8.16 (2.33)	7.98 (2.92)
T2 Writing Words, 0-8	5.91 (2.22)	5.98 (2.25)	6.09 (2.22)	6.26 (2.19)
T2 Phoneme handling, 0-12	9.31 (2.76)	9.45 (2.54)	9.53 (2.52)	9.31 (2.99)
T2 Decoding, Letter recognition, 0-10	9.39 (1.28)	9.50 (1.18)	9.51 (1.18)	9.07 (2.29)
T3 Oral Comprehension of Words, 0-15	13.41 (2.01)	13.31 (2.14)	13.39 (2.08)	13.31 (2.16)
T3 Oral Comprehension of Sentences, 0-15	13.79 (1.65)	13.77 (1.72)	13.77 (1.70)	13.69 (1.81)
T3 Writing Syllables, 0-12	9.65 (2.87)	9.34 (3.15)	9.74 (2.85)	9.71 (2.88)
T3 Writing Words, 0-12	8.62 (3.11)	8.19 (3.35)	8.61 (3.14)	8.39 (3.19)
T3 Understanding reading a sentence, 0-10	7.78 (2.45)	7.61 (2.64)	7.85 (2.41)	7.91 (2.45)
T3 Understanding reading a text, 0-8	5.97 (2.03)	5.79 (2.15)	5.98 (2.05)	5.92 (2.08)
T3 Reading a list of 60 words in 1 min, 0-93	43.49 (19.06)	41.97 (20.06)	44.89 (19.26)	44.68 (19.85)
T3 Reading 102 words of a text in 1 min, 0-136	50.98 (31.21)	48.79 (32.81)	52.68 (31.92)	45.60 (31.75)
T3 Reading a list of 60 words in 1 min, 0-60	40.49 (15.16)	39.12 (16.26)	41.69 (15.14)	40.12 (14.87)
T3 Reading 102 words of a text in 1 min, 0-102	49.04 (27.94)	46.77 (29.44)	50.49 (28.36)	49.54 (28.79)
Math mean				
T1 Writing numbers, 0-11	10.28 (1.65)	10.52 (1.36)	10.48 (1.43)	10.57 (1.32)
T1 Reading numbers, 0-10	9.70 (0.97)	9.69 (1.01)	9.64 (1.13)	9.70 (1.04)
T1 Problem solving, 0-6	3.80 (1.81)	4.08 (1.68)	4.00 (1.72)	4.08 (1.70)
T1 Enumerate quantities, 0-8	7.49 (1.15)	7.40 (1.31)	7.35 (1.40)	7.42 (1.32)
T1 Associate number to quantity, 0-60	22.99 (14.34)	21.99 (11.29)	21.90 (11.56)	22.35 (11.52)

T1 Number line, 0-6	3.06 (1.84)	3.17 (1.87)	3.15 (1.90)	3.11 (1.97)
T2 Comparing numbers, 0-40	36.58 (7.41)	31.20 (9.80)	31.23 (9.78)	31.23 (9.80)
T2 Number line, 0-10	5.42 (2.47)	5.69 (2.54)	5.77 (2.55)	6.05 (2.63)
T2 Additioning, 0-7	5.74 (1.68)	7.89 (2.53)	7.99 (2.49)	8.30 (2.39)
T2 Subtrationing, 0-7	5.09 (2.46)	6.94 (3.31)	7.10 (3.24)	7.04 (3.41)
T2 Writing numbers, 0-10	8.97 (1.87)	9.12 (1.76)	9.19 (1.70)	9.29 (1.68)
T2 Problem solving, 0-5	3.43 (1.41)	3.43 (1.40)	3.49 (1.39)	3.46 (1.52)
T3 Geometry, 0-8	5.80 (1.64)	5.97 (1.64)	5.98 (1.61)	5.98 (1.64)
T3 Number line, 0-15	7.07 (3.63)	7.01 (3.71)	7.22 (3.68)	9.66 (4.70)
T3 Additioning, 0-7	4.39 (2.20)	5.55 (2.39)	5.76 (2.30)	5.70 (2.28)
T3 Subtrationing, 0-8	3.78 (2.64)	3.64 (2.41)	3.91 (2.37)	3.76 (2.44)
T3 Mental calculus, 0-10	8.43 (2.04)	8.42 (2.08)	8.48 (2.02)	8.50 (2.00)
T3 Writing numbers, 0-10	8.29 (2.44)	8.21 (2.55)	8.47 (2.33)	8.23 (2.54)
T3 Reading numbers, 0-10	8.52 (2.03)	8.45 (2.16)	8.64 (1.96)	8.60 (2.00)
T3 Associate number to quantity, 0-16	9.69 (3.93)	9.54 (4.05)	9.75 (4.00)	-
T3 Problem solving, 0-6	4.08 (1.64)	4.06 (1.67)	4.15 (1.63)	4.11 (1.64)
Composite variables				
Language at T1	72.70 (15.68)	76.02 (15.39)	74.86 (16.39)	75.78 (16.74)
Language at T2	64.64 (13.89)	65.34 (13.65)	65.97 (13.74)	66.05 (16.24)
Language at T3	71.29 (16.08)	69.53 (17.55)	71.78 (16.28)	70.71 (15.99)
Math at T1	72.77 (13.45)	76.83 (14.02)	76.24 (14.79)	76.89 (14.34)
Math at T2	76.45 (17.78)	73.85 (18.84)	74.71 (18.66)	75.70 (18.64)
Math at T3	67.84 (17.81)	70.14 (18.79)	71.99 (17.82)	73.23 (18.45)

Table S2. Gender and age in function of the four school categories in 2019, 2020 and 2021.

Year	Variable	Category	School categories at T1			
			Private schools	Regular public schools	PE schools	HPE schools
2019	School categories, % (n)		11.7 (80163)	71.1 (488039)	10.6 (72393)	6.6 (45543)
	Gender					
		Boys, n (%)	41208 (51.4)	247890 (50.8)	36694 (50.7)	23141 (50.8)
	Age					
		Advanced, n (%)	861 (1.1)	2727 (0.6)	350 (0.5)	246 (0.5)
		Boys, n (%)	407 (47.3)	1241 (45.5)	147 (42.0)	109 (44.3)
		Late, n (%)	1697 (2.1)	11258 (2.3)	1909 (2.6)	1458 (3.2)
		Boys, n (%)	1057 (62.3)	6929 (61.5)	1179 (61.8)	886 (60.8)
2020	School categories, % (n)		12.2 (87379)	70.4 (505306)	10.7 (76517)	6.7 (48124)
	Gender					
		Boys, n (%)	44588 (51.0)	256140 (50.7)	38518 (50.3)	24348 (50.6)
	Age					
		Advanced, n (%)	963 (1.1)	2477 (0.5)	286 (0.4)	231 (0.5)
		Boys, n (%)	410 (42.6)	1089 (44.0)	132 (46.2)	100 (43.3)
		Late, n (%)	2103 (2.4)	12121 (2.4)	2125 (2.8)	1571 (3.3)
		Boys, n (%)	1249 (59.4)	7390 (61.0)	1314 (61.8)	950 (60.5)
2021	School categories, % (n)		12.0 (90011)	70.2 (526207)	10.8 (80674)	7.0 (52510)
	Gender					
		Boys, n (%)	46363 (51.5)	268664 (51.1)	41126 (51.0)	26705 (50.9)
	Age					
		Advanced, n (%)	1024 (1.1)	2792 (0.5)	313 (0.4)	197 (0.4)
		Boys, n (%)	474 (46.3)	1247 (44.7)	130 (41.5)	78 (39.6)
		Late, n (%)	2513 (2.8)	15350 (2.9)	2855 (3.5)	2128 (4.1)
		Boys, n (%)	1500 (59.7)	9262 (60.3)	1730 (60.6)	1338 (62.9)

Table S3. Among the total population entering in first grade in 2018, description of their results considering their age category (Advance; typical; late) (n = 586,936).

Variables	Advanced (Age < 69 month-old)	Typical (Age 69- to 80- month-old)	Late (Age > 80 month-old)	p
n	3519	569755	13662	-
Age at T1 (mean (SE))	66.15 (2.44)	74.44 (3.42)	85.20 (3.53)	< 0.0001
Class size (mean (SE))	17.45 (5.95)	17.22 (5.86)	16.68 (5.63)	< 0.0001
SES score (mean (SE))	105.55 (17.79)	102.48 (17.78)	96.77 (17.33)	< 0.0001
Gender - Boys, n (%)	1580 (44.9)	288574 (50.6)	8479 (62.1)	< 0.0001
Language mean				-
T1 Oral Comprehension of Words, 0-15 (mean (SE))	85.21 (14.97)	79.53 (17.91)	66.67 (20.99)	< 0.0001
T1 Oral Comprehension of Sentences, 0-14 (mean (SE))	91.03 (11.95)	86.92 (15.27)	75.30 (20.43)	< 0.0001
T1 Oral Comprehension of Texts, 0-18 (mean (SE))	80.13 (16.88)	74.02 (20.04)	59.18 (23.24)	< 0.0001
T1 Phoneme handling, 0-15 (mean (SE))	74.11 (22.53)	59.60 (25.04)	42.88 (22.66)	< 0.0001
T1 Syllable handling, 0-15 (mean (SE))	85.87 (15.66)	76.98 (20.54)	58.93 (22.77)	< 0.0001
T1 Letter-sound association, 0-10 (mean (SE))	87.82 (18.09)	74.55 (25.47)	66.20 (27.55)	< 0.0001
T1 Decoding, letter writings recognition, 0-7 (mean (SE))	79.64 (20.84)	68.36 (26.16)	56.88 (28.42)	< 0.0001
T1 Comparing letters, visuo-attentional abilities, 0-24 (mean (SE))	70.16 (25.82)	63.59 (27.85)	56.04 (28.72)	< 0.0001
T2 Oral Comprehension of Sentences, 0-14 (mean (SE))	90.81 (10.78)	87.55 (13.28)	76.75 (18.00)	< 0.0001
T2 Reading a list of 30 words in 1 minute (mean (SE))	39.81 (24.12)	25.10 (17.32)	15.77 (12.15)	< 0.0001
T2 Reading 29 words of a text in 1 minute (mean (SE))	26.61 (19.95)	15.39 (13.17)	8.77 (8.57)	< 0.0001
T2 Writing Syllables, 0-10 (mean (SE))	90.47 (14.85)	79.92 (23.34)	58.73 (30.37)	< 0.0001
T2 Writing Words, 0-8 (mean (SE))	85.57 (19.06)	74.40 (27.39)	49.42 (33.88)	< 0.0001
T2 Phoneme handling, 0-12 (mean (SE))	88.13 (17.29)	77.98 (22.69)	56.92 (25.28)	< 0.0001

T2 Decoding, Letter recognition, 0-10 (mean (SE))	97.25 (7.98)	94.11 (12.53)	85.57 (20.18)	< 0.0001
T3 Oral Comprehension of Words, 0-15 (mean (SE))	93.30 (10.58)	89.62 (13.20)	78.48 (18.17)	< 0.0001
T3 Oral Comprehension of Sentences, 0-15 (mean (SE))	94.73 (8.61)	92.17 (10.72)	82.24 (16.42)	< 0.0001
T3 Writing Syllables, 0-12 (mean (SE))	91.20 (14.37)	81.05 (23.31)	52.54 (32.08)	< 0.0001
T3 Writing Words, 0-12 (mean (SE))	84.00 (18.42)	72.43 (25.47)	44.98 (31.26)	< 0.0001
T3 Understanding reading a sentence, 0-10 (mean (SE))	88.27 (16.31)	78.35 (24.07)	53.60 (30.04)	< 0.0001
T3 Understanding reading a text, 0-8 (mean (SE))	86.37 (18.09)	75.15 (25.04)	49.08 (27.64)	< 0.0001
T3 Reading a list of 60 words in 1 min (mean (SE))	59.88 (20.01)	47.13 (20.31)	28.13 (18.75)	< 0.0001
T3 Reading 102 words of a text in 1 min (mean (SE))	54.43 (24.24)	37.84 (22.81)	18.18 (17.59)	< 0.0001
Math mean				-
T1 Writing numbers, 0-11 (mean (SE))	95.79 (11.89)	93.53 (14.76)	88.12 (21.47)	< 0.0001
T1 Reading numbers, 0-10 (mean (SE))	98.40 (6.61)	97.04 (9.50)	93.57 (15.00)	< 0.0001
T1 Problem solving, 0-6 (mean (SE))	74.86 (25.94)	63.67 (30.01)	46.83 (30.79)	< 0.0001
T1 Enumerate quantities, 0-8 (mean (SE))	95.72 (11.18)	93.69 (14.23)	88.49 (20.39)	< 0.0001
T1 Associate number to quantity, 0-60 (mean (SE))	44.80 (23.67)	38.44 (23.88)	30.95 (23.61)	< 0.0001
T1 Number line, 0-6 (mean (SE))	60.08 (29.20)	51.25 (30.60)	38.75 (28.68)	< 0.0001
T2 Comparing numbers, 0-40 (mean (SE))	95.73 (12.20)	89.86 (19.92)	78.04 (27.38)	< 0.0001
T2 Number line, 0-10 (mean (SE))	63.20 (23.16)	54.45 (24.66)	40.92 (23.97)	< 0.0001
T2 Additioning, 0-7 (mean (SE))	90.78 (16.64)	82.33 (23.66)	65.15 (31.18)	< 0.0001
T2 Substrating, 0-7 (mean (SE))	85.80 (26.41)	73.15 (34.95)	51.50 (39.22)	< 0.0001
T2 Writing numbers, 0-10 (mean (SE))	95.44 (12.30)	89.96 (18.39)	78.53 (25.87)	< 0.0001
T2 Problem solving, 0-5 (mean (SE))	79.31 (23.07)	69.01 (27.98)	49.27 (29.95)	< 0.0001
T3 Geometry, 0-8 (mean (SE))	77.33 (18.07)	72.80 (20.36)	61.24 (23.86)	< 0.0001
T3 Number line, 0-15 (mean (SE))	58.46 (23.83)	47.39 (24.16)	32.41 (20.99)	< 0.0001
T3 Additioning, 0-7 (mean (SE))	76.83 (27.01)	63.15 (31.33)	39.76 (30.31)	< 0.0001
T3 Substrating, 0-8 (mean (SE))	62.67 (31.59)	47.61 (32.96)	27.48 (27.88)	< 0.0001
T3 Mental calculus, 0-10 (mean (SE))	91.16 (14.18)	84.68 (19.99)	67.55 (28.97)	< 0.0001

T3 Writing numbers, 0-10 (mean (SE))	92.51 (16.02)	83.47 (23.95)	58.00 (31.60)	< 0.0001
T3 Reading numbers, 0-10 (mean (SE))	93.19 (13.14)	85.68 (19.90)	64.46 (27.67)	< 0.0001
T3 Associate number to quantity, 0-16 (mean (SE))	69.77 (22.35)	60.87 (24.45)	45.55 (25.58)	< 0.0001
T3 Problem solving, 0-6 (mean (SE))	81.38 (21.43)	68.44 (26.99)	43.23 (27.79)	< 0.0001
Composite variables				-
Language at T1 (mean (SE))	81.75 (11.70)	72.95 (15.55)	60.26 (16.23)	< 0.0001
Language at T2 (mean (SE))	74.09 (11.07)	64.92 (13.66)	50.28 (15.63)	< 0.0001
Language at T3 (mean (SE))	81.52 (11.58)	71.72 (15.72)	50.91 (17.57)	< 0.0001
Math at T1 (mean (SE))	78.28 (10.72)	72.94 (13.36)	64.45 (15.02)	< 0.0001
Math at T2 (mean (SE))	85.03 (12.41)	76.77 (17.54)	60.57 (20.93)	< 0.0001
Math at T3 (mean (SE))	78.14 (13.79)	68.23 (17.55)	48.86 (18.71)	< 0.0001

Table S4. Analyzing the effect of age in function of time on math and language results. Analyses were made among typical-in-age children. Time was defined as T1 = 0; T2 = 4; T3 = 12 months of school.

Fixed effects for	Math		Language	
	Math ~ Age + Time + Age*Time	p	Language ~ Age + Time + Age*Time	p
Intercept	-0.1276 (0.0086)	< 0.0001	-0.1144 (0.0086)	< 0.0001
Time	0.0175 (0.0006)	< 0.0001	0.0310 (0.0006)	< 0.0001
Age	0.0525 (0.0003)	< 0.0001	0.0472 (0.0004)	< 0.0001
Age * Time	-0.0007 (0.0000)	< 0.0001	0.0012 (0.0000)	< 0.0001

As we wanted to test if gaining more months of age (i.e., from 69 to 80 months) or being exposed to education (i.e., time = 0, 4 to 12 months) affected the results in both math and language at T1, T2 and T3, among the typical-age subgroup (i.e., 69 to 80 month-old), we implemented the following model : Math ~ Age + Time + Age*Time. Age linearly diminished with time, as starting from Age at T0 = 0.0525 (0.0003) ***, the impact on math diminished of -0.0007 (0.0000) *** for every additional month of age. The age effect was less important at the beginning of second grade (0.0525 – 0.0007 * 12 = 0.0441) compared to the beginning of first grade (0.0525). Whereas for

language, the age effect at T0 (0.0472) rose up with time ($0.0472 + 12 \times 0.0012 = 0.0616$) and was more important in second grade compared to first grade.

In addition, we examined the ANOVA p-value from the models' interactions of each variable by age group, and then, we compared their slopes: Results indicated that the typical age group of children presented a linear effect slope significantly different from both advanced and late in age children' slopes (see **Table S5**).

Table S5. Comparison of linear effect slopes between age categories and math and language at T1, T2 and T3.

Variables	Contrasts	Estimates (SE)	P value
T1 Math	Late - Typical	-0.0362 (0.0019)	< 0.0001
	Late - Advanced	-0.0132 (0.0057)	0.0523
	Typical - Advanced	0.0230 (0.0053)	< 0.0001
T2 Math	Late - Typical	-0.0256 (0.0014)	< 0.0001
	Late - Advanced	-0.0095 (0.0049)	0.1233
	Typical - Advanced	0.0160 (0.0047)	0.0017
T3 Math	Late - Typical	-0.0256 (0.0016)	< 0.0001
	Late - Advanced	-0.0064 (0.0044)	0.3221
	Typical - Advanced	0.0192 (0.0042)	< 0.0001
T1 Language	Late - Typical	-0.0293 (0.0018)	< 0.0001
	Late - Advanced	-0.0100 (0.0052)	0.1328
	Typical - Advanced	0.0193 (0.0049)	0.0002
T2 Language	Late - Typical	-0.0270 (0.0019)	< 0.0001
	Late - Advanced	0.0032 (0.0055)	0.8273
	Typical - Advanced	0.0302 (0.0052)	< 0.0001
T3 Language	Late - Typical	-0.0249 (0.0017)	< 0.0001
	Late - Advanced	-0.0033 (0.0052)	0.7999
	Typical - Advanced	0.0216 (0.0050)	< 0.0001

Table S6. Description of cognitive tests in function of school social category (Private - Public - PE - HPE) in 2018 (n = 586,936)

	Private	Regular public	PE public	HPE public	p
N students	63304	426643	58413	38576	-
Age at T1 (mean (SE))	74.61 (3.82)	74.61 (3.81)	74.79 (4.04)	74.77 (3.98)	< 0.0001
Class size (mean (SE))	19.04 (6.23)	18.14 (5.63)	11.92 (2.85)	11.87 (2.99)	< 0.0001
SES score (mean (SE))	115.06 (14.55)	106.00 (14.20)	82.05 (11.16)	72.00 (9.48)	< 0.0001
Gender - Boys, n (%)	32318 (51.1)	217196 (50.9)	29782 (51.0)	19337 (50.1)	< 0.0001
<i>Language mean</i>					-
T1 Oral Comprehension of Words, 0-15 (mean (SE))	84.08 (14.73)	80.99 (16.96)	70.91 (20.10)	65.03 (21.14)	< 0.0001
T1 Oral Comprehension of Sentences, 0-14 (mean (SE))	89.62 (12.60)	87.77 (14.46)	81.67 (18.13)	77.29 (20.78)	< 0.0001
T1 Oral Comprehension of Texts, 0-18 (mean (SE))	78.55 (17.49)	75.30 (19.36)	65.47 (21.86)	60.69 (22.76)	< 0.0001
T1 Phoneme handling, 0-15 (mean (SE))	64.62 (23.61)	60.39 (24.90)	52.28 (25.22)	49.18 (25.14)	< 0.0001
T1 Syllable handling, 0-15 (mean (SE))	80.21 (18.34)	77.83 (20.16)	70.42 (22.64)	66.69 (23.27)	< 0.0001
T1 Letter-sound association, 0-10 (mean (SE))	79.24 (22.79)	75.10 (25.12)	69.27 (27.47)	66.99 (28.45)	< 0.0001
T1 Decoding, letter writings recognition, 0-7 (mean (SE))	71.26 (24.03)	69.26 (25.61)	62.80 (28.39)	59.10 (30.14)	< 0.0001
T1 Comparing letters, visuo-attentional abilities, 0-24 (mean (SE))	66.39 (27.05)	64.59 (27.62)	57.32 (28.23)	55.33 (28.93)	< 0.0001
T2 Oral Comprehension of Sentences, 0-14 (mean (SE))	89.68 (11.57)	88.33 (12.71)	82.93 (15.25)	78.92 (17.10)	< 0.0001
T2 Reading a list of 30 words in 1 minute (mean (SE))	27.47 (17.78)	25.05 (17.34)	23.87 (17.14)	21.74 (16.50)	< 0.0001
T2 Reading 29 words of a text in 1 minute (mean (SE))	17.42 (13.77)	15.39 (13.30)	14.19 (12.38)	12.49 (11.56)	< 0.0001
T2 Writing Syllables, 0-10 (mean (SE))	83.46 (19.54)	79.56 (23.52)	78.12 (25.49)	74.34 (27.76)	< 0.0001
T2 Writing Words, 0-8 (mean (SE))	78.35 (23.74)	74.06 (27.53)	71.83 (29.79)	67.79 (32.01)	< 0.0001
T2 Phoneme handling, 0-12 (mean (SE))	80.82 (20.55)	78.05 (22.77)	74.47 (24.13)	71.25 (25.20)	< 0.0001
T2 Decoding, Letter recognition, 0-10 (mean (SE))	95.51 (10.25)	94.04 (12.51)	93.16 (14.15)	91.28 (16.59)	< 0.0001

T3 Oral Comprehension of Words, 0-15 (mean (SE))	92.60 (9.98)	90.65 (12.10)	83.65 (16.40)	78.68 (18.87)	< 0.0001
T3 Oral Comprehension of Sentences, 0-15 (mean (SE))	94.08 (8.60)	92.81 (9.98)	88.19 (13.42)	84.68 (15.78)	< 0.0001
T3 Writing Syllables, 0-12 (mean (SE))	85.08 (19.07)	80.78 (23.55)	77.63 (26.19)	73.56 (28.90)	< 0.0001
T3 Writing Words, 0-12 (mean (SE))	76.04 (22.32)	72.32 (25.56)	68.66 (28.04)	64.84 (29.97)	< 0.0001
T3 Understanding reading a sentence, 0-10 (mean (SE))	83.58 (19.38)	78.89 (23.84)	71.47 (27.16)	66.37 (29.06)	< 0.0001
T3 Understanding reading a text, 0-8 (mean (SE))	81.27 (21.19)	75.85 (24.76)	67.24 (27.35)	61.08 (28.38)	< 0.0001
T3 Reading a list of 60 words in 1 min (mean (SE))	50.93 (19.08)	47.11 (20.38)	44.00 (20.94)	40.21 (21.35)	< 0.0001
T3 Reading 102 words of a text in 1 min (mean (SE))	42.99 (22.57)	37.95 (22.98)	33.43 (22.02)	29.44 (21.43)	< 0.0001
Math mean					
T1 Writing numbers, 0-11 (mean (SE))	94.80 (12.70)	93.87 (14.26)	91.37 (17.42)	89.36 (20.06)	< 0.0001
T1 Reading numbers, 0-10 (mean (SE))	97.68 (7.83)	97.21 (9.02)	95.96 (11.66)	94.58 (14.39)	< 0.0001
T1 Problem solving, 0-6 (mean (SE))	69.51 (27.62)	65.30 (29.47)	53.08 (31.02)	47.15 (31.11)	< 0.0001
T1 Enumerate quantities, 0-8 (mean (SE))	94.87 (11.87)	94.04 (13.59)	91.53 (17.30)	89.53 (20.19)	< 0.0001
T1 Associate number to quantity, 0-60 (mean (SE))	41.38 (23.34)	39.20 (23.96)	32.99 (22.70)	31.47 (23.55)	< 0.0001
T1 Number line, 0-6 (mean (SE))	55.59 (29.82)	52.39 (30.56)	43.60 (29.77)	39.51 (29.36)	< 0.0001
T2 Comparing numbers, 0-40 (mean (SE))	93.94 (15.02)	92.01 (17.90)	88.61 (20.95)	85.44 (24.00)	< 0.0001
T2 Number line, 0-10 (mean (SE))	56.27 (23.56)	55.01 (24.71)	50.30 (24.82)	47.49 (25.01)	< 0.0001
T2 Additioning, 0-7 (mean (SE))	84.61 (21.43)	82.70 (23.34)	78.31 (26.62)	75.23 (28.43)	< 0.0001
T2 Substrationing, 0-7 (mean (SE))	75.88 (33.32)	74.22 (34.43)	66.42 (37.38)	60.46 (39.02)	< 0.0001
T2 Writing numbers, 0-10 (mean (SE))	90.89 (16.92)	90.04 (18.29)	88.48 (20.05)	86.17 (22.33)	< 0.0001
T2 Problem solving, 0-5 (mean (SE))	72.05 (26.28)	69.82 (27.76)	63.16 (29.52)	57.96 (30.05)	< 0.0001
T3 Geometry, 0-8 (mean (SE))	75.03 (19.33)	73.27 (20.20)	68.72 (21.49)	66.46 (22.33)	< 0.0001
T3 Number line, 0-15 (mean (SE))	51.07 (23.54)	48.35 (24.16)	40.58 (23.35)	36.78 (22.64)	< 0.0001
T3 Additioning, 0-7 (mean (SE))	66.54 (30.10)	64.32 (31.08)	57.73 (31.98)	52.80 (32.05)	< 0.0001

T3 Subtrationing, 0-8 (mean (SE))	50.46 (32.59)	48.73 (32.95)	42.83 (32.62)	37.74 (31.73)	< 0.0001
T3 Mental calculus, 0-10 (mean (SE))	86.63 (17.97)	84.86 (19.86)	81.48 (22.65)	78.81 (24.49)	< 0.0001
T3 Writing numbers, 0-10 (mean (SE))	86.34 (21.29)	83.47 (24.02)	79.84 (26.40)	76.08 (28.59)	< 0.0001
T3 Reading numbers, 0-10 (mean (SE))	88.11 (17.52)	85.71 (19.92)	82.64 (22.22)	79.20 (24.38)	< 0.0001
T3 Associate number to quantity, 0-16 (mean (SE))	63.88 (23.24)	61.52 (24.33)	55.31 (25.23)	52.47 (25.72)	< 0.0001
T3 Problem solving, 0-6 (mean (SE))	72.49 (25.12)	69.42 (26.74)	60.66 (28.51)	55.02 (28.84)	< 0.0001
Composite variables					
Language at T1 (mean (SE))	76.74 (13.12)	73.90 (14.98)	66.27 (17.04)	62.54 (18.09)	< 0.0001
Language at T2 (mean (SE))	67.53 (11.68)	64.93 (13.67)	62.65 (14.95)	59.69 (16.20)	< 0.0001
Language at T3 (mean (SE))	75.82 (12.66)	72.04 (15.56)	66.79 (17.69)	62.36 (19.23)	< 0.0001
Math at T1 (mean (SE))	75.64 (11.62)	73.67 (12.98)	68.09 (14.40)	65.27 (15.53)	< 0.0001
Math at T2 (mean (SE))	78.94 (15.49)	77.30 (17.29)	72.55 (19.51)	68.79 (20.88)	< 0.0001
Math at T3 (mean (SE))	71.09 (15.83)	68.76 (17.51)	63.18 (18.63)	59.41 (19.24)	< 0.0001

Table S7. Analyzing the effect of school categories in function of time on math and language results. Analyses were made among typical-in-age children. Time was defined as T1 = 0; T2 = 4; T3 = 12 months of school.

As we wanted to test if going to a specific subgroup of school (i.e., private, regular public, PE or HPE) or being exposed to education (i.e., time = 0, 4 to 12 months) affected the results in both math and language at T1, T2 and T3, among the typical-age subgroup (i.e., 69- to 80-month-old), we implemented the following models:

	Math Math ~ Categories + Time + Categories*Time	p	Language Language ~ Categories + Time + Categories*Time	p
Fixed effects				
<i>Intercept</i>	0,1783 (0,0065)	< 0.0001	0,2197 (0,0069)	< 0.0001
Categories – Regular Public	-0,0008 (0,0003)	0,0024	0,0037 (0,0003)	< 0.0001
Categories – PE	-0,0896 (0,0070)	< 0.0001	-0,1322 (0,0074)	< 0.0001
Categories – HPE	-0,4596 (0,0088)	< 0.0001	-0,5017 (0,0094)	< 0.0001
Time	-0,6482 (0,0098)	< 0.0001	-0,7026 (0,0104)	< 0.0001
Categories – Regular Public * Time	0,0006 (0,0003)	0,0481	-0,0046 (0,0003)	< 0.0001
Categories – PE * Time	0,0069 (0,0004)	< 0.0001	0,0037 (0,0004)	< 0.0001
Categories – HPE * Time	0,0055 (0,0004)	< 0.0001	-0,0000 (0,0004)	0.9621

The more the SES score diminished (and school type lowered in the SES range), the worst results there were. School categories were classed as followed from the worst effect of school categories on math and language over time: HPE > PE > regular public > private schools. PE schools had an effect on math 5x lower compared to regular public, and HPE effect was 8x lower on the level in math. Overall, the global effect of time on math or language is negative, as their mean from T1 to T2 and to T3 are decreasing – these results should be carefully interpreted as none of the tests in math or in language were identical between T1 to T3, even with normalized and gaussianized variables. From T1 to T3 (i.e., time), regular public (-0.0008 + 0.0006*12 = 0.0064) PE (-0.0896 + 0.0069*12 = -0.0068) and HPE (-0.4596 + 0.0055*12 = -

0.3936) tended to progress more compared to private and regular public schools (i.e., the interactions were positive and improve from T1 to T3). HPE schools were not affected by the school category on their results in language (i.e., results non-significant). However, there is a bias here, as mentioned earlier, several scores were saturated and children going to PE and HPE schools tended to have a biased opportunity to progress more (they were farther in level compared to the two other school categories). Overall, children performed better in private and regular public schools, and globally children tended to worsen they level in math from T1 to T3 in PE and HPE schools.

Table S8. Analysis zooming on T1-T2 and on T2-T3 in math and language, and the school categories effects on the results.

	Math				Language			
	T1 to T2 Math ~ Categories + Time + Categories*Time	p	T2 to T3 Math ~ Categories + Time + Categories*Time	p	T1 to T2 Language ~ Categories + Time + Categories*Time	p	T2 to T3 Language ~ Categories + Time + Categories*Time	p
Fixed effects								
<i>Intercept</i>	0.2128 (0.0070)	0.0000	0.0971 (0.0073)	0.0000	0.2485 (0.0075)	0.0000	0.1531 (0.0075)	0.0000
Categories – Regular Public	-0.0205 (0.0009)	0.0000	0.0070 (0.0004)	0.0000	-0.0120 (0.0008)	0.0000	0.0099 (0.0003)	0.0000
Categories – PE	-0.1163 (0.0075)	0.0000	-0.0233 (0.0079)	0.0030	-0.1393 (0.0080)	0.0000	-0.1141 (0.0081)	0.0000
Categories – HPE	-0.5416 (0.0095)	0.0000	-0.2592 (0.0101)	0.0000	-0.6269 (0.0101)	0.0000	-0.1928 (0.0102)	0.0000
Time	-0.7287 (0.0105)	0.0000	-0.4513 (0.0112)	0.0000	-0.8378 (0.0112)	0.0000	-0.3703 (0.0114)	0.0000
Categories – Regular Public * Time	0.0172 (0.0009)	0.0000	-0.0061 (0.0004)	0.0000	-0.0004 (0.0009)	NS (0.6365)	-0.0063 (0.0004)	0.0000
Categories – PE * Time	0.0562 (0.0013)	0.0000	-0.0129 (0.0006)	0.0000	0.0798 (0.0012)	0.0000	-0.0267 (0.0005)	0.0000
Categories – HPE * Time	0.0539 (0.0014)	0.0000	-0.0138 (0.0006)	0.0000	0.0818 (0.0013)	0.0000	-0.0328 (0.0005)	0.0000

Both in math and in language, PE and HPE “progressed” between T1 and T2 compared to private and regular public schools that did not significantly modify their math or language trajectory. Their level improved with time between T1 and T2 and SES gaps were reduced. Whereas, from T2 to T3, PE and HPE’ levels worsened significantly (about 2 to 3 times more) compared to private and regular public and the SES gaps widened between school categories.

Figure S2. Math and language at T1 considering all ages available in advance-in-age and late-in-age children. The linear relationship observation for typical age children was not seen for the children in advance (represented in dark blue) neither for those already one year late (represented in green). Results in Math and language were presented using the normalized percentage of success data (range: 0-100).

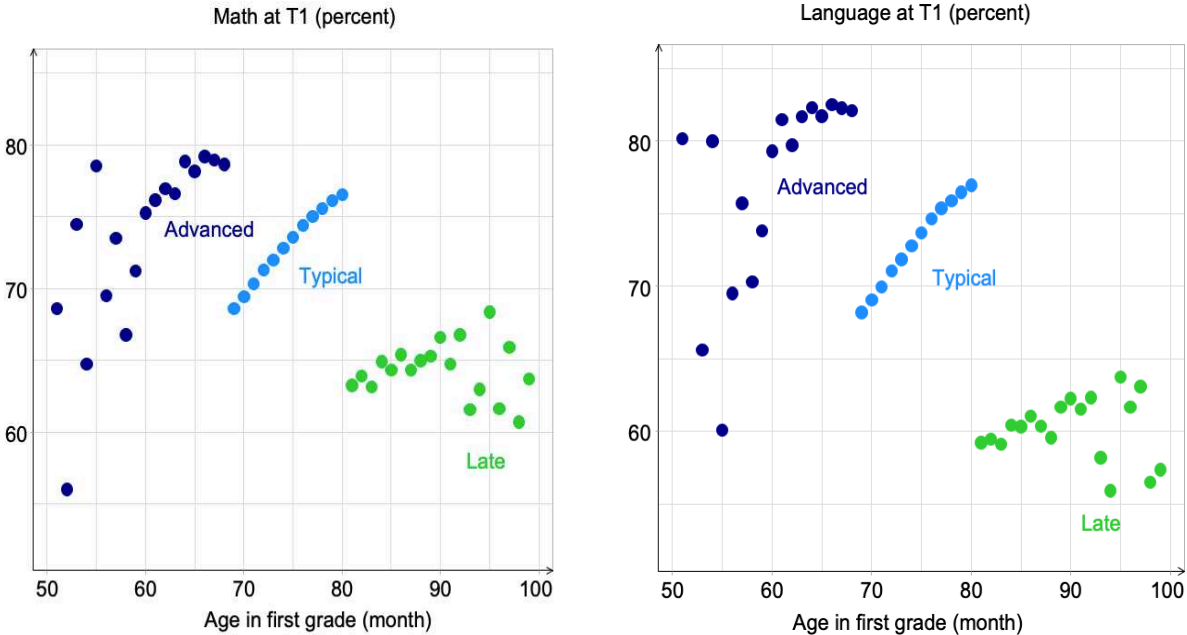


Figure S3. Panel of Math and Language in function of Age at T1, T2 and T3 in 2018, 2019, 2020 and 2021.

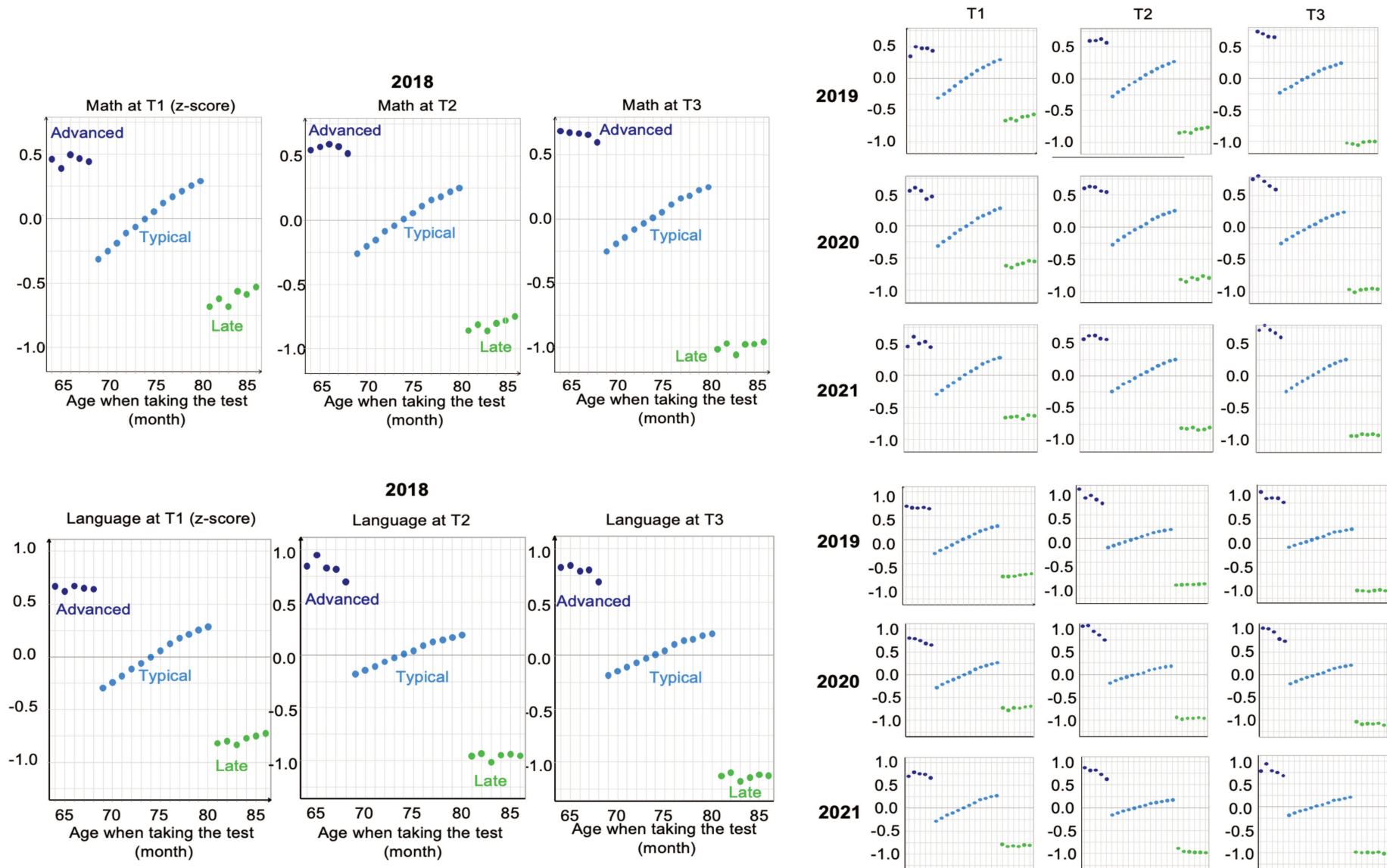
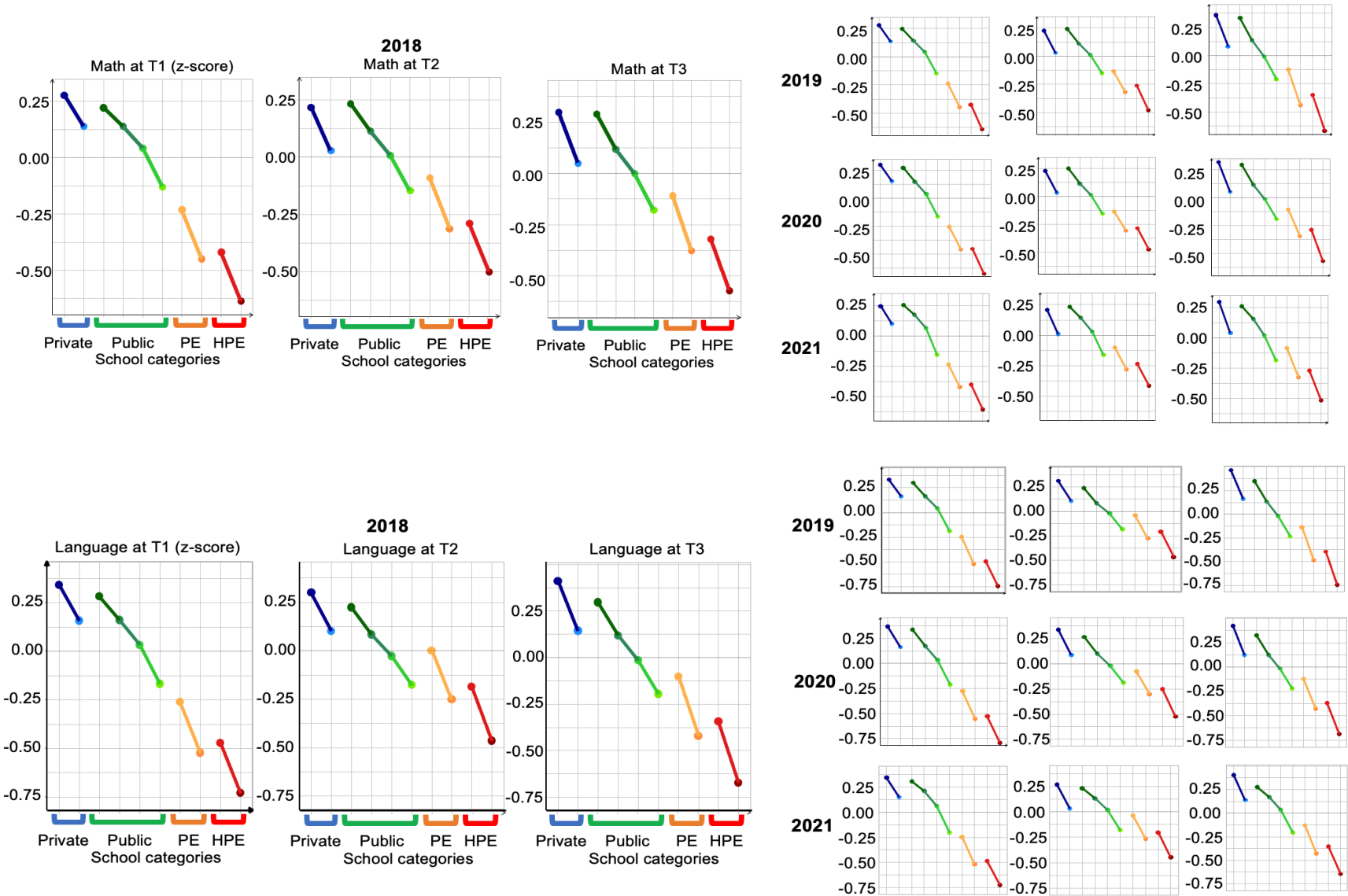


Figure S4. Panel of Math and Language in function of SES and types of school at T1, T2, and T3 in 2018, 2019, 2020 and 2021.



Chapter 3. Data sciences in education and learning how to read: predictors and learning patterns

I) Introduction

A. French students decline in language levels in the last two decades.

As introduced at the beginning of this work, the decline in language proficiency among young students, particularly in reading comprehension, has become a matter of concern in France. The international assessments of students' reading abilities, such as PISA and PIRLS have highlighted a worrisome trend: while other OECD countries have seen either a stability or an improvement in reading performances, France has witnessed a gradual decline (Colmant et al., 2008; OECD, 2018b). In 2012, 19% of the French 15-year-olds were low performers in reading literacy, compared to 15% in 2000 (*PISA 2012 Results*, 2012). Between 2000 and 2012, the proportion of children with severe difficulties increased by 4% in France, while it decreased by 2% on average in OECD countries (OECD, 2018b). According to the Department of National Statistical Institute in France (INSEE), about 18-20% of students in 2016 left primary schools with poor reading comprehension skills compared to 11 to 15% at the end of the 1990s (Daussin et al., 2011). The decline in reading comprehension is particularly evident in priority education schools, where nearly a third of students struggle with written words, marking a substantial increase over the past decade (Daussin et al., 2011). In addition, between 2003 and 2017, results of assessments on French 17-year-old children revealed a stability and indicated on average 23.1% of inefficient readers in France, 11.5% of children presented with high reading difficulties overall and 5.2 to 6.3% were illiterate (Daussin et al., 2011). As expected, the percentage of reading difficulties decreased with higher education: from 48.7 percent among those with no more than a college education to 4.8 percent among those who report having at least some general or technological education in high school. In the French overseas territories, the percentages of reading difficulties were much higher: around 30% in Guadeloupe,

Martinique and Reunion, 46% in French Guyana and 73% in Mayotte (De la Haye et al., 2018).

B. Reading and Reading comprehension: required skills and identified predictors.

The development of children's words **reading** skills (i.e., decoding words) involves the capacity to transform written words into spoken language, and this skill is typically evaluated by measuring the accuracy and speed of their oral reading (i.e., an ability called fluency). However, it is important to note that while efficient word reading is a crucial step, it is insufficient alone for the development of **reading comprehension** (Castles et al., 2018; Lervåg et al., 2018a).

On the one hand, a strong scientific consensus and body of evidence ended the 'reading wars' by documenting the importance of 'identifying the individual words' – and for that, the importance of phonics abilities (i.e., associate letters to sound, phonemes to graphemes) associated to a rich vocabulary comprehension, as skills needed for the development of reading abilities (Castles et al., 2018). Both skills are subordinate to the child's oral language abilities and thus requires acuteness of both auditory and phonetic representations (Gentaz et al., 2013, 2015a; Spencer et al., 2014), more particularly a good level of phonological awareness (PA; i.e., an individual's awareness of the sound structure of language) (Bianco et al., 2010; Melby-Lervåg et al., 2012). Acute auditory functioning matters as syllabic decomposition into its phonemes is dependent on the sound recovery of the word. The latter is accelerated if the word already exists in the child's vocabulary, and if the knowledge of the syntactic structure of the language can predict the type of words in a given context (Kirby et al., 2008; Melby-Lervåg et al., 2012). Moreover, the development of reading skills hinges on proficient visual acuity and attentional abilities. Similarly, as when needed to decode, a student must adeptly discern and isolate individual letters within a sequence of characters, necessitating precise control over their eye movements to pinpoint the correct location. This gradual progression involves the student shifting from a sequential, letter-by-letter reading approach (i.e., referred to as the serial mode) to a more advanced stage where the letters within a word are processed simultaneously (i.e., termed the parallel mode) (Dehaene & Cohen, 2011; Kolinsky et al., 2018). To

acquire quality reading, a child must also focus their attention on the correct letter while reducing interference from neighboring letters (Rayner K., 2016).

On the other hand, in 1990, Hoover and Gough introduced the 'Simple View of Reading,' which initially conceptualized **reading comprehension** as the result of two fundamental components: decoding words and oral language comprehension. These two elements were identified as the key factors contributing to the development of individual reading comprehension (Grainger et al., 2016; Hoover & Gough, 1990). Nonetheless, studies revealed that the relative importance of decoding versus language comprehension varied based on the developmental stage of students (i.e., kindergarten, early school years, later school years) and the complexity of the text (Lonigan et al., 2018b). As students' decoding skills advance and they tackle more intricate texts, oral language comprehension gains significance over decoding (Catts et al., 2005; Hoover & Gough, 1990; Tilstra et al., 2009). However, there are unresolved questions about reading comprehension related factors and results are inconsistent: the relative predictive weight and importance of different language skills remains unclear regarding the development of reading comprehension. Indeed, previous studies targeted different populations, varying in size (~ 35 to 300 children to meta-analysis of ~ 30,000 children), varying in age (starting at 4 to 7.5 years of age and adults (i.e., analyzing expert readers to identify the reading comprehension predictors is one strategy that many papers adopted), varying in SES score (from various background) and varying in their orthographic language codes (e.g., English, Italian, French ...): for instance, studies nuanced the Simple view of reading, identifying a predominance of either (1) word decoding (Kendeou et al., 2009; Lauterbach et al., 2017), (2) oral language abilities (Bianco et al., 2012; Massonnié et al., 2019; Pinto et al., 2016), (3) richness of vocabulary (Currie & Cain, 2015; Dong et al., 2020; Quinn et al., 2015; Roth et al., 2002), (4) listening or language comprehension (Hogan et al., 2014; Kim, 2016; Lervåg et al., 2018b) or (5) a combination of all with the working memory (Perfetti & Stafura, 2014), as the main predictors of reading comprehension abilities. These disparate findings justify the need to establish a more precise overview of predictive weights on reading comprehension capacities among first and second graders. We have a unique opportunity to do so, thanks to the Evalaide national

programme, on a large and complete cohort of children followed for 12 months, four years in a row from 2018 to 2022.

C. Reading and reading comprehension difficulties: different predictors.

An additional strategy to explore predictors would be to analyze children with reading comprehension difficulties in second grade and identify the factors specifically associated with them compared to the general population. At this stage, it is too early (i.e., at the beginning of second grade) to diagnose the most common reading disorders called 'Dyslexia' (i.e., a difficulty in learning to decode print and to transform it into speech, problems with accurate or fluent word reading, poor decoding, and poor spelling 'that must have persisted for at least 6 months, despite the provision of interventions that target those difficulties) which manifest in children with normal intelligence and social behavior, adequate oral comprehension skills, and no sensory problems (vision, hearing (Snowling et al., 2020)) and it is characterized by a poor association between graphemes and phonemes, as well as an inability to quickly grasp a word in its entirety (Sprenger-Charolles L. & Colé, 2013). The child reads slowly and makes errors that are persistent and approximately 5 to 7 % of children are affected (Peterson & Pennington, 2012). Dyslexia has a strong genetic component, but it is also modulated by factors such as the transparency of the writing system, and the socio-economic background. Currently, we do not fully understand the exact mechanisms responsible for these disorders. Conversely, studies with beginning readers have shown that a wide range of predictors allow to identify children-at-risk of developing reading comprehension difficulties and these predictors vary with the child development (Adlof et al., 2010, 2017; Bianco et al., 2014). Additionally, vocabulary knowledge and syntactic skills have been linked to oral comprehension challenges, as children with limited vocabulary or difficulties in understanding sentence structures may struggle to grasp the meaning of spoken language (Castles et al., 2018; Nation et al., 2010; Storch & Whitehurst, 2002). Moreover, socioeconomic factors such as low family income and limited access to language-rich environments have been identified as contributors to oral comprehension difficulties and to reading comprehension difficulties, emphasizing the importance of addressing these disparities in early

childhood education (Chen et al., 2018; Gentaz et al., 2013, 2015b; Stanovich, 1986). Low socio-economic status (SES) had a notable impact on predictors of reading comprehension, particularly in first-grade children with varying decoding skills. For example, phonological awareness significantly influences reading comprehension in children with poor and average decoding skills, while listening comprehension plays a more substantial role in children with good decoding skills (Billard et al., 2009; Fluss et al., 2009; Gentaz et al., 2013, 2015b). Poor readers from low socioeconomic backgrounds present a similar pattern to the classic dyslexic population (Billard et al., 2010).

D. Aims of this work and research hypothesis

For the first time, with the help of massive data of the entire population of French first graders, four years in a row, we can answer several questions regarding learning how to reach a good level of reading and of reading comprehension, more precisely targeting the identification of specific related cognitive domains, but also the existence of subgroups with similar learning pathways, and the role of schooling in addressing language progresses at the national level. Our strategy was to begin with a systematic exploration of language assessments, examining their interrelationships and variations across three time points (T1, T2, and T3). This initial analysis prompted several key inquiries. First, we postulated that certain cognitive domains might exhibit high correlations, such as the relationship between oral comprehension of individual words and oral comprehension of complete texts. This hypothesis was empirically explored within our study. Second, our study aimed at identifying individual predictors associated with both achieving proficient reading and reading comprehension between the first and second grade. Our investigation also aimed at detecting subgroups of children with similar skill profiles, which could inform future targeted interventions on language development: we asked whether distinct groups or patterns existed among children, delineating their learning pathways in both oral and written language. Subsequently, we sought to identify which specific cognitive domains of language were associated with later reading comprehension difficulties. Furthermore, we analyzed factors that predicted high levels of reading comprehension at the classroom level (i.e.,

class size, heterogeneity of level in a class, mixity, first of class in language being a boy or a girl). For instance, to comprehensively address these research questions, our study considered not only language predictors but also demographic factors such as age and socioeconomic status (SES) scores.

II) Material and methods

A. Materials

All the data management and cognitive assessments' descriptions were presented in Chapter 2. On this chapter 3, we focused on language assessments at T1, T2 and T3 registered between 2018 to 2022, including the including the Covid-19 specific year of 2019-2020, where the French first graders were in lockdown (i.e., off from school) for 52 consecutive days, and the French kindergarteners were away from school during 42 consecutive days.

In this chapter 3, we analyzed children that began first grade on the year of their 6th birthday, considered as “typical age” children, and did not present results of more specific populations of advance-in-age and late-in-age children in this work.

B. Methods

1) Correlation matrices

Implementing correlation matrices, we explored how related all the language assessments were correlated between each other. For this purpose, we used the package *Hmisc* and the *rcorr* and *corrplot* functions on the *software R*.

2) Principal Component Analyses (PCA)

Principal Component Analysis is a method based on reducing the number of variables of a large dataset, by transforming the latter into smaller data sets that concentrate most of the information needed to analyze. In order to identify the principal components of the data, we computed eigenvectors and eigenvalues (i.e., which are linear algebra concepts) from the covariance matrix.

Note that eigenvectors and eigenvalues always come as a pair, and their number equals the number of the dataset dimensions. Eigenvectors of the covariance matrix represent the directions (i.e. principal components) of the axes where there is the most variance (i.e., the most information). Eigenvalues are simply the coefficients attached to eigenvectors, which give the amount of variance carried in each Principal Component. By ranking the eigenvectors in order of their eigenvalues (i.e., highest to lowest) we obtain the principal components in order of significance.

Therefore, **principal components** are the new variables designed as linear combinations of the initial variables of the data set. These combinations are made in a way that the principal components are uncorrelated and most of the information within the initial variables is compressed into the **first components**. The PCA process aims at gathering the maximum of information into the first component, then, adding the maximum of remaining information into the second component and so on, **each dimension representing the direction of data with the maximal amount of explained variances**.

To summarize, PCA allowed us to explore the children's distribution and correlation between each language score at each period. These analyses revealed that at each period the average language score was highly correlated with the first principal dimension, meaning that it best summarized the overall variability between children's level. Overall, we performed the PCA per period and share the results of all language items at T1, T2 and T3.

Before starting PCA, using raw scores, we supposed that data were missing at random (MCAR) and dealt with missing data by using the package *missMDA*. We imputed missing data using the command *imputePCA* on the *R software*. We used all the raw grades to complete missing periods (period for which we had no information at all, for example because the child was not present the day the exam was made) (Josse & Husson, 2016).

Note that the *missMDA* package provided a powerful tool that relied on a model that considered the inherent relationships between individuals and variables in the dataset. The specific modeling methods employed depend on the nature of the data, with Principal Component Analysis (PCA) being used for continuous data, Multiple Correspondence Analysis (MCA) for categorical data, and Mixed Data Factorial Analysis (MDFA) for datasets that contain a mix of both continuous and categorical data. We followed different steps for this MDA imputation: Firstly, an initial imputation where missing values were replaced with the mean of the observed values for the corresponding variable, a step that provided with a preliminary complete dataset. Then, a PCA was conducted on this initial dataset, identifying the line that minimizes the distances between data points and their perpendicular projections onto this line, known as perpendicular errors. This line also maximized the spread of point projections along it. After, the values predicted by PCA for the missing data were used as replacements for the missing values in the dataset. Following this imputation, a new PCA was performed on the dataset with the updated values. These iterative steps were repeated until convergence was attained. Convergence was achieved when the new values predicted by PCA were either identical or very close to the values predicted by the previous PCA. This iterative process ensured that the imputed values aligned closely with the underlying structure of the data, and the final dataset was a reliable representation of the complete dataset while preserving the relationships between variables and individuals.

PCA were then performed using the R package *FactoMineR* (Lê et al., 2008) on the first grader whole-population, both in 2018, 2019 and 2020. Inside each population, one PCA was implemented per period (T1, T2, T3) using all the raw score results in language. Note that when performing the PCA, all covariates were scaled and centered on zero.

3) Multilevel linear regression model

Studying environmental effects on individuals underlies each element's independency and their belonging to different statistical units: some are micro-units and others are macro-units, all of them are hierarchized with a micro-unit being the individual, an intermediate unit being the class, and a macro-unit being the school. These different contexts are nested. In this study, children were taught within classes, all nested within schools, and therefore experienced a variety of stimuli at different levels. Due to these different environments, data contained natural groupings which had an impact on the individual's performance in language. These multiple levels induced observations that were not independently sampled from one another (Bressoux, P., 2010).

Facing such data, we implemented multilevel linear mixed models allowed to overcome these two limitations of conventional models accounting for sources variation in the data and not assuming independently sampled data. Language nested patterns were introduced in the intercept and in the slope at the class levels.

All continuous independent covariates were centered and reduced before calculation of means in language at T1, T2 and T3, as well as before calculating heterogeneity of level at T1. To obtain orthogonal measures for gender effect, each boy was attributed a score of 0.5 and each girl was attributed a score of -0.5. Among all independent covariates, only gender and school categories remained non-scaled. Multilevel Linear mixed model, fit by maximum likelihood, were performed using the R package *lmerTest*, and allowed to estimate several individual and environmental parameters regarding the level in reading comprehension at T3 and the level of reading at T3. The global effectiveness of classes is represented by the model's class-level variance intercept of a parameter estimate. We considered the population of a typical age when entering in first grade (i.e., 69- to 80-month-old), and selected classes that allowed to calculate the class characteristics (such as boys-girls proportion, class heterogeneity, first of class in language).

4) Anova

ANOVA is a statistical technique used to analyze the variation among means in multiple independent groups or cohorts to determine if there are statistically significant differences between them. It is particularly useful as we want to compare the means of three or more independent groups on the same assessment or dependent variable (i.e., 4 groups which are the 4 studied years – 2018, 2019, 2020 and 2021). We implemented ANOVA models on the population having a typical age when entering in first grade at school (i.e., 69 to 80 months), and used gaussianized data for applying the function *anova-test* of the package *rstatix* in R software.

5) Alluvial

An alluvial graph, also known as a Sankey diagram or flow diagram, represents a powerful tool for visualizing and understanding the progression of language skills or any categorical data over time, showing the flow or transition of data. It's often used to represent changes or progress over time or across different conditions. Alluvial graphs use a series of connected vertical columns (nodes) to display the evolution or movement of data categories from one state to another. We implemented Alluvial graphs as they are particularly useful for tracking the trajectories of progress in language between three periods (T1, T2, and T3). We implemented alluvial models and graphs on the population having a typical age when entering in first grade at school (i.e., 69 to 80 months), and used gaussianized data for applying the function *alluvial* of the package *alluvial* in R.

III) Results

Across the following chapter, our approach was multifaceted. Firstly, we provided a detailed characterization of children performances in language subdomains at T1, T2 and T3. Next, through the utilization of both Principal Component Analysis (PCA) and multilevel models, we sought to determine whether there were observable groups or patterns of children with analogous learning trajectories in oral and written language at T1, T2 and T3. Our inquiry aimed at addressing this critical question: “Were there any identifiable groups or patterns of children with similar learning pathways in oral and written language?”.

For instance, we implemented multilevel models to estimate the significance of predictors associated with enhanced performance in both reading abilities and reading comprehension. Subsequently, we conducted comparative analyses, contrasting these characteristics with those of other children, both within the remaining four quintiles and with those in the highest-performing quintile. The objective was to pinpoint specific attributes and learning needs unique to the group facing difficulties in each domain.

Were the learning processes similar between low achievers and high achievers?”. We also delved into assessing their capacity for improvement, gauging their ability to transition from lower quintiles of performance to higher ones. A particular emphasis was placed on examining these transitions within the context of PE (Priority Education) and HPE (Higher Priority Education) public schools, where children encountered more pronounced challenges compared to other school categories.

Finally, we wondered if school had any impact on their progress by observing on one hand the summertime effect (i.e., the long 2-month holiday break from school) and on the other hand, by observing the COVID natural experiment of an absence of school in 2019 and 2020 compared with 2018 and 2021. Our focus remained on the population of children of typical age when beginning first grade (i.e., 69- to 80- months old). By adopting this comprehensive approach on these massive data, we aimed at

shedding light on the intricacies of learning pathways for reading comprehension at T3 and learning pathways for children with difficulties to provide valuable insights into the factors influencing linguistic skills development in struggling learners.

A. Descriptive analyses of correlation matrices

In Chapter 2 entitled 'Data Management and Description,' all language assessments conducted at three distinct time points (namely T1, T2, and T3) were detailed, as well as their correlations. Between all three periods, only one test was identical between T1 and T3: the oral comprehension of words – which would have been interesting to explore in detail but unfortunately, we did not have any access to the specific words answered by children either at T1 nor at T3, we only had access to their global scores on oral comprehension of words.

All the other language tests changed in their content between T1, T2 and T3, as they fit the children's learning expectations between T1, T2 and T3 (e.g., Oral comprehension of sentences presented more complex sentences to understand from T1 to T3). Two language items presented with abnormal higher scores only in 2018, compared to 2019, 2020 and 2021: oral comprehension of texts at T1 (i.e., 13.27 (3.64) vs. 8.24 (2.26) in 2019) and comparing letters at T1 (i.e., 15.23 (6.69) vs. 10.06 (2.72)) (see Chapter 2 - **Table S1**). Unfortunately, the DEPP technical team could not explain these differences, and we supposed they could be due to a bias created only in 2018, as this was the first year of national tests implementation, where ~ 150,000 less students were included compared to other years ($n = \sim 600,000$ in 2018 vs. $n = \sim 750,000$ in 2019, 2020 and 2021), notably, lesser PE and HPE students were included in 2018 (i.e., 10.0 % PE in 2018 vs. ~ 10.7% in other years, and 6.6% HPE in 2018 vs. ~6.7 to 7.0% in other years – see Chapter 2 - **Table S1**).

Figure 9 showed the distribution of all language tests at T1, T2 and T3, when all tests previously went through normalization and ranged between 0 to 100 percent of success. A focused analysis of the distribution of reading assessments in 2018 yielded valuable insights, notably the rapid saturation of some tests that did not allow to

measure a proper variability in our population, but also, these distributions shared hints about the relative difficulty levels of these assessments (see **Figure 9**).

Notably, when considering all four years (2018, 2019, 2020, and 2021), it became apparent that children entering first grade did not exhibit all required skills in language, either in oral comprehension, in meta phonology or in decoding skills. However, it is important to have in mind that the diverse performances observed among children should not be automatically interpreted as test failures. Indeed, in France, first grade marks the beginning of formal education, and teachers are required to follow a national curriculum that allows flexibility in implementing various subdomains throughout the first year of school, whereas kindergarten does not follow a national formal teaching curriculum in language nor in math.

Briefly, France makes a strong distinction between “maternelle” (which includes kindergarten and preschool and is mandatory from age 3) versus “école” (school). Maternelle is mostly based on play and socializing activities, and teachers and their unions are adamant that its young children should benefit from a playful environment largely devoid of the pressures associated with formal schooling (Ministère de l’Education Nationale, 2023). Therefore, children were not necessarily exposed to the same level of linguistic skills development as they were in their first-grade experience and being exposed for the first time to new exercises in language can explain lower results rather than failure in the specific language domain.

Having these programme and teaching disparities in mind, among all language tests assessed at T1, T2 and T3, we were able to ‘estimate’ which tests were achieved relatively more easily (i.e., performances were high for a large part of children) which were, at the meantime, tests that saturated and made it hard to distinguish children’s level (see **Figure 9**).

Notable examples of such tests included ‘sentence comprehension at T1’ and ‘letter recognition at T2’ (respectively of 87% and 94% of test success in Chapter 2 - **Table 10** in 2018). Conversely, other tests presented greater challenges, with a greater

variability and no test saturation such as 'phoneme handling at T1' (i.e., 59% of test success in Chapter 2 - **Table 10** in 2018) (see **Figure 9**). A detailed visual representation and a replication of these findings were found and presented for the years 2019, 2020, and 2021 in **Figure S5**, **Figure S6** and **Figure S7** in the supplementary materials.

The three means in language at T1, T2 and T3, composed of normalized scores' means per period, presented with wide gaussianized distributions, allowing to discriminate children' levels in language more easily at T1, T2 and T3 compared to a reduced distribution or any saturated item (see **Figure 9**). In addition, and as explained in chapter 2, we identified abnormalities in the data collection for the four tests of reading words and texts at T2 and T3. Presented on **Figure 9**, these tests underwent a data transformation (keeping a threshold where 97% of individuals answered) and normalization with a range of 0 to 100, whereas raw tests, keeping official maximums (i.e., 30, 29, 60 and 102 respectively), were presented on **Figure 10**.

As a reminder, reading words and texts were individual assessments made with an adult that registered the number of correct read words or texts were performed in one minute, on a list of respectively 30 words, 29 words in texts, 60 words and 102 words in texts at T2 and T3. Distribution of raw scores showed an important number of children facing difficulties in reading words and texts in one minute, particularly 12 months after the beginning of learning how to read (see **Figure 10**). These difficulties will be explored below in the subchapter "difficulties in reading and reading comprehension".

Figure 9. Distribution of normalized language assessments at T1, T2 and T3 (range 0-100) in 2018.

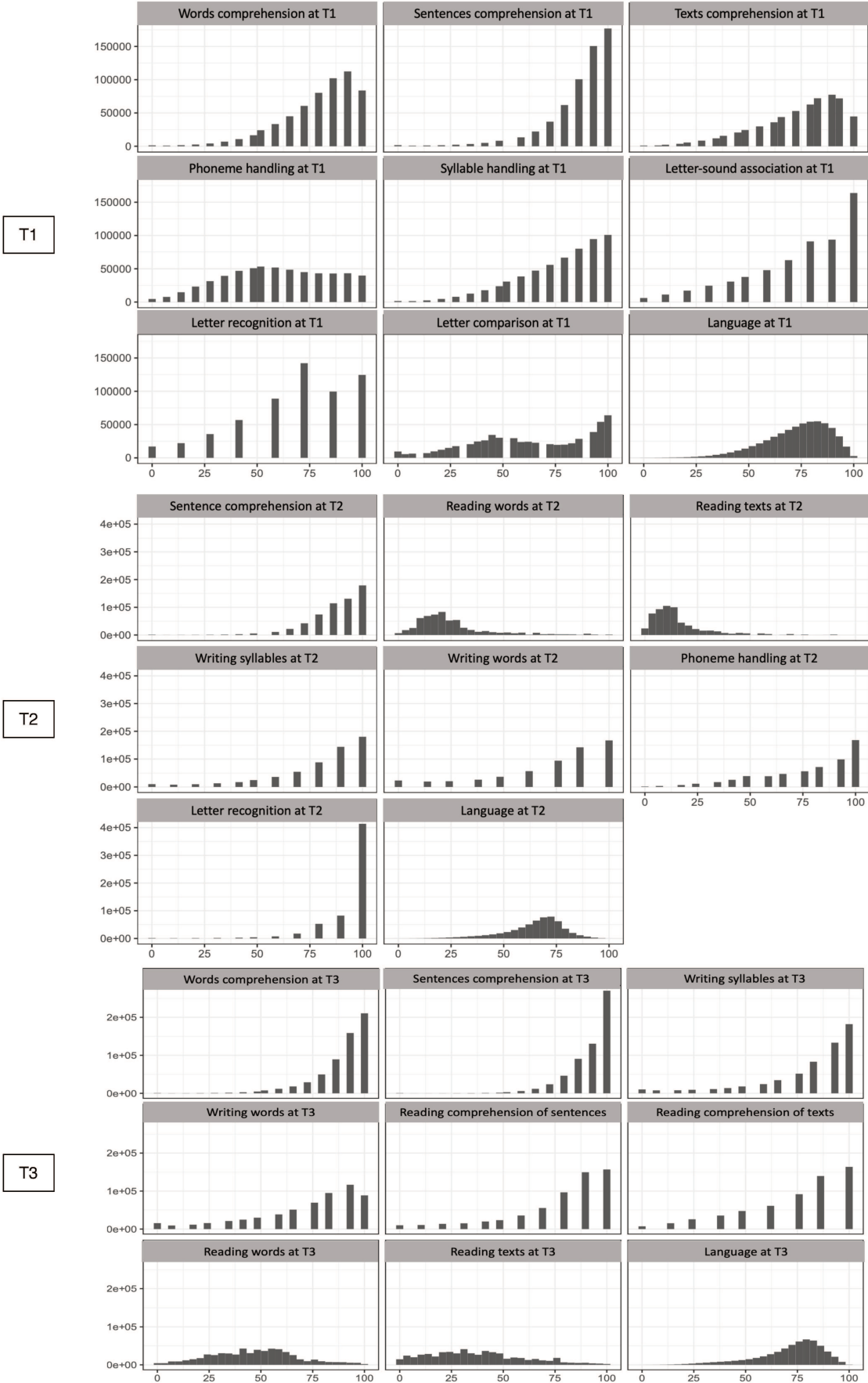
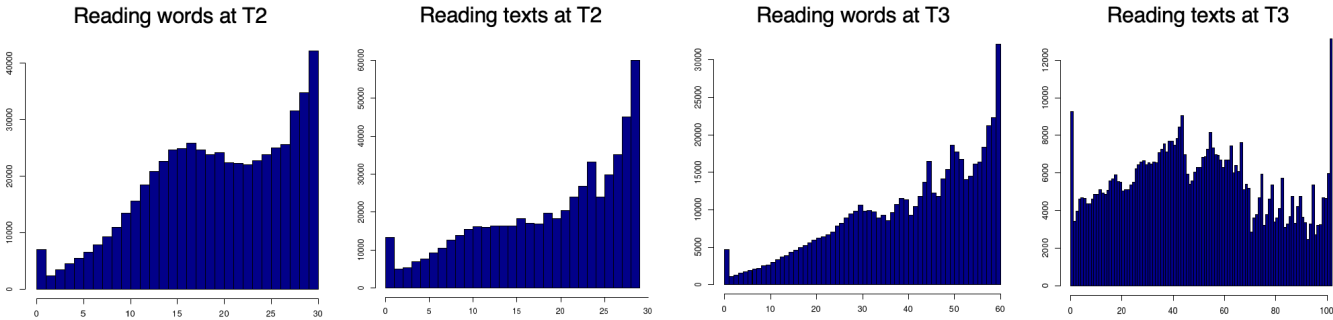
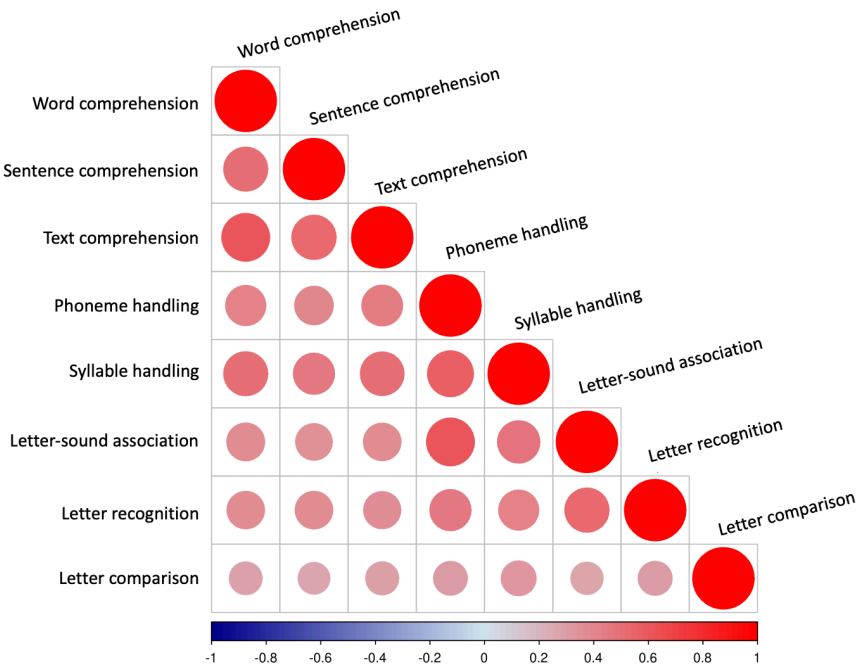


Figure 10. Distribution of raw scores in reading words and texts at T2 and T3 in 2018 for all typical-age children.



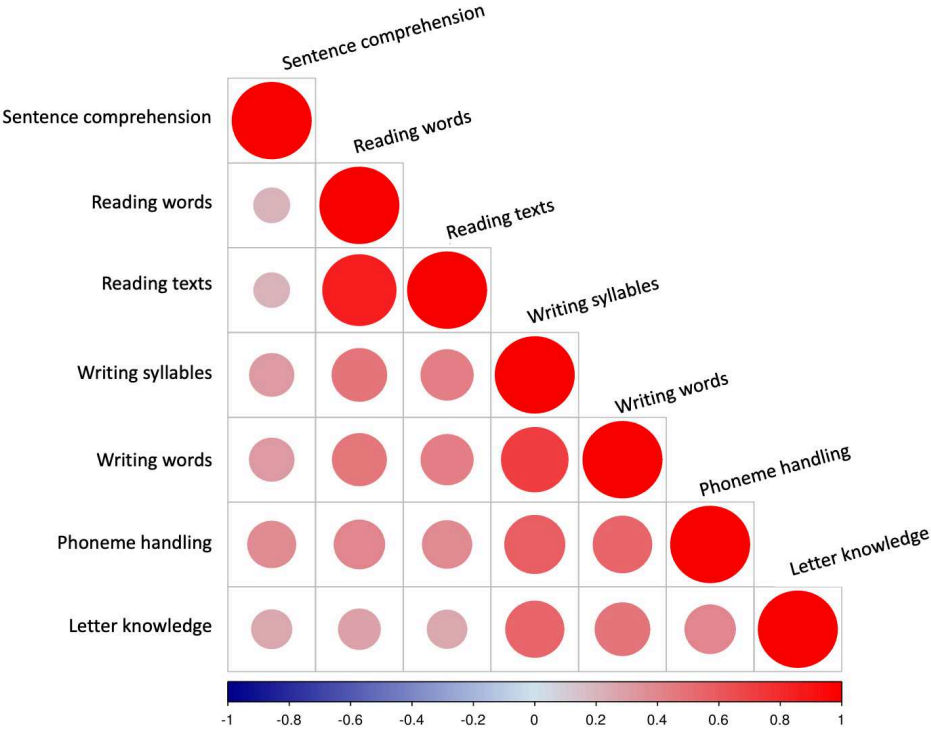
Pursuing the goal to analyze relations between variables in language, we delved into the correlation matrix of language variables at T1, we were able to discern distinctive patterns. First, the three oral comprehension tests exhibited stronger correlations among themselves compared to the other tests at T1, as depicted in **Figure 11**. We observed a second pronounced correlation cluster encompassing phoneme handling, syllable handling, and letter-sound association, which distinguished them from the remaining variables. These consistent findings were reaffirmed through replication in the subsequent years, namely 2019, 2020, and 2021, as visually represented in **Figure S8** within the Supplementary Material (SOM).

Figure 11. Correlations between language exercises at T1 in 2018.



At T2, we observed distinct patterns in the variables' correlation matrix. Firstly, reading words and texts displayed a strong mutual correlation, and they also exhibited correlations with phoneme handling and writing words. This interconnectedness was noteworthy and indicative of shared attributes among these variables. Conversely, letter-sound association demonstrated a notably higher correlation with writing syllables, phoneme handling, and writing words compared to its correlations with either oral comprehension of sentences or reading abilities. These findings were visually represented in **Figure 12**. Importantly, these patterns persisted and were consistently replicated in the subsequent years—2019, 2020, and 2021, as evidenced in **Figure S9** within the Supplementary Material (SOM).

Figure 12. Correlations between language exercises at T2 in 2018.



At T3, oral comprehension of words and sentences correlated significantly with reading comprehension items, while writing syllables, writing words, and reading abilities showed stronger intercorrelations. Both reading comprehension of sentences and reading comprehension of texts maintained consistent and equally strong correlations with all items, ranging from 0.41 to 0.58 (as seen in **Figure 13**). These patterns were

replicated in subsequent years, 2019, 2020, and 2021, as shown in **Figure S10** within the Supplementary Material (SOM).

Figure 13. Correlations between language exercises at T3 in 2018.

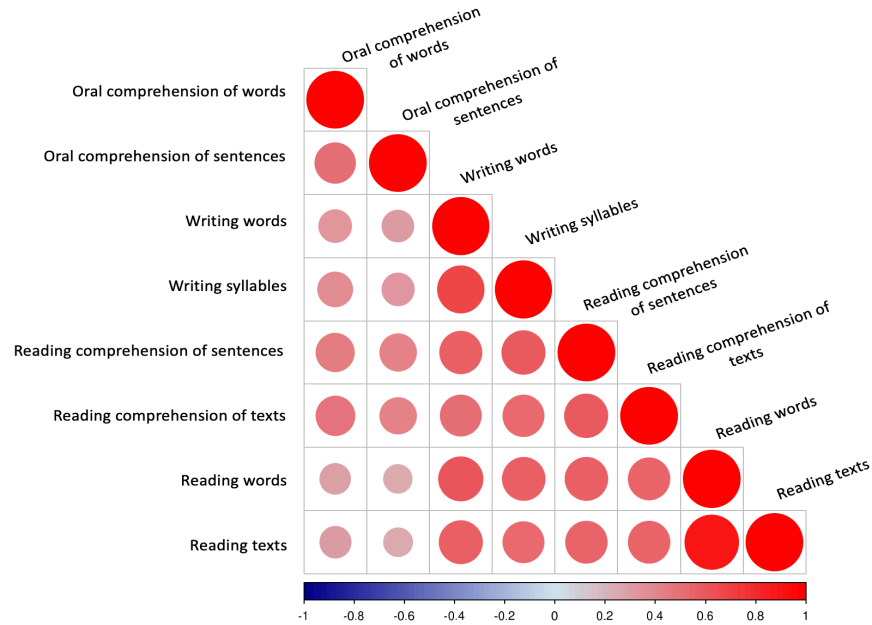
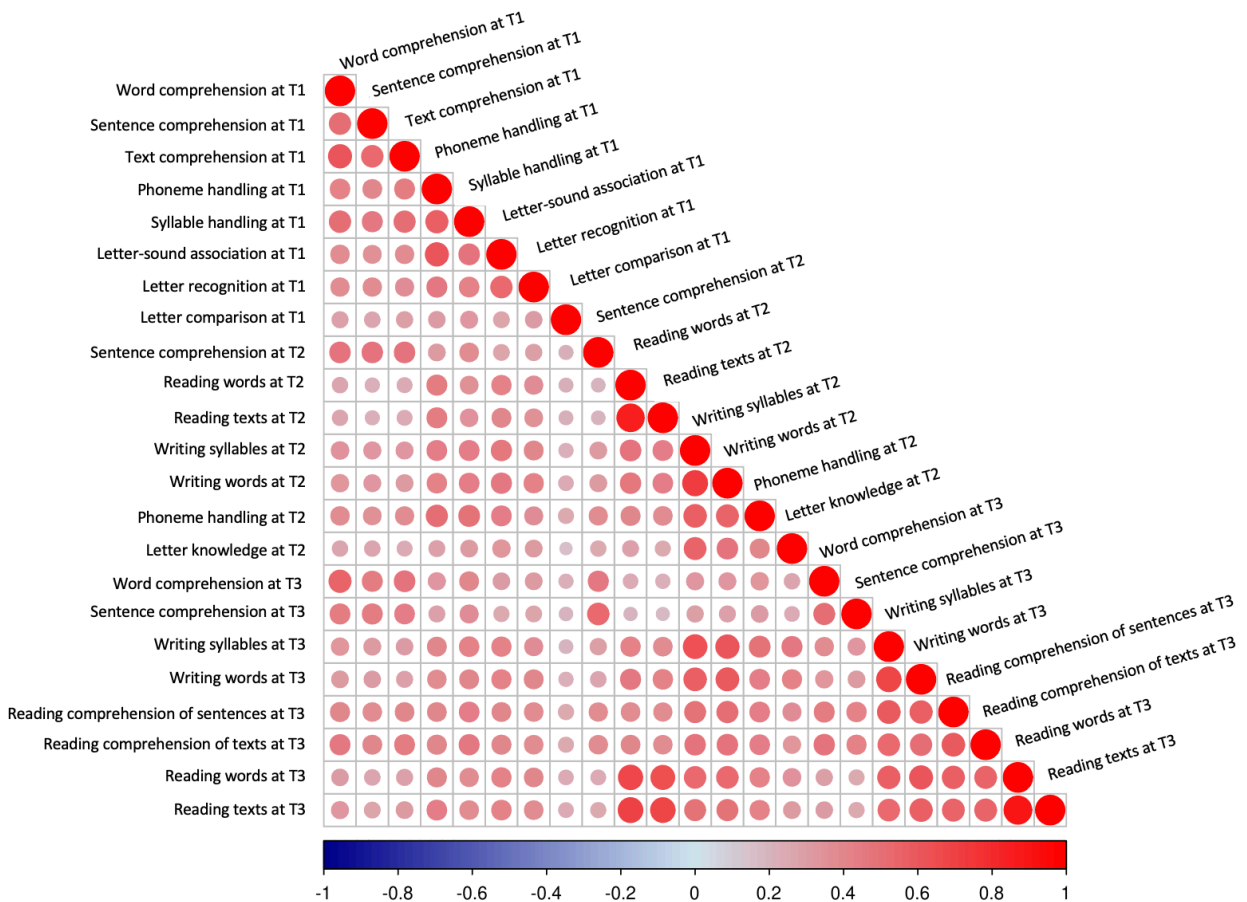


Figure 14. Correlations between language exercises at T1, T2 and T3, results for the 2018 cohort.



Combining language variables from T1, T2, and T3, we identified three distinct groups based on their strong correlation coefficients. The first group comprised oral comprehension variables, the second group included phoneme and syllable handling along with letter-sound association, and the third group consisted of decoding variables, specifically ‘Letter recognition’ and ‘Letter knowledge,’ as depicted in **Figure 14**.

In summary, as anticipated, all language variables exhibited positive and significant correlations with one another. To refine the identification of subgroups characterized by variables evolving in the same direction, mobilizing similar cognitive skills, and further refine the common difficulties experienced by children, we implemented a Principal Component Analysis (PCA).

B. Descriptive analysis using Principal component analysis – PCA

The PCA helped us answering the question “which language subgroups of variables were correlated (i.e., predicted) a similar skills domain in language at T1?”. Each subgroup – defined as a common arrow direction on the PCA graph – predicts similar cognitive domain in language.

Tips for results interpretation of PCA

Principal Component Analysis focuses on data found in several domains: observations (or individuals) in rows, and variables in columns. This methodology allows to:

- Find group of individuals that are similar according to all the variables.
- Visualize correlation between variables.
- Find synthetic variables.
- Detect outliers.
- Reduce dimensionality and find the core messages of the data.

The PCA figures will include the following 3 parts:

- 1) **The first circle** represents 2 dimensions and all the items measured in the PCA.
 - the dimension 1 allocates data of children from left (children who failed the test) to the right (children who performed well on the test); all the dots on the right of the vertical axis are children who succeeded in the domain.
 - the dimension 2 allocates data from above and below the horizontal line, where children above the line will be good in the tests whose arrows belong to the same side (up or down). The data located at the opposite side of the arrow (e.g., in diagonal), indicate the children having difficulties in this specific item.
- 2) **The second figure** on the right, represents the cloud of dots. Every dot is a child, positioned in the different dimensions according to his results in each dimension calculated. It is necessary to use both the circle and the cloud of dots in order to interpret correctly the PCA.
- 3) **The last figure** underneath represents barycenters of categorial variables, with their confidence interval (i.e., the ellipse surrounding the barycenter). Here, we need to use all three figures to interpret the barycenter. Example: see interpretations below.

Good students in language were similar (i.e., the cloud of dots is condensed on the right of the figure, after the 0 vertical line), and there was more variability in children with difficulties (i.e., the cloud of dots is sparser on the left of the figure) (see **Figure 15**). Three groups of children with specific difficulties appeared: A group of children with oral comprehension (of words, sentences, and texts) difficulties, a second group with syllable handling difficulties, and a third group with letter-sound associations and phoneme handling difficulties (see **Figure 15**).

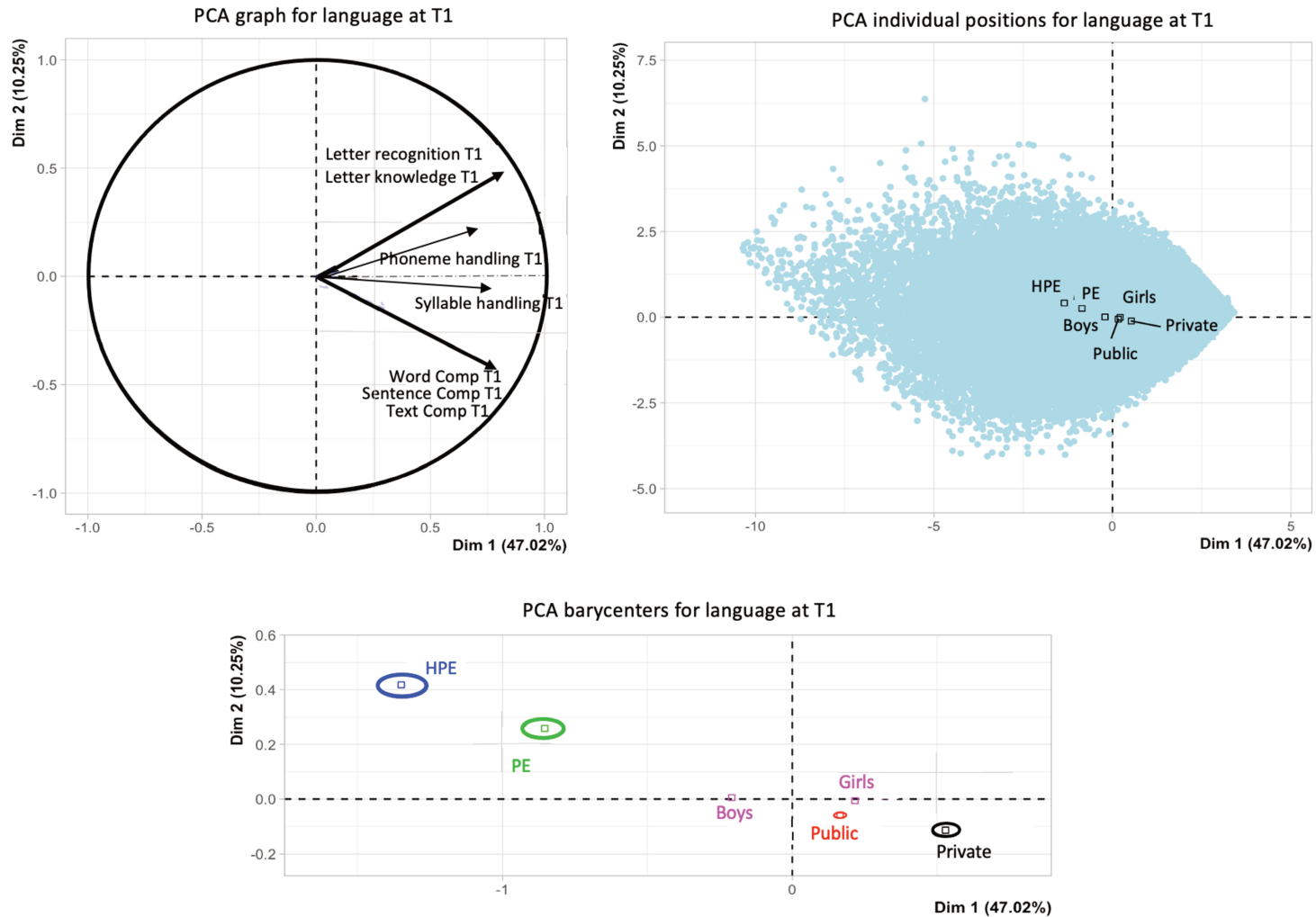
Additionally, we noted that girls performed significantly better than boys at T1 (i.e., girls' barycenter was located on the right part of the diagram from the vertical 0 line, compared to boys which were located on the negative side compared to 0). Also, social background modulated performances: PE and HPE both had more difficulties in the

oral comprehension domain at T1 compared to public and private schools (i.e., PE and HPE' barycenters were located on the opposite of the oral comprehension arrows, and on the left side compared to public and private schools' barycenters).

The PCA graph showed and confirmed there was a saturation of some language tests at T1, explaining the “arrow-form” on the right of the figure (i.e., many children are gathered in this arrow, with very similar results in language items, closed to the horizontal 0 line). Finally, the top left graph with arrows, indicated that children tended to perform differently in three main different groups of assessments in language: oral comprehension, syllable handling and phoneme handling and, decoding skills (i.e., when children perform similarly in a specific domain, their arrow superposed to each other and indicate a similar direction and amplitude – this latter information is given by the length of the arrow).

Altogether, these first PCA graphs allowed to differentiate children regarding their skills in three main domains of language at T1. These results were similar and replicated in 2019, 2020 and 2021 (see **Figure S11** in SOM).

Figure 15. Principal component analysis of language items at T1 in 2018.



*These figures resulted from the collaborative work with Dr Bénédicte Colnet, INRIA.

Figure 16. Principal component analysis of language items at T2 in 2018.

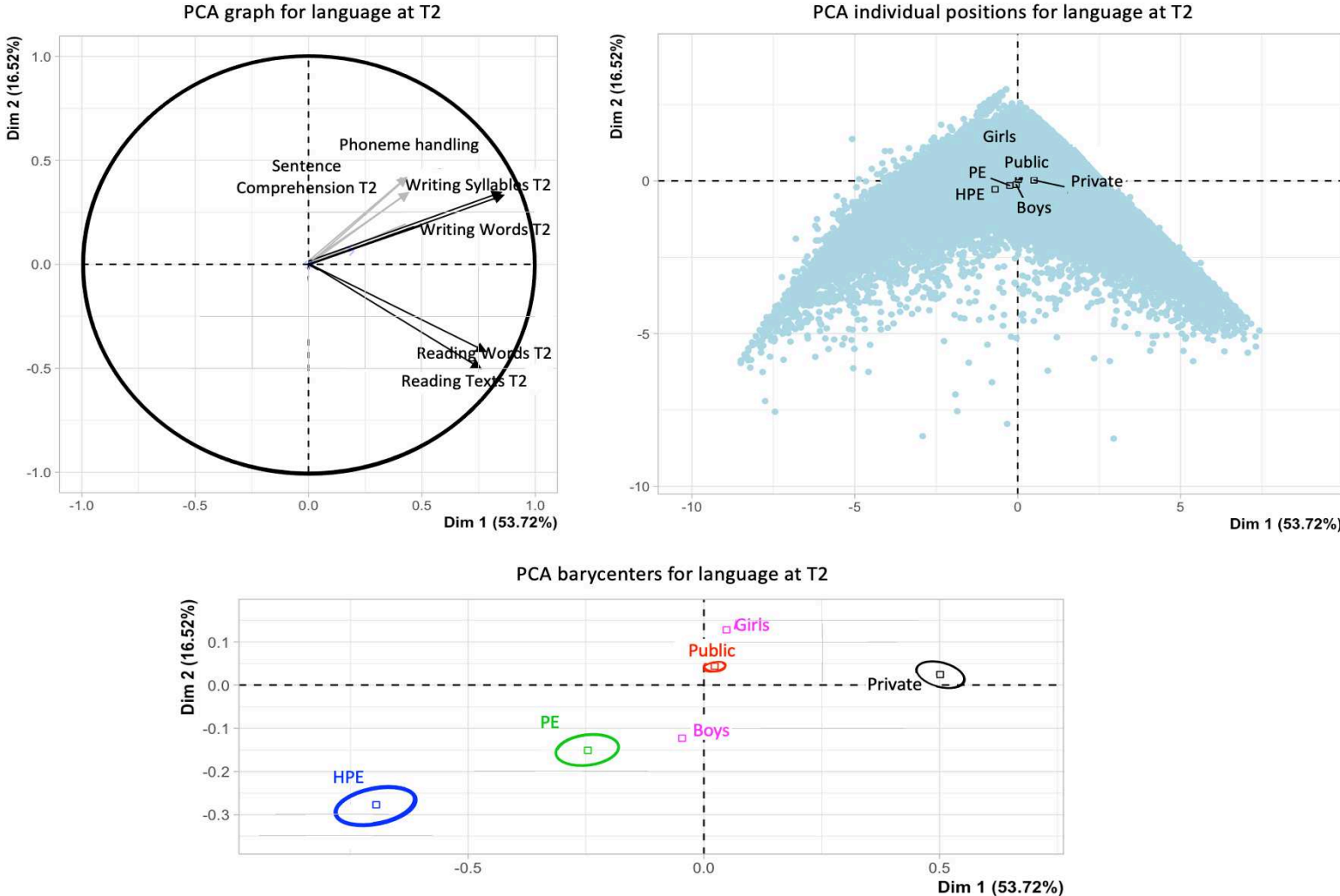
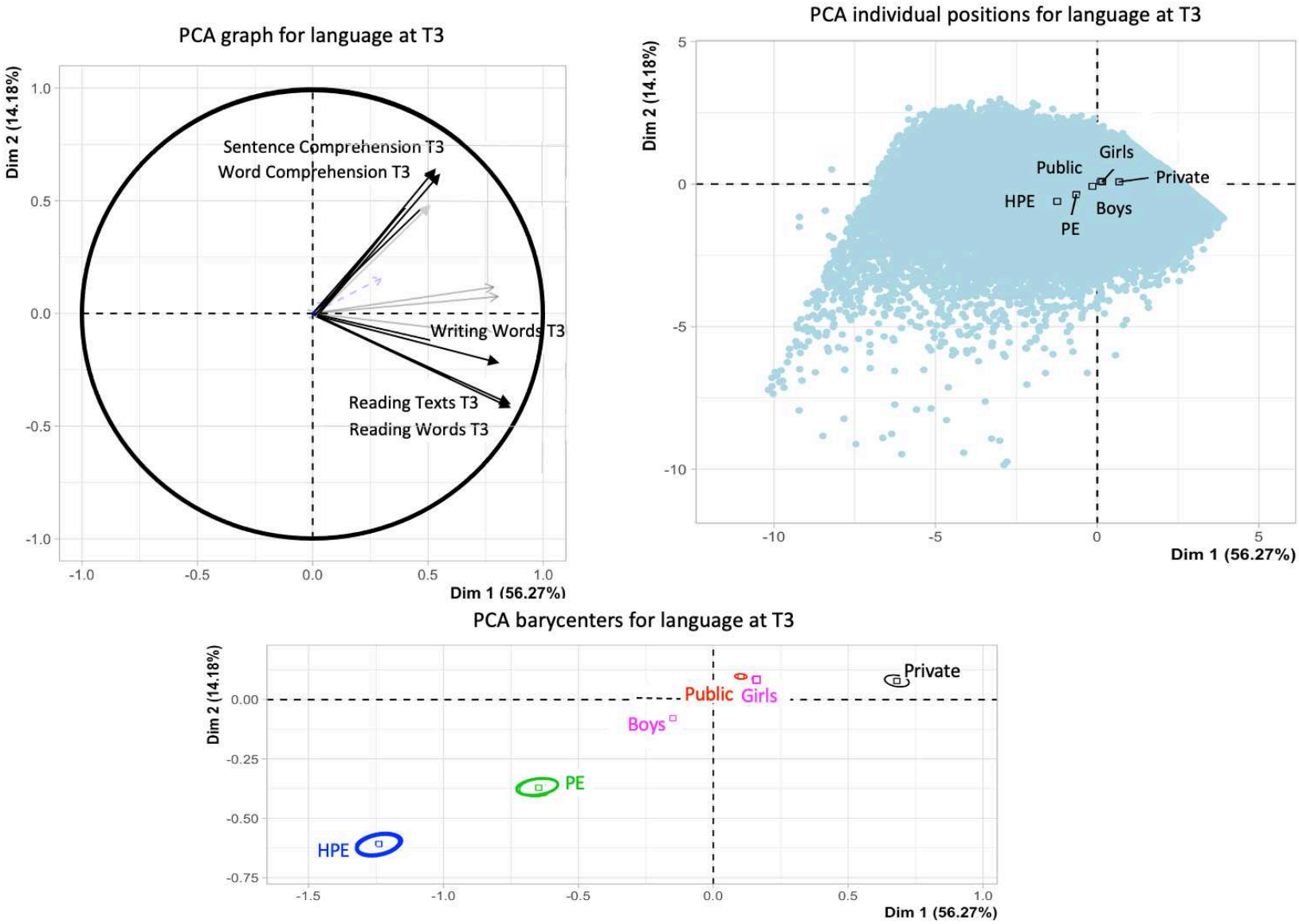


Figure 17. Principal component of language items at T3 in 2018.



At T2, the highest variability between children among all language assessments at T2 belonged to two groups: (1) the group that struggled a lot in writing words, writing syllables, phoneme handling and, oral comprehension at T2; (2) the group that overperformed in reading words and texts at T2 (see **Figure 16**).

More precisely, there was a saturation of tests in writing, syllables/phoneme handling and oral comprehension at T2 (i.e., representing the “upper-roof form” of the data of the right, and there was a saturation of tests in reading at T2 as well, among the worst level of readers mainly, explaining this “roof-like form” on the left of the graph. Girls tended to perform better in *writing syllables, words and phoneme handling* than *reading* compared to boys (i.e., Girls’s barycenter position was located in the top right quarter, in the same direction than the writing and phoneme items. On the contrary, boys tended to perform better at T2 than at T1 (i.e., Boys’ T1 barycenter was of -0.25 on **Figure 15** vs. Boys’ T2 barycenter was of -0.03 on **Figure 16**) and boys presented with more difficulties at T2 in writing and phoneme handling than in reading (i.e., Boys’ barycenter was located on the lower left quarter of the PCA figure, opposite from the writing and phoneme handling arrows).

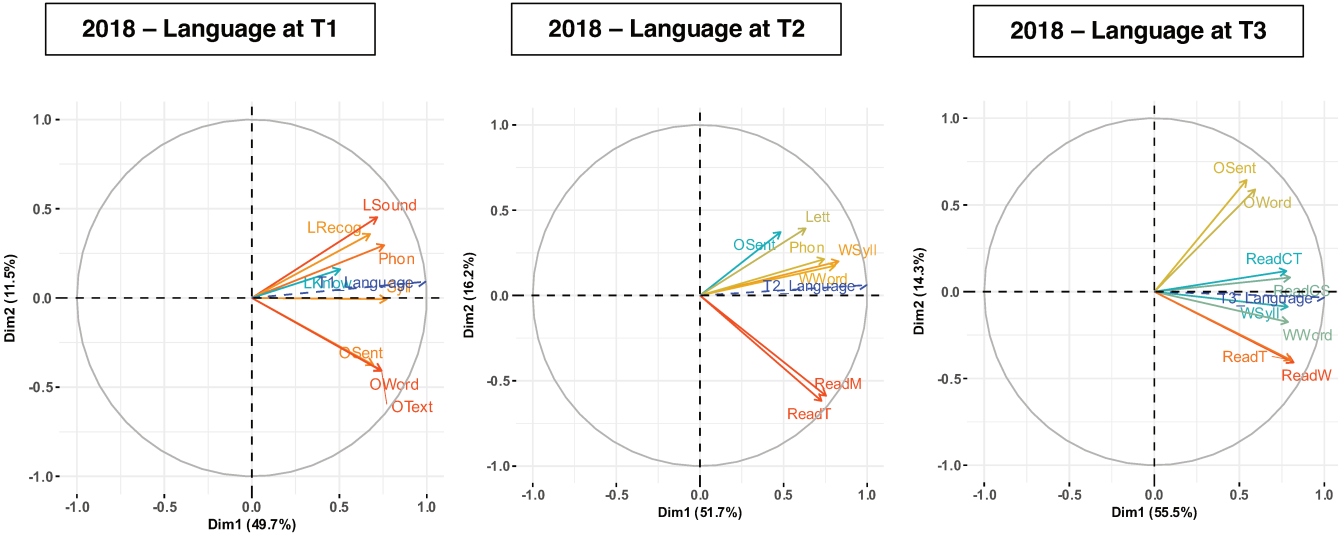
Overall, girls and boys performed more similarly at T2 than at T1 (i.e., Boys and girls’ barycenters are closer on the dimensional graph). The reduction of this gender gap could be explained by the boys’ s higher level in reading abilities at T2 compared to girls’s. A specificity of these two latter tests, was their time-limitation (i.e., reading abilities were measured in less than a minute). In line with the literature, boys tended to perform better in competitive exercises (Buser, 2014; Buser et al., 2021; Stoevenbelt et al., 2023) and one could interpret their specific higher level in reading due to this competitive assessment aspect.

Regarding school categories, HPE and PE children presented with more difficulties in *writing words and syllables and phoneme handling and oral comprehension* rather than *reading* (i.e., PE and HPE’ barycenters were lagging behind in the lower left corner of the PCA graph, opposite from the writing and phonology arrows). These results were similar and replicated in 2019, 2020 and 2021 (see **Figure S12** in SOM).

At T3, the best students performed in all language exercises and were quite similar in skills, compared to children presenting with difficulties, that were sparser on the diagram and therefore, were more different in their difficulties profiles (see **Figure 17**). A distinct group of children struggled with *oral comprehension of words and sentences* at T3. Less children presented with difficulties in *reading* abilities. Gaps between types of school barycenters were sparser at T3 compared to T2, indicating that differences between levels for HPE, PE, regular public and private schools were larger. These level gap could find an explanation related to the summertime school break effect of a lack of schooling for 2 long months in France. This effect was identified as being predominant among lower SES score population of children, compared with higher SES score children. Note that there was a saturation of best results in the test of *oral comprehension* at T3, explaining the “roof-like” form on the right of the graph, as well as among the worst performers in *reading* level at T3, explaining the “roof-like form” on the left of the graph. These results were similar and replicated in 2019, 2020 and 2021 (see **Figure S13** in SOM).

To summarize, all previous results using both correlation matrices and PCA, led us towards similar conclusions regarding specific subgroups in language abilities: At T1, three groups emerged with a dominance of oral comprehension on one hand, a dominance of meta-phonology on the second, and a third group with a better level at decoding skills; at T2, reading abilities were dominant for a group and meta-phonology and decoding skills were on a second group; at T3, reading abilities were dominant for a group and oral comprehension belong to a second distinct one. Summarized results were presented on **Figure 18**.

Figure 18. Summarize of three PCA in language at T1, T2 and at T3 in 2018. OWord = Oral comprehension of words; OSent = Oral comprehension of sentences; OText = Oral comprehension of texts; LSound = Letter-sound association; LRecog = Letter recognition; Phon = Phoneme handling; Syll = Syllable handling; LKnow = Letter knowledge; ReadM = Reading words; ReadT = Reading texts; WWord = Writing words; WSyll = Writing syllables; Lett (T2) = Letter-sound association; ReadCS = Reading comprehension of sentences; ReadCT = Reading comprehension of Texts.



In order to affine these relationships and precisely identify the specific language assessment that would present more weight (i.e., that would be the best predictor) of the later reading comprehension and reading abilities, we implemented multilevel regression models using these language items.

C. Predicting reading and reading comprehension at T3 with multilevel regression linear models

As reading abilities at T2 and T3 comprised both reading items of words and texts in one minute, we will now define “reading abilities” as the abilities to decode words, without any confirmation of words and texts understanding. On the other hand, “decoding” comprised letter identification and letter knowledge at T1 and comprised letter recognition at T2. Finally, reading comprehension comprised both assessments of reading comprehension of sentences and of texts at T3.

In this part, we explored predictive models of (1) oral comprehension at T2, (2) phonological awareness at T2, (3) decoding at T2, (4) reading abilities at T2 with variables at T1 (see **Table 12**). These outcomes were defined as follow: The oral comprehension of sentence assessment at T2 corresponded to the oral comprehension variable at T2; phoneme awareness at T2 was used as the outcome for the second model; decoding at T2 corresponded to the letter knowledge assessment at T2; reading abilities at T2 was computed as the mean of reading words at T2 and reading texts at T2. All these variables were normalized and gaussianized.

As expected, oral comprehension abilities at T2 were most highly predicted by oral comprehension or words, sentences, and texts at T1 (respectively $\beta = 0.1710^{***}$; 0.1811^{***} ; 0.1898^{***}) and were predicted with a lesser magnitude by syllable handling at T1 ($\beta = 0.0760^{***}$). A higher SES score was associated with higher skills in oral comprehension at T2 (β SES score = 0.0648^{***}). Girls presented with an advantage compared to boys ($\beta = -0.0580^{***}$) in oral comprehension abilities at T2 (see **Table 12**).

Both phoneme awareness abilities at T2 and decoding abilities (i.e., letter recognition) at T2 were highly predicted by phoneme handling at T1, syllable handling and by letter-sound association abilities at T1, and less predicted by letter recognition at T1. Surprisingly, reading abilities at T2 (i.e., defined as the mean of reading both words and texts in 1 minute) were highly predicted by the phoneme awareness at T1 as well as the decoding abilities at T1, and were almost non predicted by the oral comprehension tasks at T1 (i.e., the coefficients related to oral comprehension items were close to zero, compared to the higher coefficients for other language domains) contrary to the latest systematic review on reading predictors which emphasized the role of word comprehension (Castles et al., 2018).

This difference could either be due to (1) the timing of the measure here 4 months after the beginning of formal reading teaching in first grade vs. by the end of second grade and older in the systematic review, (2) the lack of teaching and practicing the learning of enough vocabulary in class (i.e., too early to measure the oral comprehension of

words) (3) or to the assessments themselves (i.e., not precise enough, or needed more time to perform at the exercise). In addition, boys presented with an advantage for reading abilities at T2 compared to girls. It is the only language exercise for which boys presented with an advantage compared to girls (see **Table S18** in Chapter 4).

This advantage could be mainly due to the “competitive” aspect of these exercises, that were the only exercises which were time-limited, as shown elsewhere (Tsui & Mazzocco, 2007). Globally, girls performed better than boys at T2 only in oral comprehension, whereas boys outperformed girls in phoneme awareness, decoding and reading at T2 (see **Table 12**).

Table 12. Multilevel models for different language exercises at T2 in 2018.

	Model 1 Oral comprehension at T2, Parameter estimate (SD)	p	Model 2 Phoneme awareness at T2, Parameter estimate (SD)	p	Model 3 Decoding at T2, Parameter estimate (SD)	p	Model 4 Reading at T2, Parameter estimate (SD)	p
Intercept	0.0293 (0.0020)	< 0.0001	-0.0104 (0.0024)	< 0.0001	-0.0101 (0.0023)	< 0.0001	-0.0763 (0.0026)	< 0.0001
Age	0.0120 (0.0012)	< 0.0001	-0.0114 (0.0010)	< 0.0001	-0.0078 (0.0011)	< 0.0001	-0.0138 (0.0010)	< 0.0001
Gender (Girls < 0; Boys > 0)	-0.0580 (0.0023)	< 0.0001	0.0205 (0.0019)	< 0.0001	0.0201 (0.0021)	< 0.0001	0.1511 (0.0020)	< 0.0001
SES score	0.0648 (0.0018)	< 0.0001	-0.0103 (0.0024)	< 0.0001	-0.0059 (0.0023)	0.0094	0.0444 (0.0026)	< 0.0001
Language	-	-	-	-	-	-	-	-
Oral comprehension of words at T1	0.1710 (0.0016)	< 0.0001	0.0622 (0.0013)	< 0.0001	0.0596 (0.0014)	< 0.0001	0.0185 (0.0014)	< 0.0001
Oral comprehension of sentences at T1	0.1811 (0.0014)	< 0.0001	0.0536 (0.0012)	< 0.0001	0.0563 (0.0013)	< 0.0001	0.0097 (0.0012)	< 0.0001
Oral comprehension of texts at T1	0.1898 (0.0015)	< 0.0001	0.0158 (0.0013)	< 0.0001	0.0217 (0.0014)	< 0.0001	-0.0231 (0.0014)	< 0.0001
Phoneme handling at T1	0.0075 (0.0017)	< 0.0001	0.2559 (0.0015)	< 0.0001	0.2169 (0.0016)	< 0.0001	0.2313 (0.0015)	< 0.0001
Syllable handling at T1	0.0760 (0.0016)	< 0.0001	0.1935 (0.0013)	< 0.0001	0.1800 (0.0014)	< 0.0001	0.1347 (0.0014)	< 0.0001
Letter-sound association at T1	-0.0063 (0.0016)	0.0001	0.1672 (0.0014)	< 0.0001	0.1521 (0.0015)	< 0.0001	0.2320 (0.0014)	< 0.0001
Recognizing letter writing at T1	0.0302 (0.0014)	< 0.0001	0.0879 (0.0012)	< 0.0001	0.0924 (0.0013)	< 0.0001	0.1478 (0.0013)	< 0.0001
Comparing letters at T1	0.0244 (0.0014)	< 0.0001	0.0482 (0.0013)	< 0.0001	0.0504 (0.0013)	< 0.0001	0.0711 (0.0013)	< 0.0001
Variables per class	-	-	-	-	-	-	-	-
First of class in Language is a boy at T1	-0.0008 (0.0016)	NS (0.6123)	0.0002 (0.0022)	NS (0.9321)	-0.0002 (0.0020)	NS (0.9081)	-0.0016 (0.0024)	NS (0.5027)
Class size	-0.0099 (0.0017)	< 0.0001	-0.0279 (0.0023)	< 0.0001	-0.0244 (0.0021)	< 0.0001	-0.0380 (0.0025)	< 0.0001

Boys – Girls ratio per class	0.0023 (0.0016)	NS (0.1357)	-0.0029 (0.0021)	NS (0.1721)	-0.0009 (0.0020)	NS (0.6566)	0.0009 (0.0023)	NS (0.7111)
Heterogeneity of language at T1	0.0002 (0.0016)	NS (0.8996)	0.0119 (0.0022)	< 0.0001	0.0041 (0.0020)	0.0429	0.0278 (0.0024)	< 0.0001
Variance (intercept per class)	0.0366	-	0.1242	-	0.0973	-	0.1549	-
Residual	0.6760	-	0.4681	-	0.5483	-	0.4860	-

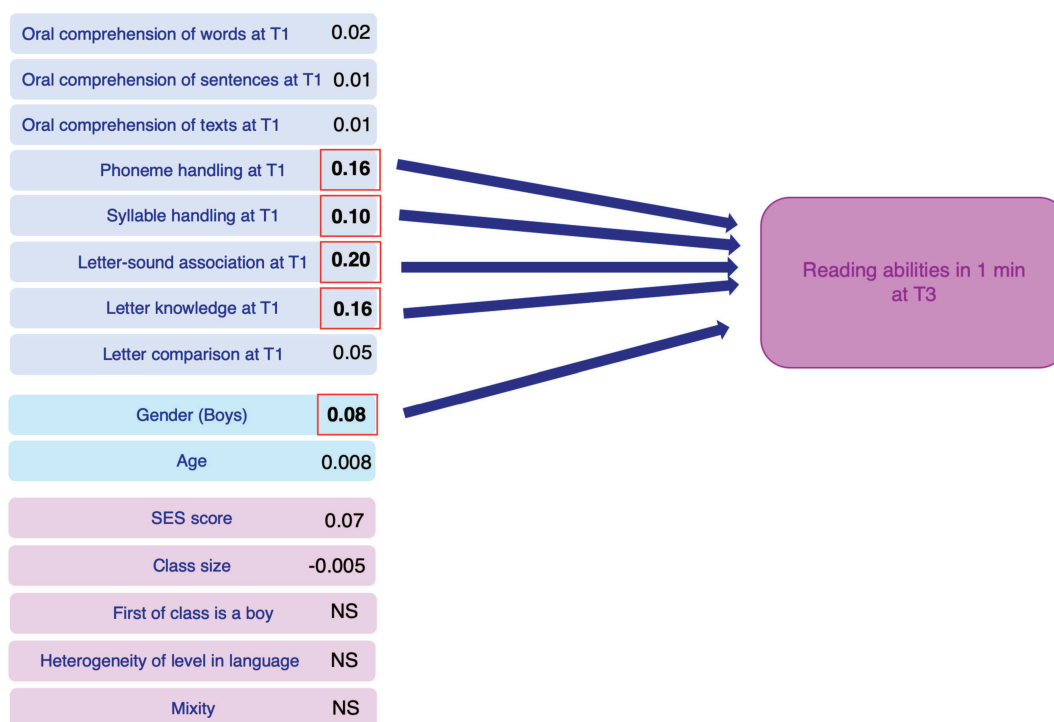
1) Reading words and texts at T3 in one minute.

As mentioned earlier, we made a difference between two outcomes: (1) reading (i.e., defined as the mean of reading words and texts at T3 in one minute) and (2) reading comprehension (i.e., defined as the mean of reading comprehension of sentences and texts at T3 – an exercise not time-limited).

Firstly, we explored the predictors for **reading** abilities at T3. The following models were established thanks to progressive models, beginning with the simplest and adding more dependent variables (see the progressive models introduced in **Table S9** in SOM).

As showed in **Table 13** and in **Figure 19**, boys performed better than girls in reading assessments at T3. This language domain was the only one were boys performed better than girls, both at T2 and T3 probably due to the time-limited competitive aspect of these tests as explained above (see **Table S18** in Chapter 4).

Figure 19. Predictors for reading words and texts in one minute in second grade.



In addition, the higher the SES score was, the higher the results in reading were. A larger class size tended to reduce the chance of having a good level of reading at T3, despite the fact that smaller classes were mainly in priority education schools as these multilevel models estimated the effect magnitude when all other parameters equal to zero. Finally, the best predictors for reading abilities at T3 were letter-sound association, phoneme handling, letter knowledge and syllable handling at T1. Neither the first of class being a boy, nor the boys-girls ratio per class, nor the heterogeneity of level in language in the class had any influence on the reading level at T3. These results were similar in magnitude between 2018, 2019, 2020 and 2021, only the SES score had a higher magnitude in 2019, which could be explained by the Covid-19 year where children attended less days of school overall, therefore, relying more on families for instruction and enlarging SES score gaps in reading levels (see **Table 13** and see cohorts' comparisons below).

Table 13. Multilevel regression model testing “reading at T3” as the main outcome among children of typical age.

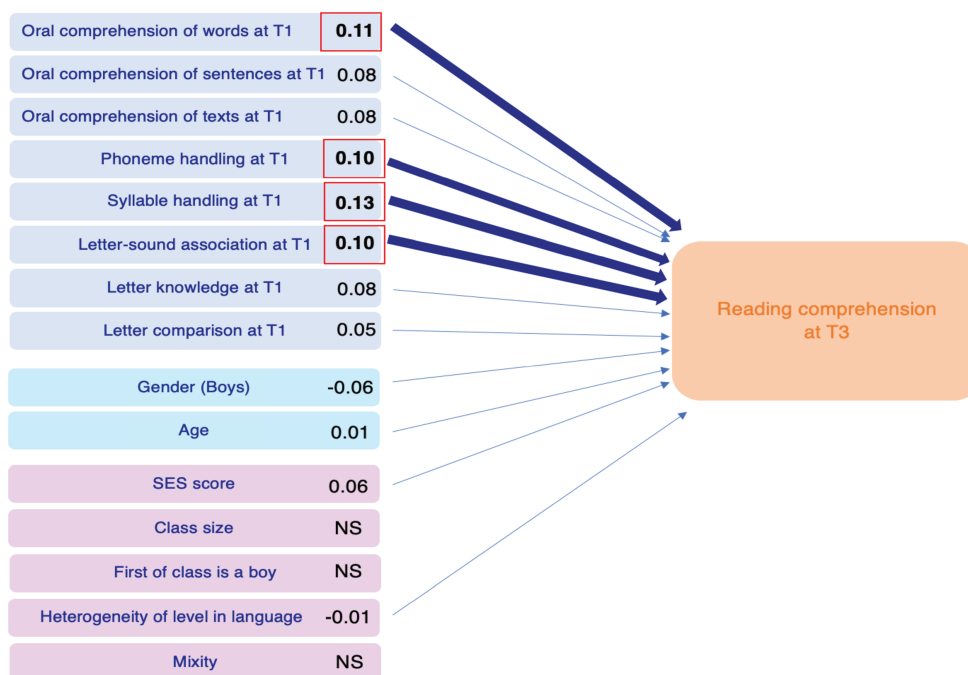
Variables	Reading at T3							
	2018		2019		2020		2021	
N	569,771		665,632		695,449		722,230	
N group (classes)	39,573		46,671		49,010		49,701	
Fixed effects	Parameter estimates (sd)	p	Parameter estimates (sd)	p	Parameter estimates (sd)	p	Parameter estimates (sd)	p
<i>Intercept</i>	0.0420 (0.0023)	< 0.0001	0.0389 (0.0020)	< 0.0001	0.0413 (0.0019)	< 0.0001	0.0484 (0.0018)	< 0.0001
Age at T1 (month)	0.0074 (0.0012)	< 0.0001	0.0085 (0.0011)	< 0.0001	0.0033 (0.0010)	0.0014	0.0072 (0.0010)	< 0.0001
Gender (Boys)	0.0862 (0.0021)	< 0.0001	0.0802 (0.0019)	< 0.0001	0.0829 (0.0019)	< 0.0001	0.0975 (0.0018)	< 0.0001
Oral comprehension of words at T1	0.0324 (0.0014)	< 0.0001	0.0403 (0.0013)	< 0.0001	0.0235 (0.0013)	< 0.0001	0.0074 (0.0013)	< 0.0001
Oral comprehension of sentences at T1	0.0160 (0.0013)	< 0.0001	0.0172 (0.0012)	< 0.0001	0.0176 (0.0012)	< 0.0001	0.0079 (0.0011)	< 0.0001
Oral comprehension of texts at T1	0.0118 (0.0014)	< 0.0001	0.0156 (0.0012)	< 0.0001	0.0077 (0.0012)	< 0.0001	-0.0041 (0.0012)	0.0006
Phoneme handling at T1	0.1589 (0.0016)	< 0.0001	0.1596 (0.0014)	< 0.0001	0.1545 (0.0014)	< 0.0001	0.1466 (0.0013)	< 0.0001
Syllable handling at T1	0.1194 (0.0014)	< 0.0001	0.0909 (0.0013)	< 0.0001	0.0810 (0.0013)	< 0.0001	0.0694 (0.0013)	< 0.0001
Letter-sound association at T1	0.1783 (0.0014)	< 0.0001	0.2071 (0.0012)	< 0.0001	0.2113 (0.0013)	< 0.0001	0.1644 (0.0012)	< 0.0001
Letter knowledge at T1	0.1543 (0.0013)	< 0.0001	0.1630 (0.0012)	< 0.0001	0.1699 (0.0012)	< 0.0001	0.1399 (0.0011)	< 0.0001
Letter comparison at T1	0.0820 (0.0013)	< 0.0001	0.0563 (0.0011)	< 0.0001	0.0595 (0.0011)	< 0.0001	0.0430 (0.0010)	< 0.0001

First of class is a boy in language	-0.0018 (0.0020)	NS (0.3758)	0.0001 (0.0017)	NS (0.9703)	0.0030 (0.0017)	NS (0.0778)	0.0024 (0.0015)	NS (0.1151)
SES score	0.0701 (0.0021)	< 0.0001	0.1160 (0.0018)	< 0.0001	0.0692 (0.0018)	< 0.0001	0.0593 (0.0016)	< 0.0001
Class size	-0.0071 (0.0021)	0.0007	-0.0041 (0.0018)	0.0225	-0.0050 (0.0018)	0.0059	0.0045 (0.0016)	0.0043
Boys-Girls ratio per class	-0.0010 (0.0019)	NS (0.6040)	-0.0032 (0.0016)	0.0496	-0.0084 (0.0016)	< 0.0001	-0.0016 (0.0015)	NS (0.2802)
Heterogeneity of language level at T1	-0.0091 (0.0019)	< 0.0001	-0.0012 (0.0016)	NS (0.4645)	-0.0028 (0.0016)	NS (0.0838)	-0.0005 (0.0014)	NS (0.7137)
Random effects								
Between-class variance (Level 2)								
Intercept variance	0.1046		0.08579		0.08903		0.06702	
Within-class variance (Level 1)	0.5678		0.55260		0.56373		0.55268	
Deviance (-2 log L)	1344007.7		1546904		1630746		1670212.3	

2) Reading comprehension abilities at T3

As mentioned above, reading comprehension at T3 was composed of the mean of reading comprehension of sentences at T3 and reading comprehension of texts at T3 which were both normalized and gaussianized, all results were presented in the following **Table 14** and in **Figure 20**.

Figure 20. Predictors of reading comprehension in second grade.



The model was established thanks to a progressive model, beginning with the simplest and adding more dependent variables (see the progressive models introduced in **Table S10** in SOM). In terms of reading comprehension abilities at T3, girls outperformed boys. A higher SES score was positively linked to higher reading comprehension results at T3. However, class size did not show any significant association with reading comprehension levels at T3. When considering all assessments at T1, along with explanatory variables such as age and class size, several key predictors of reading comprehension at T3 emerged. Among these, oral comprehension of words at T1 had the highest predictive weight for reading comprehension at T3, overpassing oral comprehension of sentences and texts.

Additionally, with similar magnitudes, phoneme handling, syllable handling, and letter-sound association at T1 played pivotal roles in predicting reading comprehension outcomes (details were presented in **Table 14**). Notably, factors such as being the first boy in the class or the boys-girls ratio did not exhibit any association with reading comprehension. However, in contrast to reading abilities, a greater variability in language proficiency levels at T1 within a class was associated with poorer reading comprehension outcomes. In simpler terms, children in classes with a wide range of language levels at T1 tended to perform less successfully in reading comprehension at T3. These findings remained consistent across multiple years, as indicated in the corresponding table for 2019, 2020, and 2021 and found the similar higher predictive weight for SES score in 2019.

Combining the subgroups identified thanks to correlations and PCAs, we defined oral comprehension at T1 as the mean of oral comprehension of words, sentences and texts at T1; meta phonology at T1 as the mean of phoneme handling, syllable handling and letter-sound association at T1, and visual decoding at T1 as the mean of letter knowledge and letter comparison at T1. The following models implemented these three latter variables in 2018, 2019, 2020 and 2021. Meta phonology at T1 was the most predictive domain for later reading comprehension at T3, followed by oral comprehension at T1, and then, decoding at T1. These results were replicated every year (see **Table 15** and in **Figure 21**).

 ***Tips for results interpretation of multilevel models***

Different questions that multilevel models helped us solve, and how to interpret multilevel model parameters

Model significance: The model significance is not measured by the “p” but rather by its decrease of deviance all along the different progressive models (see below).

Model function of variance: Multilevel models do not present with any unique R^2 because here, an R^2 is calculated for each level (N.B., a R^2 for a model usually represents the percentage of variance that the model explains). The multilevel model presents a function of variance, where each element of the model has a variance associated to it.

Table 14. Multilevel regression model testing “reading comprehension at T3” as the main outcome and all the T1 assessments.

Variables	Reading comprehension at T3							
	2018		2019		2020		2021	
N	569,771		665,632		695,449		722,230	
N group (classes)	39,573		46,671		49,010		49,701	
Fixed effects	Parameter estimates (sd)	p	Parameter estimates (sd)	p	Parameter estimates (sd)	p	Parameter estimates (sd)	p
<i>Intercept</i>	-0.0296 (0.0017)	< 0.0001	-0.0288 (0.0015)	< 0.0001	-0.0278 (0.0015)	< 0.0001	-0.0209 (0.0015)	< 0.0001
Age at T1 (month)	0.0112 (0.0010)	< 0.0001	0.0078 (0.0009)	< 0.0001	0.0170 (0.0009)	< 0.0001	0.0191 (0.0010)	< 0.0001
Gender (Boys)	-0.0600 (0.0018)	< 0.0001	-0.0582 (0.0016)	< 0.0001	-0.0576 (0.0016)	< 0.0001	-0.0447 (0.0017)	< 0.0001
Oral comprehension of words at T1	0.1112 (0.0012)	< 0.0001	0.1115 (0.0011)	< 0.0001	0.1077 (0.0011)	< 0.0001	0.0915 (0.0012)	< 0.0001
Oral comprehension of sentences at T1	0.0796 (0.0011)	< 0.0001	0.0798 (0.0010)	< 0.0001	0.0882 (0.0010)	< 0.0001	0.0647 (0.0011)	< 0.0001
Oral comprehension of texts at T1	0.1092 (0.0012)	< 0.0001	0.0859 (0.0010)	< 0.0001	0.0837 (0.0010)	< 0.0001	0.0719 (0.0011)	< 0.0001
Phoneme handling at T1	0.0951 (0.0013)	< 0.0001	0.1131 (0.0012)	< 0.0001	0.1065 (0.0012)	< 0.0001	0.1265 (0.0012)	< 0.0001
Syllable handling at T1	0.1282 (0.0012)	< 0.0001	0.1158 (0.0011)	< 0.0001	0.1087 (0.0011)	< 0.0001	0.0968 (0.0012)	< 0.0001
Letter-sound association at T1	0.0962 (0.0012)	< 0.0001	0.1221 (0.0011)	< 0.0001	0.1155 (0.0011)	< 0.0001	0.0860 (0.0011)	< 0.0001
Letter knowledge at T1	0.0880 (0.0011)	< 0.0001	0.0874 (0.0010)	< 0.0001	0.0865 (0.0010)	< 0.0001	0.0574 (0.0010)	< 0.0001
Letter comparison at T1	0.0508 (0.0011)	< 0.0001	0.0530 (0.0009)	< 0.0001	0.0514 (0.0009)	< 0.0001	0.0270 (0.0009)	< 0.0001
First of class is a boy in language	-0.0011 (0.0015)	NS (0.4429)	0.0003 (0.0013)	NS (0.7960)	0.0002 (0.0012)	NS (0.8859)	0.0007 (0.0013)	NS (0.5795)
SES score	0.0686 (0.0016)	< 0.0001	0.1032 (0.0014)	< 0.0001	0.0679 (0.0013)	< 0.0001	0.0819 (0.0013)	< 0.0001
Class size	-0.0000 (0.0015)	NS (0.9785)	0.0004 (0.0014)	NS (0.7792)	-0.0022 (0.0013)	NS (0.0925)	0.0018 (0.0013)	NS (0.1682)

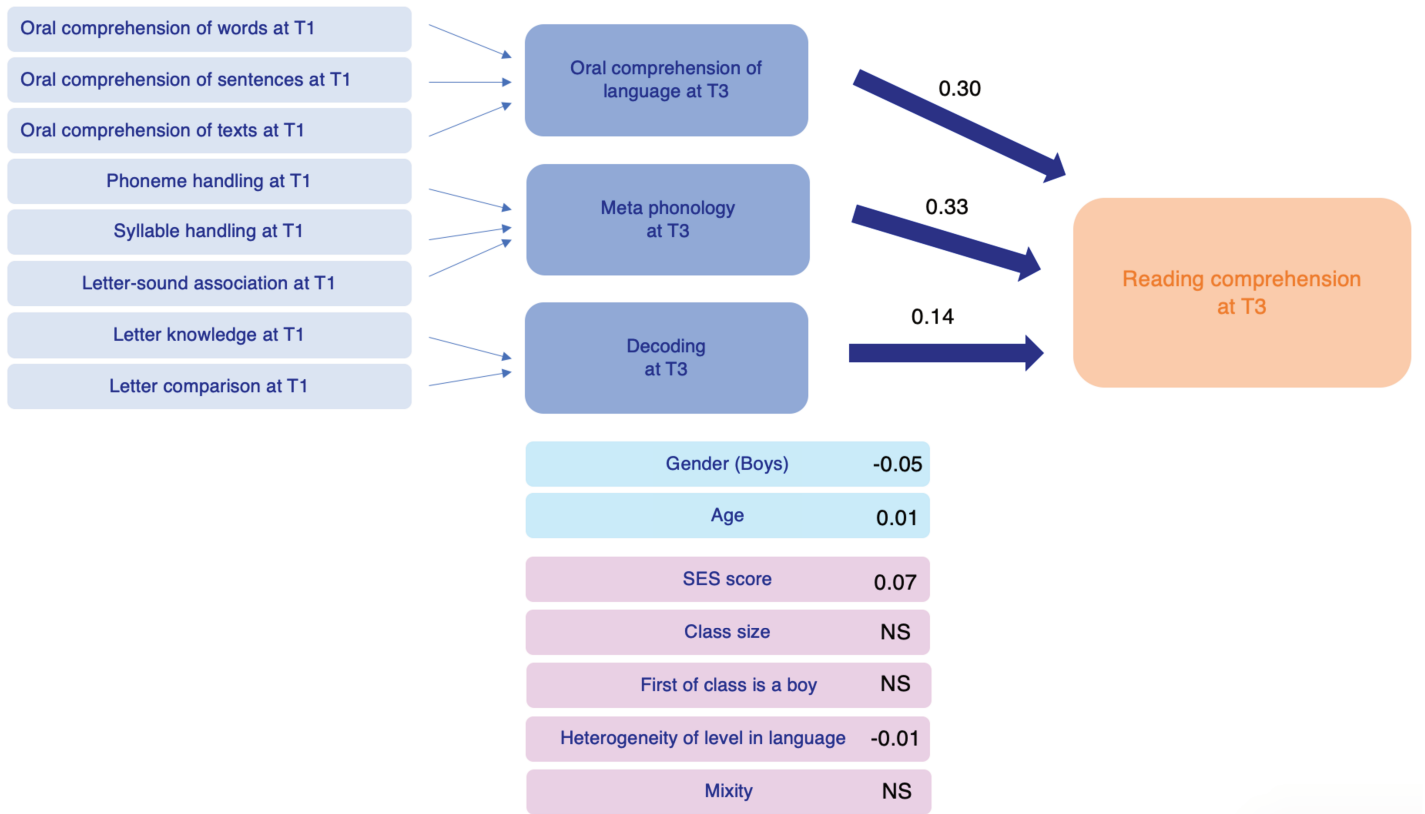
Boys-Girls ratio per class	0.0014 (0.0014)	NS (0.3271)	-0.0011 (0.0013)	NS (0.3860)	-0.0043 (0.0012)	0.0004	0.0002 (0.0012)	NS (0.8973)
Heterogeneity of language level at T1	-0.0178 (0.0014)	< 0.0001	-0.0131 (0.0012)	< 0.0001	-0.0144 (0.0012)	< 0.0001	-0.0001 (0.0012)	NS (0.9510)
Random effects								
Between-class variance (Level 2)								
Intercept variance	0.0482		0.0422		0.03922		0.04081	
Within-class variance (Level 1)	0.4099		0.4098		0.40487		0.48056	
Deviance (-2 log L)	1146740.8		1336002.7		1385873.8		1559082.2	

Table 15. Multilevel regression model assessing Reading comprehension at T3 with composite variables at T1.

Variables	Reading Comprehension at T3							
	2018		2019		2020		2021	
N	569,771		665,632		695,449		722,230	
N group (classes)	39,573		46,671		49,010		49,701	
Fixed effects	Parameter estimates (sd)	p	Parameter estimates (sd)	p	Parameter estimates (sd)	p	Parameter estimates (sd)	p
Intercept	-0.0278 (0.0017)	< 0.0001	-0.0270 (0.0015)	< 0.0001	-0.0264 (0.0015)	< 0.0001	-0.0194 (0.0015)	< 0.0001
Age at T1 (month)	0.0116 (0.0010)	< 0.0001	0.0084 (0.0009)	< 0.0001	0.0173 (0.0009)	< 0.0001	0.0197 (0.0010)	< 0.0001
Gender (Boys)	-0.0563 (0.0018)	< 0.0001	-0.0548 (0.0016)	< 0.0001	-0.0551 (0.0016)	< 0.0001	-0.0417 (0.0017)	< 0.0001
Oral comprehension at T1	0.3038 (0.0015)	< 0.0001	0.2749 (0.0014)	< 0.0001	0.2774 (0.0014)	< 0.0001	0.2272 (0.0014)	< 0.0001
Meta phonology at T1	0.3237 (0.0015)	< 0.0001	0.3581 (0.0014)	< 0.0001	0.3377 (0.0014)	< 0.0001	0.3163 (0.0015)	< 0.0001
Decoding at T1	0.1368 (0.0014)	< 0.0001	0.1379 (0.0013)	< 0.0001	0.1355 (0.0012)	< 0.0001	0.0797 (0.0013)	< 0.0001
First of class is a boy in language	-0.0012 (0.0015)	NS (0.4120)	0.0003 (0.0013)	NS (0.7923)	0.0002 (0.0012)	NS (0.8933)	0.0006 (0.0013)	NS (0.6155)
SES score	0.0695 (0.0016)	< 0.0001	0.1042 (0.0014)	< 0.0001	0.0687 (0.0013)	< 0.0001	0.0836 (0.0013)	< 0.0001
Class size	0.0004 (0.0015)	NS (0.7726)	-0.0003 (0.0014)	NS (0.8339)	-0.0027 (0.0013)	0.0439	0.0006 (0.0013)	NS (0.6255)
Boys-Girls ratio per class	0.0015 (0.0014)	NS (0.3000)	-0.0010 (0.0013)	NS (0.4132)	-0.0042 (0.0012)	0.0004	0.0001 (0.0012)	NS (0.9287)
Heterogeneity of language at T1	-0.0194 (0.0014)	< 0.0001	-0.0125 (0.0012)	< 0.0001	-0.0140 (0.0012)	< 0.0001	0.0012 (0.0012)	NS (0.3360)
Random effects								
Between-class variance (Level 2)								
Intercept variance	0.0493		0.0420		0.0392		0.03987	
Within-class variance (Level 1)	0.4105		0.4106		0.4054		0.48178	
Deviance (-2 log L)	1148093.7		1337082.3		1386813.0		1560238.6	

Overall, meta-phonology at T1 predicted reading comprehension with the highest weight, followed by oral comprehension at T1 and way behind, visual decoding at T1 with a lower predictive weight, as presented in **Table 15** and in **Figure 21**.

Figure 21. Predictors of reading comprehension in second grade.



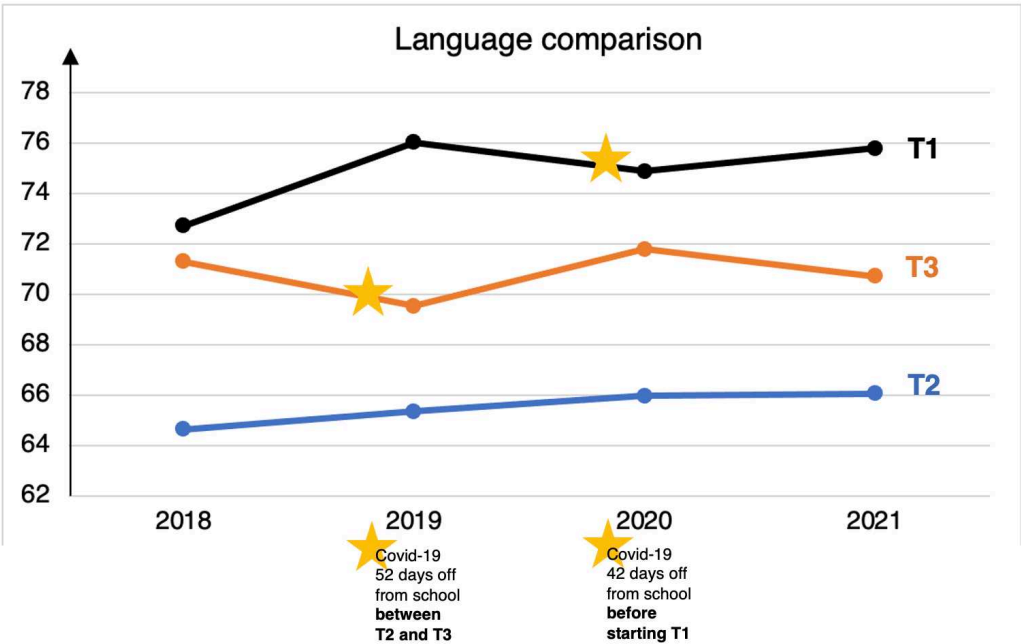
D. Comparison between cohorts and estimation of the COVID-19 impact on reading acquisition and estimation of progress.

Covid-19 pandemic justified political measures of school closure and explained a reduced school exposure during 52 consecutive days (between T2 and T3 for the children of 2019) and of 42 consecutive days during kindergarten for the cohort who started first grade in 2020, both happening before the 2 months summertime break. The year 2020 was also disrupted, though to a lesser extent (i.e., schools were closed for 15 days, between T2 and T3). Therefore, this unintentional natural experiment allowed to see the effect of schooling on learning how to read.

Firstly, we compared all years (2018, 2019, 2020, 2021) regarding their language levels, using normalized variables in percentage of success ranging from 0 to 100 (see **Figure 21**). As explained in chapter 2, since the tests vary across T1 to T3, comparing the overall language proficiency levels between T1, T2, and T3 within each of the four cohorts is not feasible (e.g., comparing T1 and T2 levels in language in 2018), whereas we can compare direct year-to-year for the same period (e.g., we can compare language levels at T2 between 2018, 2019, 2020 and 2021).

Overall, from 2018 to 2021, we noticed children presented with an improvement of their language level at T1, except for the year 2020 (i.e., cohorts of children that were exempt from school for 42 days by the end of 2019, before beginning first grade). In addition, we noted a progress in language level at T2 for every cohort from 2018 to 2021. Finally, we noted a drop in language level at T3 in 2019 compared to 2018, followed by a catch up in level at 2020 and another drop in 2021 at T3. The lesser exposure to school in 2019 could explained the drop of level in language at T3 in 2019 compared to the other years. However, it cannot explain the more recent drop in level at T3 in 2021. All differences were significant and tested with ANOVA presented in SOM (see **Table S11**).

Figure 21. Levels in language at T1, T2, T3 in 2018, 2019, 2020 and 2021.



For instance, we must keep in mind that the concept of progress was associated with biases here for three main reasons: (1) As, even when measuring the same cognitive domain (e.g., phoneme awareness), most tests changed from T1 to T3, therefore making it impossible to compare T1 to T2 to T3; (2) as the best students in Evalaide had excellent results in most tests they had a lesser potential for progression compared to lower performers, and (3) as some tests were saturated, we were not able to measure progress for the children with good levels. Therefore, we compared year-to-year identical exercises but could not measure a proper “progress slope” for every child in Evalaide (i.e., through T1 to T3).

However, regarding reading abilities and reading comprehension at T3, we were able to compare results in between cohorts with the aim at identifying if children progressed similarly or not regarding the exposure of school they received (see **Figure 22 – A** and **Figure 23 - A**).

Both reading words and texts at T3 and reading comprehension of sentences and of texts at T3 presented a drop in level in 2019 compared to the years 2018, 2020 and 2021 (**Figure 22-A** and **Figure 23-A**). Especially, for reading words and texts, this drop in level was not noticed at T2 but noticed at T3 (**Figure 22-A**).

In addition, we were able to compare, within the cohorts 2018, 2019, 2020 and 2021, how children performed when belonging to a type of school (i.e., private vs. regular public vs. PE vs. HPE) (see **Figure 22 – B** and **Figure 23-B**). Results obtained per test among school categories were shown in detail in **Table S4** in chapter 2. Overall, in 2018, either in reading abilities at T2 and T3 or in reading comprehension at T3, children presented with a higher level when going to private and regular public schools with median to high SES scores, while other children of regular public schools with lower SES scores or going to PE and HPE schools presented a lower level (see **Figure 22 – B** and see **Figure 23-B**). More particularly, a larger drop in level was noted in 2019 for PE and HPE schools compared to private and regular public schools. All differences were significant and tested with ANOVA presented in SOM (see **Table S11** and **Table S12**).

Figure 22. Comparing Reading abilities for words and texts at T2 and at T3 (A) in 2018, 2019, 2020 and 2021 and (B) between school categories in 2018, 2019, 2020 and 2021.

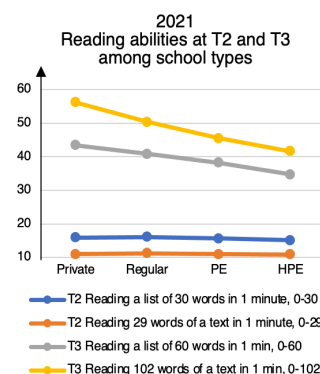
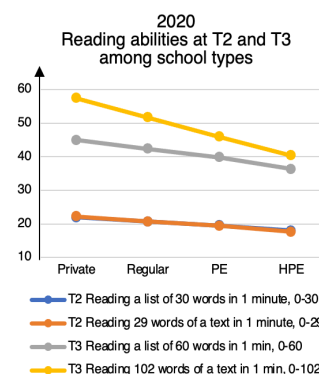
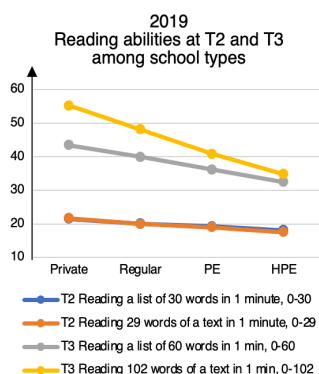
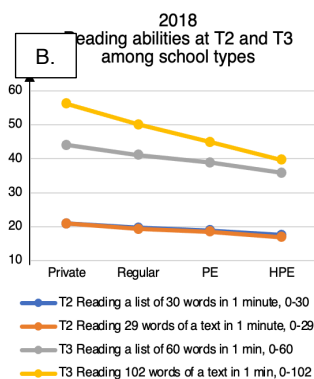
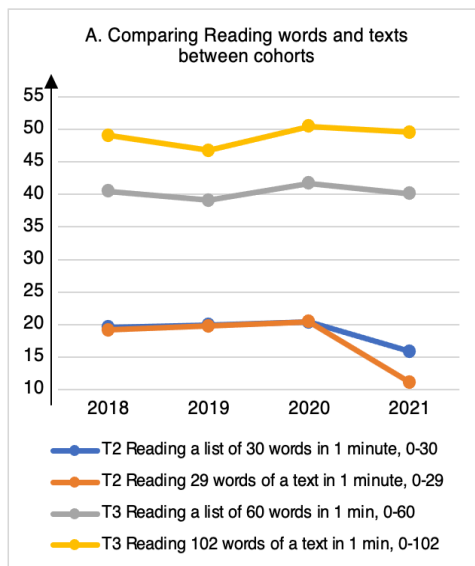
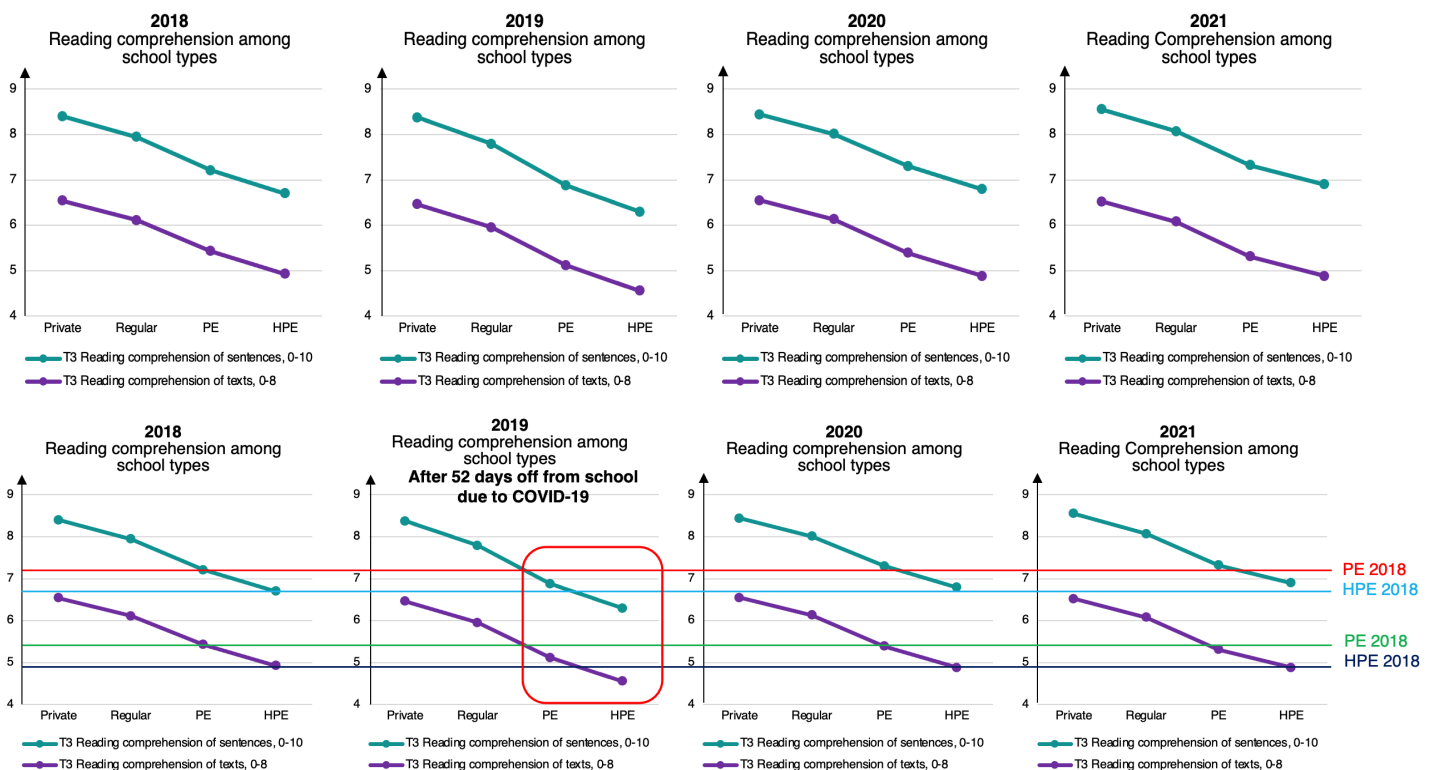
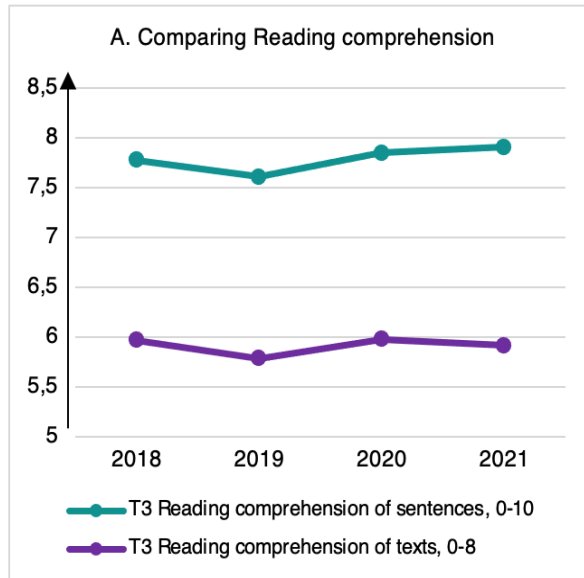
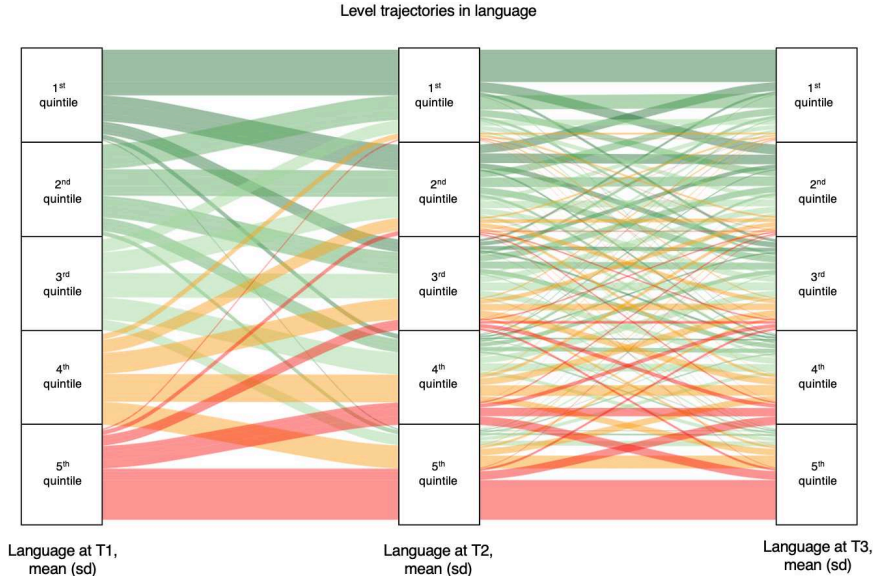


Figure 23. Comparing Reading comprehension of sentences and of texts at T3 (A) in 2018, 2019, 2020 and 2021 and (B) between school categories in 2018, 2019, 2020 and 2021, where the same figure was represented twice, without any indications (up figure) and with four different horizontal lines representing the level of reading comprehension among PE and HPE schools for reading comprehension of sentences and of texts (down figure).



Finally, to have a better idea of the language trajectories and affine our focus of identification of children with more difficulties in language domains, we implemented alluvial analyses, where language was the normalized variable represented in percentage of success (ranging from 0 to 100) and was defined, for each period, as the mean of all assessments belonging to this period (T1, T2 or T3). Colors represented children' quintiles of level at T1: The green population represented children with a better level in language, compared to the orange category (4th quintile of level in language) and the red category represented children with the most difficulties in language at T1. We noted that a large part of children belonging to the lowest level quintile remained reached the lowest quintile of level at T2 and remained among the lowest quintile of level at T3 (see **Figure 24**).

Figure 24. Level trajectories in language in 2018. Language was the mean of all assessments of a period of time, all assessments were normalized.



E. Children presenting with difficulties in Reading comprehension at T3

As we aimed at understanding which specificities children present when belonging to the lower quintile in reading comprehension at T3, we focused the following subchapter on this population and analyzing only the typical-age-in-first-grade population of first

and second graders (except for the very first analysis on age categories that was on the general population with all ages).

1) Who are the children who struggled in Reading comprehension at T3 (lowest quintile)

When we examined the age in T1 of the children who were in the lowest percentile for reading comprehension one year later in T3, there was a linear relationship between age and their performance at T3 (see **Figure 25**).

59.73% of children who entered a year later in first grade (i.e., the year of their 7th or 8th birthday) belong to the latest quintile of level in reading comprehension at T3 (see **Table 16** and **Table 17**). When comparing the four school types, the proportion of children with higher difficulties in reading comprehension at T3 was higher among late-in-age children of private schools compared to regular, PE and HPE schools (i.e., 7.7% vs. 5.2 to 7.2 %, see **Table 17**). Children with a year in advance were largely ahead in level in reading comprehension compared with typical age children. In addition, late-in-age children were lagging way behind in reading comprehension level compared to typical age children.

Figure 25. Reading comprehension at T3 in function of age in month

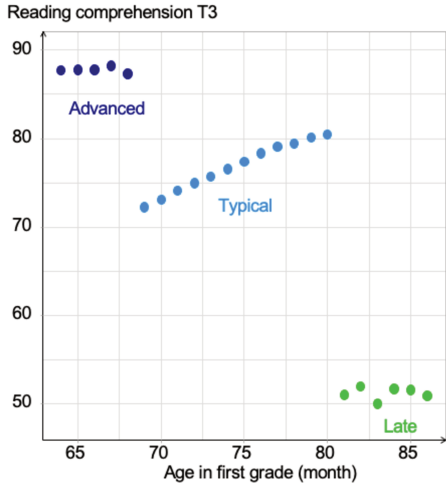


Table 16. Age characteristics of children belonging to the latest quintile vs. the rest, and vs. the best quintile in reading comprehension at T3 in 2018.

	Latest quintile 20% with most difficulties in reading comprehension at T3	Other 4 quintiles 80% other students in reading comprehension at T3	Best quintile 20% most advanced in reading comprehension at T3
N	120971	459159	72309
Age at T1, month (mean (SD))	74.54 (4.42)	74.66 (3.67)	75.04 (3.53)
Age in categories,			
1 year in advance (n, (% per column))	215 (0.2)	3251 (0.7)	846 (1.2)
Typical age (n, (%))	112836 (93.3)	450569 (98.1)	71293 (98.6)
1 year late (n, (%))	7920 (6.5)	5339 (1.2)	170 (0.2)
Class size (mean (SD))	16.75 (5.64)	17.43 (5.84)	17.69 (5.92)
SES (mean (SD))	95.16 (18.02)	104.32 (17.23)	108.68 (16.26)
Gender - Boys (n, (% per column))	69369 (57.3)	225499 (49.1)	30762 (42.5)

Table 17. Description of children belonging to the worst reading comprehension quintile at T3 regarding their age categories and their school types

	Latest quintile 20% with most difficulties in reading comprehension at T3 N = 120971			
	Private schools	Regular public schools	PE public schools	HPE public schools
n	7221	81296	17655	14799
Age, mean (SD)	74.57 (4.58)	74.50 (4.41)	74.70 (4.55)	74.57 (4.28)
Age categories				
Age – 1 year in advance (n, (% per column))	28 (0.4)	99 (0.1)	48 (0.3)	40 (0.3)
Age – Typical (n, (%))	6634 (91.9)	75886 (93.3)	16333 (92.5)	13983 (94.5)
Age – 1 year late (n, (%))	559 (7.7)	5311 (6.5)	1274 (7.2)	776 (5.2)
Part of children in school type X, belonging to				

Age – 1 year in advance (% per line), n = 215	13.02	46.04	22.33	18.60
Age – Typical (%), n = 112836	5.88	67.25	14.47	12.39
Age – 1 year late (%), n = 7920	7.06	67.06	16.09	9.80
Class size (mean (SD))	18.37 (6.14)	18.51 (5.29)	12.00 (2.93)	11.95 (3.08)
SES (mean (SD))	110.43 (13.94)	101.72 (13.99)	79.46 (10.94)	70.37 (8.98)
Gender - Boys (n, (% per column))	4411 (61.1)	46937 (57.7)	9969 (56.5)	8052 (54.4)

As found in the previous chapter, great disparities of level existed in reading comprehension at T3 between the four school types, with a large advance in level for private schools and regular public schools (i.e., either above the median or above the mean, only a few students belong to the lowest quintile), while a great delay for both PE and HPE schools (i.e., a majority of students were below the median and a large part of them belong to the lowest quintile, see **Figure 26**).

Children going to private school had 4 times more chances to belong to the best quintile in reading comprehension at T3 compared to children going to HPE schools and 40.03% of children going to HPE schools belong to the last quintile of level in reading comprehension at T3 (vs. 13.13% for private schools) (see **Table 18**).

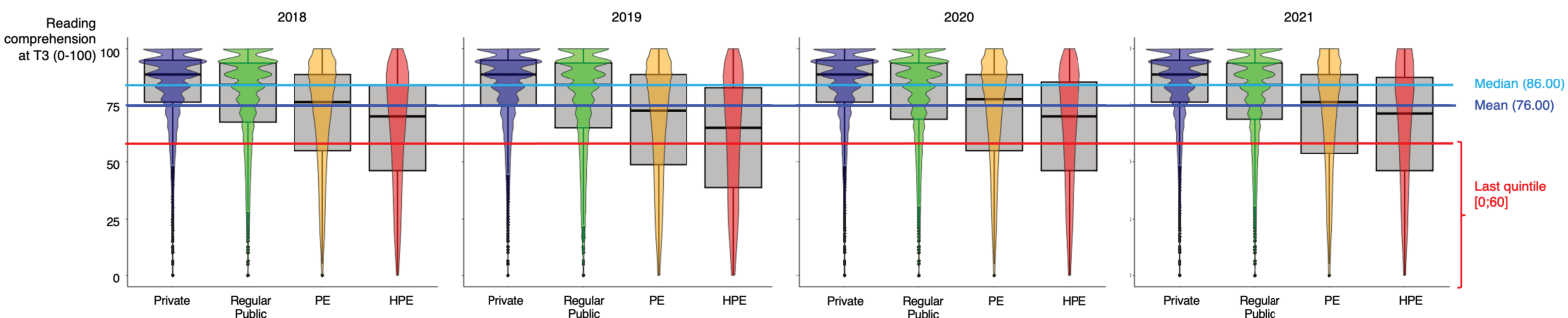
Table 18. Describing children belonging to the latest quintile in reading comprehension at T3 with other 4 quintiles and with the best quintile, among typical-age children.

	Latest quintile 20% with most difficulties in reading comprehension at T3	Other 4 quintiles 80% other students in reading comprehension at T3	Best quintile 20% most advanced in reading comprehension at T3
n	112836	379276	71293
School categories			
Private (n, (% per column))	6634 (5.9)	43896 (11.6)	10111 (14.2)
Regular public (n, (%))	75886 (67.3)	279426 (73.7)	55285 (77.5)
PE public (n, (%))	16333 (14.5)	35007 (9.2)	4150 (5.8)
HPE public (n, (%))	13983 (12.4)	20947 (5.5)	1747 (2.5)
Part of children in private schools (n = 50530) which belong to X, (% per line)	13.13	86.87	20.00

Part of children in regular public schools (n = 355312) which belong to X, %	21.36	78.64	15.56
Part of children in PE public schools (n = 51340) which belong to X, %	31.81	68.19	8.08
Part of children in HPE public schools (n = 34930) which belong to X, %	40.03	59.97	5.00

In addition, when comparing the four cohorts, we noted a more important part of children belonging to PE and HPE were under the mean and belong to the lowest quintile in reading comprehension at T3 in 2019 compared to the other three cohorts, indicating that PE and HPE children were more sensitive to the lack of school during Covid-19 in 2019 regarding their reading comprehension performances.

Figure 26. Reading comprehension average scores at T3 (in percentage of success) per school categories and SES score.



Among children struggling with reading comprehension at T3 (i.e., belonging to the worst quintile), children going to HPE public schools presented with worst results for each language test at T1, T2 and T3 compared to children going to PE public schools, especially in both reading words and texts domains at T2 and T3 – but also in math at T1, T2 and T3, even when presenting with similar ages, gender ratio and class size (see **Table 19**).

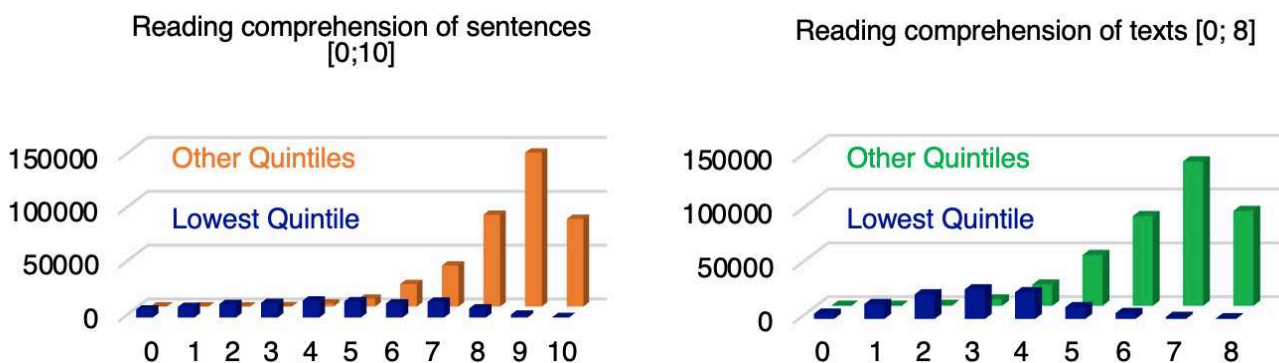
Table 19. Description of difficulties presented by children of PE and HPE public schools that belong to the worst quintile in reading comprehension at T3 in 2018.

	Students belonging to the worst reading comprehension quintile at T3 and going to PE public schools	Students belonging to the worst reading comprehension quintile at T3 and going to HPE public schools	p
n	16333	13983	
Age at T1, month (mean (SD))	73.90 (3.41)	73.99 (3.38)	< 0.0001
Class size (mean (SD))	12.00 (2.95)	11.93 (3.08)	< 0.0001
SES (mean (SD))	79.46 (11.00)	70.30 (8.97)	< 0.0001
Gender - Boys (n, (% per column))	9130 (55.9)	7571 (54.1)	< 0.0001
T1 Oral Comprehension of Words, 0-15 (mean (SD))	8.83 (3.05)	8.15 (3.08)	< 0.0001
T1 Oral Comprehension of Sentences, 0-14 (mean (SD))	9.96 (2.96)	9.45 (3.18)	< 0.0001
T1 Oral Comprehension of Texts, 0-18 (mean (SD))	9.49 (3.89)	8.98 (3.97)	< 0.0001
T1 Phoneme handling, 0-15 (mean (SD))	5.80 (3.28)	5.67 (3.33)	< 0.0001
T1 Syllable handling, 0-15 (mean (SD))	8.49 (3.35)	8.26 (3.38)	< 0.0001
T1 Letter-sound association, 0-10 (mean (SD))	5.37 (2.80)	5.32 (2.87)	< 0.0001
T1 Decoding, letter writings recognition, 0-7 (mean (SD))	3.29 (2.05)	3.16 (2.12)	< 0.0001
T1 Comparing letters, visuo-attentional abilities, 0-24 (mean (SD))	11.32 (6.77)	11.17 (6.89)	< 0.0001
T2 Number of words read out of a list of 30 words in 1 minute, range 0-100 (mean (SD))	13.94 (9.84)	13.04 (9.41)	< 0.0001
T2 Number of words read out of a list of 29 words of a text in 1 minute, 0-195 (mean (SD))	14.03 (12.51)	12.83 (11.45)	< 0.0001
T3 Number of words read out of a list of 60 words in 1 min at T3, 0-93 (mean (SD))	25.22 (15.21)	23.57 (15.32)	< 0.0001
T3 Number of words read out of a list of 102 words in a text in 1 min at T3, 0-136 (mean (SD))	21.60 (19.28)	19.84 (18.95)	< 0.0001
T3 Understanding reading a sentence, 0-10 (mean (SD))	5.18 (2.44)	5.05 (2.48)	< 0.0001
T3 Understanding reading a text, 0-8 (mean (SD))	3.93 (1.55)	3.79 (1.53)	< 0.0001
T3 Reading comprehension, 0-100, mean (SD)	39.24 (15.17)	37.69 (15.65)	< 0.0001

2) What are the characteristics and learning patterns specificities of children belonging to the lowest quintile in reading comprehension?

Among children with highest difficulties in reading comprehension at T3 (i.e., mean of reading comprehension of sentences and reading comprehension of texts at T3), more than half of the children (52.15%, n = 15,747 children) understood 4 or less sentences among the 10 sentences presented to them for reading comprehension at T3, compared to more than half of the children that understood 8 to 9 sentences at T3 among children of the four other quintiles of level in reading comprehension at T3 (see **Table S13 to Table S17** and **Figure 27**).

Figure 27. Distribution of raw scores in reading comprehension of sentences and texts at T3 in 2018 for all typical-age children, comparing children of the lowest quintile in reading comprehension at T3 vs. the others.



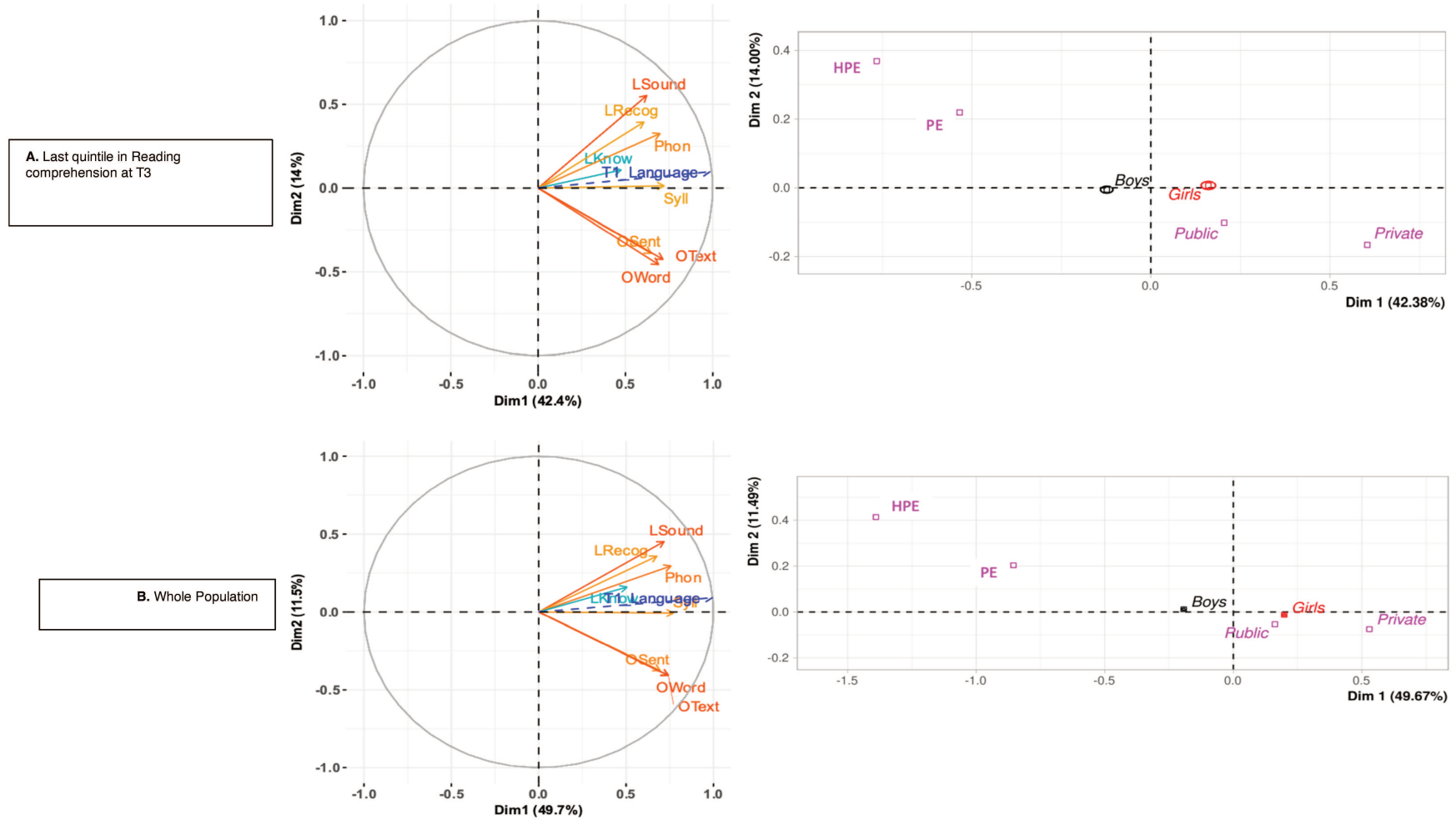
Among children with highest difficulties in reading comprehension at T3, more than half of the children (between 36.86 to 61.18, n = 50,287 children) understood 2 to 3 or less texts among the 8 texts presented to them for reading comprehension at T3, compared to more than 41.39 to 76.77% of children that understood 6 to 7 texts out of 8 texts at T3 among children of the four other quintiles of level in reading comprehension at T3.

As we wanted to explore if any patterns of learning in language were existing among children with difficulties in reading comprehension at T3, we decided to compare the PCA of language assessments at T1 among children with difficulties in reading comprehension at T3 with, on one hand, (1) PCA of language assessments at T1 of

the whole population of typical-age children (see **Figure 28**) and on the other hand, (2) with the PCA of children belonging to the worst quintile in oral comprehension at T1 and to children belonging to the worst quintile in meta phonology at T1 (see **Figure 29**).

Variables and their dimensions for the population with difficulties in reading comprehension at T3 were similar to the one of the general population (see **Figure 28**) and different from both PCAs of children with difficulties in oral language comprehension and in meta phonology (see **Figure 29**). Similar barycenters' positions were found for children with reading comprehension difficulties at T3 and for the whole population, with a large advantage for private and regular public schools (vs. a delay for PE and HPE schools) and an advantage for girls over boys.

Figure 28. Principal component analysis of language assessments at T1 among children facing difficulties in reading comprehension at T3.

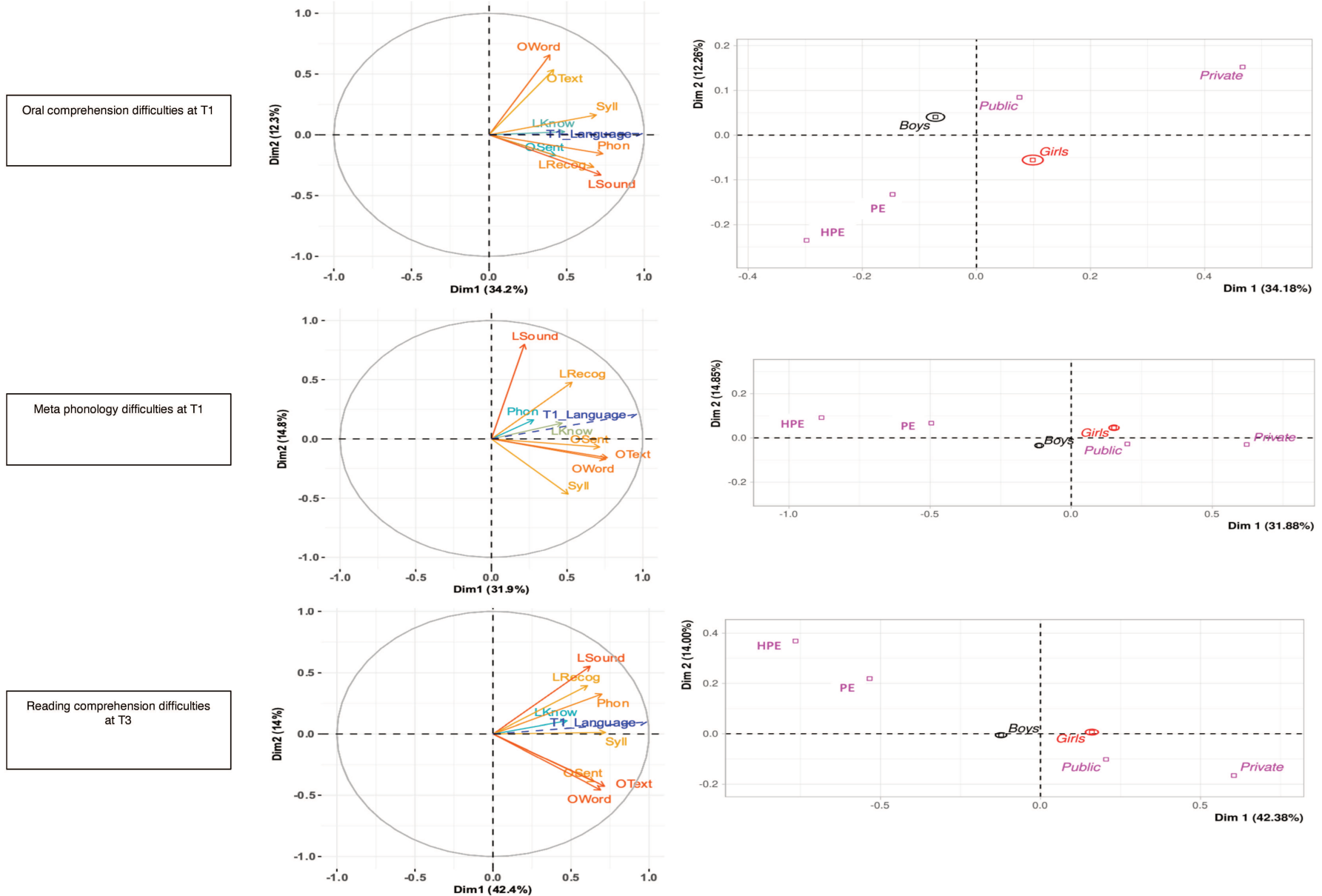


Then, when comparing the PCAs between the three following groups of difficulties (i.e., (1) Children belonging to the lowest quintile in oral comprehension at T1; (2) Children belonging to the lowest quintile in meta phonology at T1; (3) Children belonging to the lowest quintile in reading comprehension at T3), we found that very specific dimensions defined each population, and did not identify a common description of variables associated with the three types of difficulties (see **Figure 29**).

Particularly, we found that disparities of level between the four school categories (i.e., with a larger barycenters distance between private schools and PE and HPE public schools) is located on the oral comprehension (for words and texts) axis – which depend on linguistic immersion since birth, whereas more formal scholar domains (i.e., phoneme handling, syllable handling and letter knowledge) are less discriminant between schools (see **Figure 29**). Both among the three populations of lowest quintiles in language, Girls performed better than boys in average, and more particularly in meta phonology and in decoding, compare to oral comprehension (see the right part of the **Figure 29**).

Language assessments in first grade were categorized in three language subgroups that evolved in the same direction in the typical-age whole population with (1) oral language comprehension on one hand, (2) letter-sound association and letter knowledge and phoneme handling and (3) syllable handling on the other, whereas we observed 4 more distinct sub-groups in the reading comprehension, oral comprehension and meta phonology lowest quintiles with (1) oral comprehension of words and texts on one hand, (2) letter-sound association and letter knowledge, (3) phoneme and (4) syllable handling (see **Figure 28** and **Figure 29**).

Figure 29. Comparing all three PCA between (1) children with oral comprehension difficulties at T1, (2) children with meta phonology difficulties at T1 and (3) children with difficulties in reading comprehension at T3.



At last, when comparing both PCA of children with oral comprehension difficulties (P1) and PCA of the whole population (P2, introduced in Chapter 3), we noted that the better performance of girls in meta phonology for P1 was not found in the whole population (P2). Barycenters' gender gap was smaller (0.2) compared to the whole population (i.e., 0.4 on dimension 2). Similarly, barycenters' gaps were closer (0.8 in dimension 1) compared to 1.9 in the whole population. Disparities between the four school categories were similar (0.4 vs. 0.5 on dimension 2).

As both oral comprehension of words at T1 and meta phonology domains were identified as predictors of later reading comprehension at T3 (see **Table 14**), we wanted to investigate whether having difficulties in these two domains changed the predictions on later reading comprehension at T3. For instance, we implemented, on one hand, a multilevel model among selected children of the latest quintile in oral comprehension of language at T1, and on the other hand, we implemented a multilevel model among selected children of the latest quintile in meta phonology at T1 (see **Table 20**).

Class size was negative and significantly associated with later success in reading comprehension at T3 among children facing oral language difficulties at T1, whereas class size was not associated with reading comprehension abilities at T3 for children facing meta phonology difficulties at T1. This result suggested that children belonging to smaller classes when facing difficulties in oral comprehension, makes children progress in reading comprehension in second grade. Girls that presented meta phonology difficulties at T1 succeeded more in reading comprehension at T3 compared to girls with difficulties in oral language comprehension at T1. In both groups, SES score had a strong positive influence on the later reading comprehension abilities at T3. Neither the gender of the first of class, nor the boys-girls ratio per class were associated with the reading comprehension level at T3 (see **Table 20**).

Table 20. Multilevel models analyzing – among children with the highest difficulties in oral language comprehension at T1(left model) and among children with the highest difficulties in meta phonology at T1 (right model) – the predictors at T1 associated with a better level of reading comprehension at T3 in 2018.

Variables	Reading comprehension at T3			
	population of children with oral language difficulties at T1		population of children with meta phonology difficulties at T1	
n	112,979		114,823	
	<i>Estimate, β</i> (standard error)	<i>p</i>	<i>Estimate, β</i> (standard error)	<i>p</i>
Intercept	-0.0004 (0.0032)	< 0.0001	-0.0089 (0.0032)	0.0057
Age	0.0064 (0.0025)	0.0109	0.0076 (0.0025)	0.0026
Gender	-0.0473 (0.0052)	< 0.0001	-0.0501 (0.0052)	< 0.0001
T1 Oral Comprehension of Words, 0-15 (mean (SD))	0.0840 (0.0027)	< 0.0001	0.1135 (0.0034)	< 0.0001
T1 Oral Comprehension of Sentences, 0-14 (mean (SD))	0.0922 (0.0027)	< 0.0001	0.0928 (0.0031)	< 0.0001
T1 Oral Comprehension of Texts, 0-18 (mean (SD))	0.0719 (0.0027)	< 0.0001	0.1174 (0.0033)	< 0.0001
T1 Phoneme handling, 0-15 (mean (SD))	0.0650 (0.0032)	< 0.0001	0.0428 (0.0026)	< 0.0001
T1 Syllable handling, 0-15 (mean (SD))	0.1453 (0.0030)	< 0.0001	0.1219 (0.0028)	< 0.0001
T1 Letter-sound association, 0-10 (mean (SD))	0.1305 (0.0033)	< 0.0001	0.0646 (0.0027)	< 0.0001
T1 Decoding, letter writings recognition, 0-7 (mean (SD))	0.1503 (0.0031)	< 0.0001	0.1587 (0.0028)	< 0.0001
T1 Comparing letters, visuo-attentional abilities, 0-24 (mean (SD))	0.0542 (0.0028)	< 0.0001	0.0574 (0.0028)	< 0.0001
First of class in language is a boy	-0.0038 (0.0033)	NS (0.2504)	-0.0042 (0.0033)	NS (0.2007)
SES score	0.1254 (0.0037)	< 0.0001	0.1037 (0.0037)	< 0.0001
Class size	-0.0157 (0.0036)	< 0.0001	-0.0040 (0.0035)	NS (0.2546)
Boys-Girls ratio per class	0.0022 (0.0033)	NS (0.5043)	0.0036 (0.0033)	NS (0.2744)
Heterogeneity of language at T1 in the class	0.0075 (0.0034)	0.0284	0.0086 (0.0034)	0.0114

IV) Discussion

A. Main results and discussion

In this work, the whole population of France was longitudinally analyzed – and this, for four years in a row for a total of ~3 million children, using the tests administered by the French government as our outcome measures. Our approach has benefits that go beyond merely having a large n . First, it allowed us to be certain of the representativeness of our data. This is in sharp contrast to previous educational studies that are always based on subsamples (i.e., they test only a small, often unrepresentative sample of a country's school population), and there is always a doubt as to how representative they are. Second, not only it included children going to regular public and private schools, but it also included lower SES populations of children all in the meantime, a global approach often not considered in smaller study samples that focus on specific populations. Third, we had the opportunity to include both individual-level, class-level, and school-level data in the same study, which again, is not frequent.

Thanks to precise language assessments measured as early as the beginning of first grade in primary school (i.e., oral comprehension assessments, meta phonological assessments, and decoding skills), this study provided insight into the theoretical debate about the language predictors for reading comprehension in second grade. Notably, as we were able to detect the fine correlations between all specific cognitive language assessments in first grade and their common dimensions, we identified that among the general typical-age population, language assessments in first grade were categorized in three major groups of language performance (and therefore, groups of difficulties as well) which were (1) oral language comprehension on one hand, (2) letter-sound association and letter knowledge and phoneme handling on the other and (3) syllable handling.

Furthermore, we were able to detect different groups of difficulties, most of the time concerning a specific language domain (e.g., oral comprehension), which is needed to (1) better anticipate on every child's needs at the individual level and (2) on every child's needs at the classroom level to implement detection of children-at-risk of

difficulties, in line with others work (Adlof et al., 2017; Lauterbach et al., 2017). PCAs showed how all the language domains were positioned in several dimensions and therefore helped us identifying (1) children with oral comprehension difficulties at T1, (2) children with letter-sound association and phoneme handling difficulties at T1, (3) children with letter recognition difficulties at T1, (4) children with syllable handling at T1, (5) and children with crossed difficulties. Specifically, as these subgroups presented with language performances that varied and were distinct in their dimensions, it was possible to define their specific reinforcement needs, an important information which matters to provide future tailored support in the specific domains where they face difficulties, as suggested by others (Al Otaiba et al., 2011; Catts et al., 2016).

In addition, as several class- and school-level data were analyzed, we were able to identify the alarming delay and needs for children with a lower SES score, notably in oral language comprehension, compared to the other children. Also, we found that disparities of level between the four school categories (i.e., with a larger barycenters distance between private schools and PE and HPE public schools) was located on the oral comprehension (for words and texts) axis – which depend on linguistic immersion since birth, whereas more formal scholar domains (i.e., phoneme handling, syllable handling and letter knowledge) are less discriminant between schools. Furthermore, the school categories gap in language level worsened after the Covid-19, where school was closed for 52 days before the summertime break in 2019: when comparing language levels between 2018, 2019, 2020 and 2021, the longest children went to school, the best results they obtained.

Furthermore, we identified that needs according to the children's age varied importantly: indeed, advance-in-age, typical-in-age and late-in-age children's needs differed. Even if they were older, in average typical-in-age children performed worse in language and math compared to advance-in-age children, whereas late-in-age children struggled way more than both typical-in-age and advance-in-age children in both language and math. As we focused our analyses on typical-in-age children here, and found many different teaching needs subgroups, further research should also

explore the two other subgroups of age and precisely describe their difficulties and learning needs.

In addition, boys were more numerous among children with difficulties in reading comprehension, no matter which school type they belong to.

By leveraging this extensive dataset, we have been able to confirm that predictors for reading abilities and for reading comprehension differed as shown elsewhere (Castles et al., 2018): Predictors for reading abilities at T3 were firstly 'letter-sound association' and 'phoneme handling', both presenting with the highest predictive weight and secondly, were 'syllable handling', and 'letter-knowledge', and their predictive weight were similar between 2018, 2019, 2020 and 2021. This result contrasted with most research on reading abilities which identified phoneme awareness as the highest predictor of later reading abilities (i.e. decoding words and words in texts) (Clayton et al., 2020; Cunningham & Carroll, 2011; Sprenger-Charolles et al., 2003; Sprugevica & HØien, 2003).

On one hand, we could claim that we found more predictive weight for letter-sound association compared to phoneme handling due to a difference in the content and in the context of the tests, as notably in developmental psychology studies made on smaller populations, tests are very specific and take more time to assess a child compared to national generalized tests 'assessing phoneme awareness' in a classroom context and are passed in less than 5 minutes. On the other hand, the massive data, replicated 4 exhaustive population of children in a row, gave a 'new information' and indicated the high place of letter-sound association as higher than phoneme awareness among the reading abilities predictors. Also, and as seen earlier, the time window varied among the different studies, with more weight for decoding skills in preschool and more weight for oral comprehension of words by first and second grade. The time window of our study (i.e., from beginning of first grade and followed for 12 months) could have been associated with a dominant predictive weight for letter-sound association compared to phoneme handling.

Secondly, and consistent with some prior research, we have established that, meta-phonological skills (encompassing syllable and phoneme manipulation and letter-sound association) and oral comprehension of words (but not sentences nor texts) at T1 served as strong predictors of later reading comprehension abilities in second grade (T3) in line with several studies (Cain & Oakhill, 2014; Dong et al., 2020; Kendeou et al., 2009; Lervåg et al., 2018a; Massonnié et al., 2019; McBride–Chang & Kail, 2002). However, in other studies, meta-phonology and letter knowledge carried more substantial predictive weight for reading comprehension (H. Hjetland et al., 2017; H. N. Hjetland et al., 2019; Leppänen et al., 2008; Ozernov-Palchik et al., 2017) but were mainly directed among preschool children (i.e., younger than those in our study). Others argued that background knowledge served more as a major predictor of reading comprehension (R. Smith et al., 2021). Overall, even with variations between 2018 to 2021, our model comprised oral comprehension of words and meta phonological abilities as major landmarks for later reading comprehension.

Furthermore, our models enabled us to assess various parameters related to the learning environment (i.e., classroom characteristics) that influenced, or not, reading comprehension. Children with difficulties in reading comprehension at T3 tended to be younger (among the typical-age children) or to belong to late-in-age category of children, tended to be associated with a lower SES school type (i.e., PE and HPE), and tended to be boys (57% vs. 43%) compared to the general population. As described previously, comparing the PCAs between the three groups of difficulties, we found that very specific dimensions defined each population, and did not identify a common description of variables associated with the three types of difficulties. Remarkably, and contrary to math (see Chapter 4), our investigation revealed that none of the classroom characteristics significantly modified the level of reading comprehension, except for SES scores. A higher SES score was associated with superior reading comprehension at T3. For instance, children going to private schools had four times more chances to belong to the best quintile of level in phoneme handling, compared to children going to HPE schools. Also, among children with difficulties – either in oral comprehension of language or in meta phonology – belonging to a higher SES score was an important predictor for a better reading

comprehension level at T3. Additionally, the heterogeneity of language proficiency at T1 within a class was linked to a lower level of reading comprehension at T3. In other words, greater disparities in language proficiency within the same class at T1 hindered children's progress in reading comprehension. However, concerning decoding abilities at T3, class size did matter: Smaller classes and higher SES scores were both linked to fastest reading speed at T3 (i.e., more words read per minute).

It was very tempting to study the different progression trajectories in reading and reading comprehension between the four school categories to identify which school environment was more beneficial to children in difficulties. However, we were facing a large bias: As most tests were highly performed by children, it was not easy for them to perform any better, whereas children belonging to lower SES school categories had way more progress possibilities. In addition, as tests differed from T1 to T2 to T3, we were not able to compare the 'progress' of children for every language tests, but rather to observe how they performed in each domain per period of time.

Lastly, the unique context of the COVID-19 pandemic presented a natural experiment, allowing us to compare the challenging year of 2019 characterized by a substantial absence from school (52 days off plus a two-month summer break) and 2020 with three weeks off from school and the usual two-month summer break to the other years. Our results showed a significant drop of level in both reading, reading comprehension and language in 2019 that could be attributed to less exposure to the formal teaching of reading at school due to the natural experiment of Covid-19. Covid-19 event was the only major change in schooling exposure in between all the four years and as children presented with similar results before the Covid-19 happened (i.e., at T1 and T2) and did not differ in individual nor in environmental characteristics.

Notably, the summertime vacation was described in other studies as being associated with a widening of SES gaps in language and math (Shinwell & Defeyter, 2017). To summarize our findings, our study provided a more nuanced understanding of the essential components required for effective reading comprehension and reading,

parameters which are important for future research and implementation of reinforcement interventions in language.

B. Conclusion

A complete longitudinal follow-up of children in language, with assessments based on cognitive evidence for the developing child, allowed us to precise the predictors of developing reading comprehension skills in first and 2nd grade, and opened the possibility to identify patterns of difficulties in this domain as well as to implement a specific training to prevent the development and long-term implementation of reading delays. In addition, our results measured the direct effect of schooling, of class size and of other classroom parameters for the development of reading abilities in 2nd grade.

V) Supplementary materials

Figure S5. Language assessments' distributions in 2018, 2019, 2020 and 2021 at T1

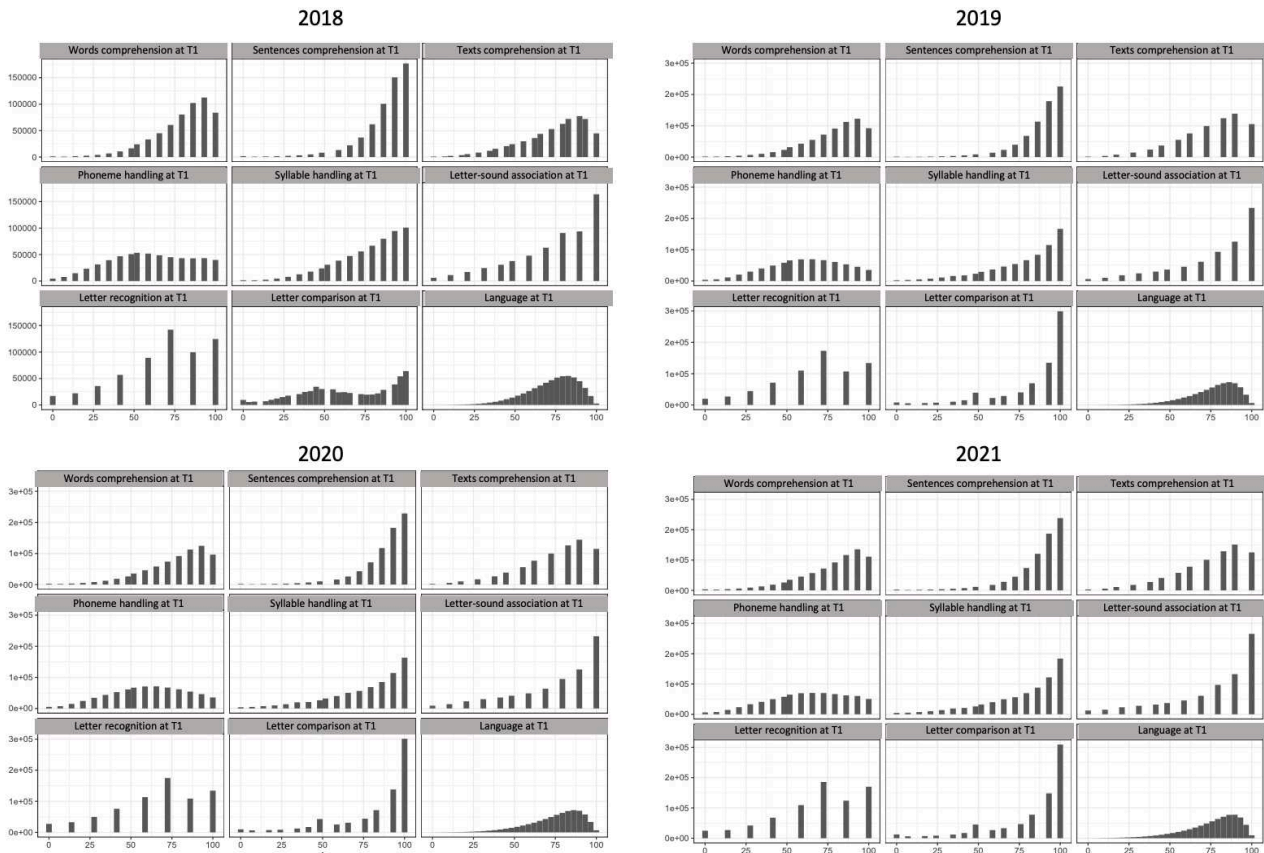


Figure S6. Language assessments' distributions in 2018, 2019, 2020 and 2021 at T2

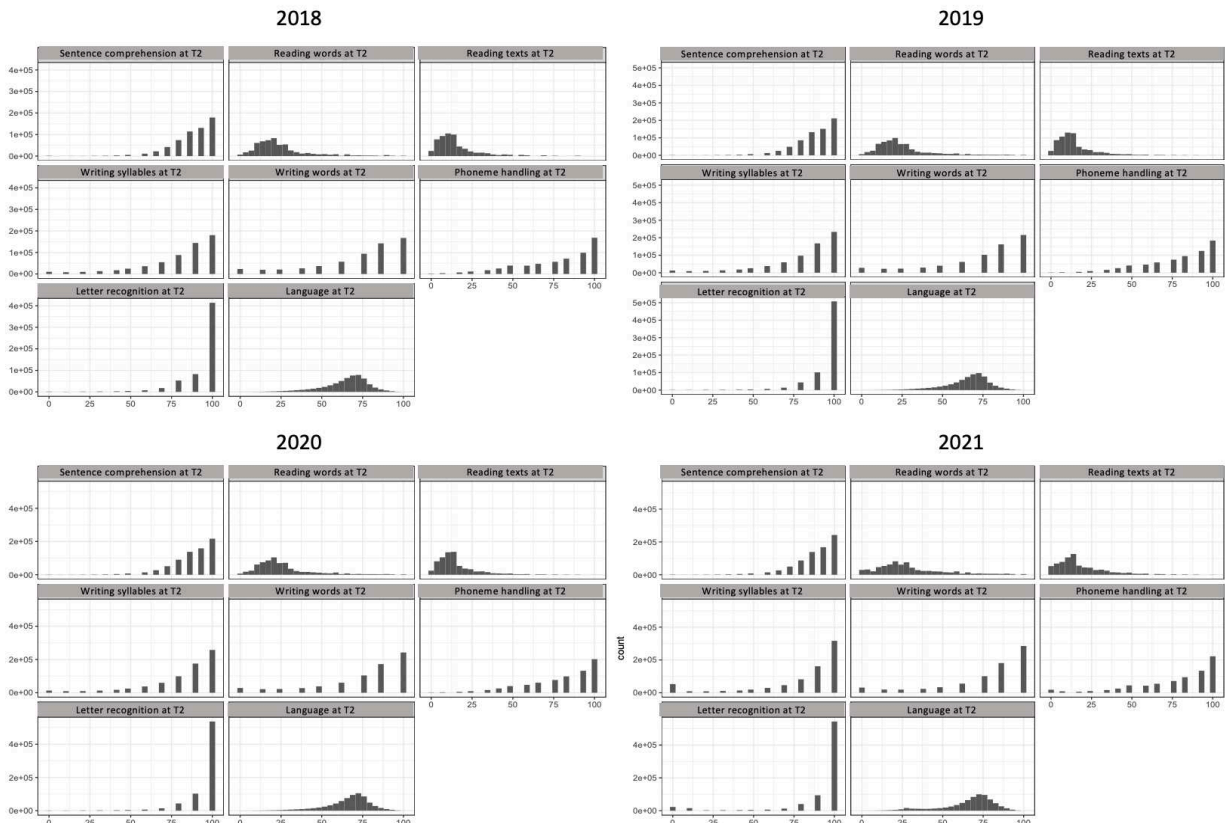


Figure S7. Language assessments' distributions in 2018, 2019, 2020 and 2021 at T3

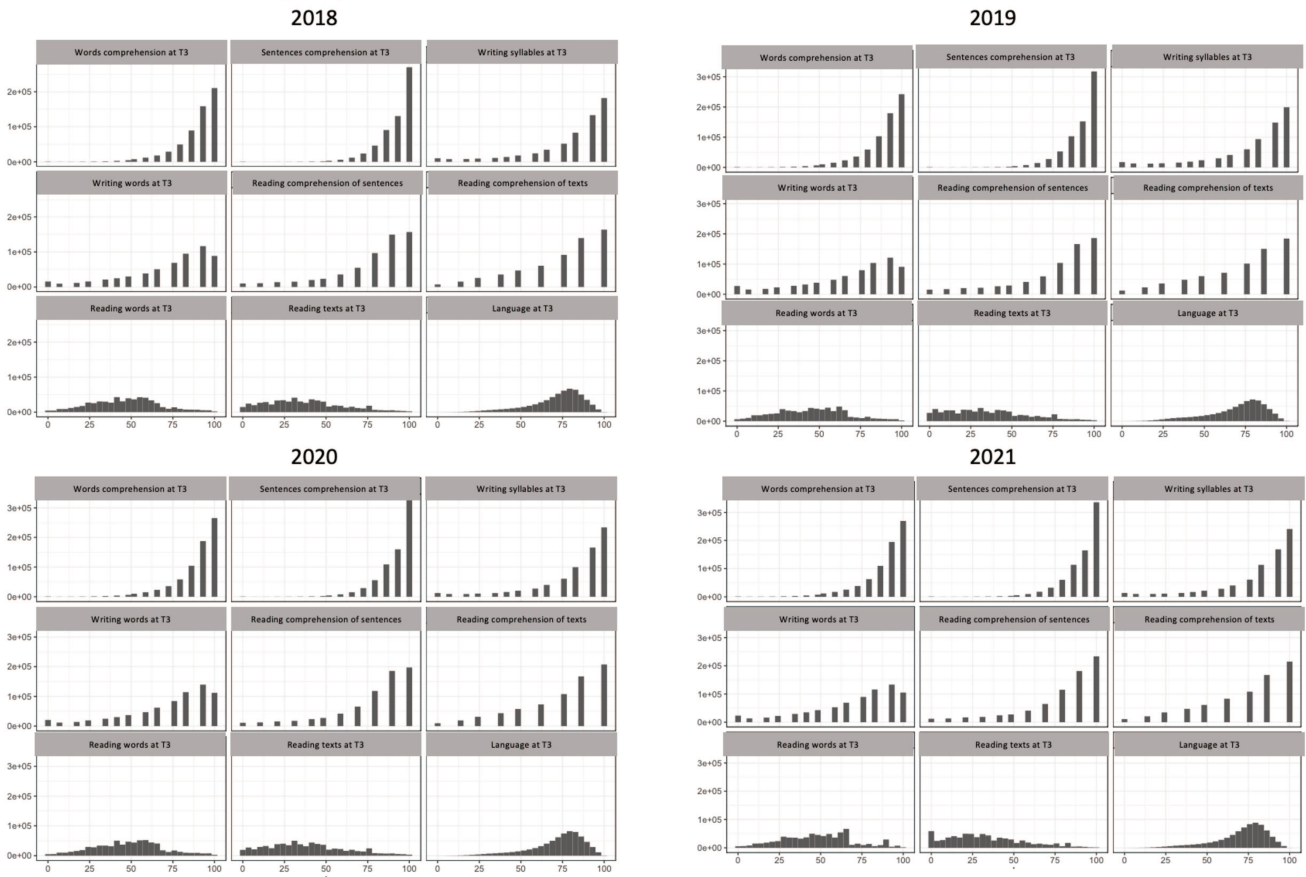


Figure S8. Correlation panels for language at T1 in 2018, 2019, 2020 and 2021

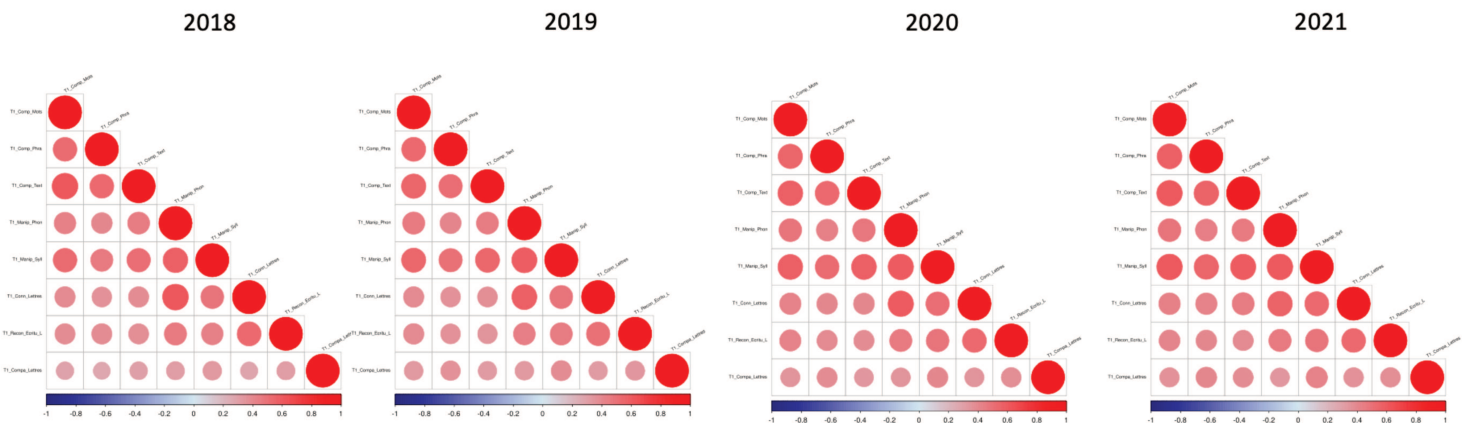


Figure S9. Correlation panels for language at T2 in 2018, 2019, 2020 and 2021

2018

2019

2020

2021

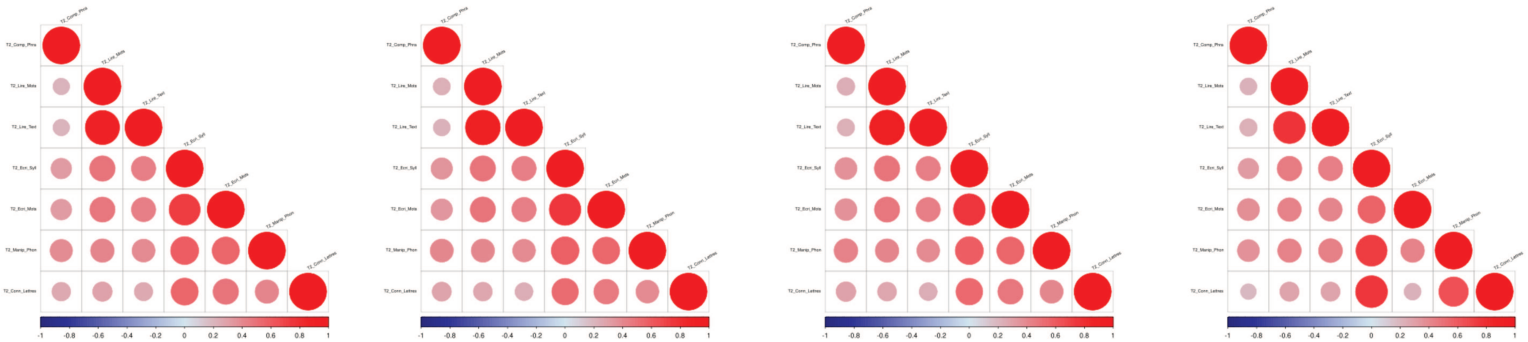


Figure S10. Correlation panels for language at T3 in 2018, 2019, 2020 and 2021

2018

2019

2020

2021

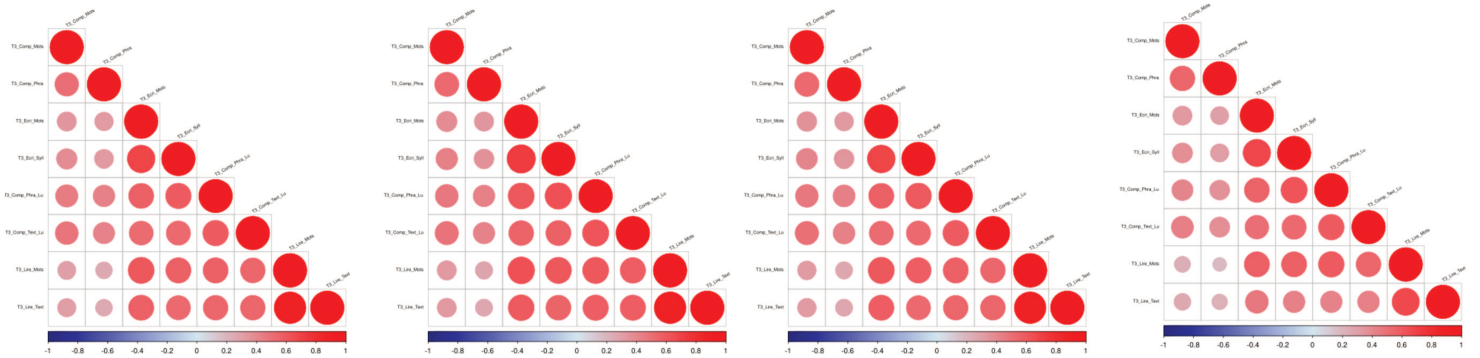


Figure S11. PCA results for language at T1, T2 and T3 in 2019

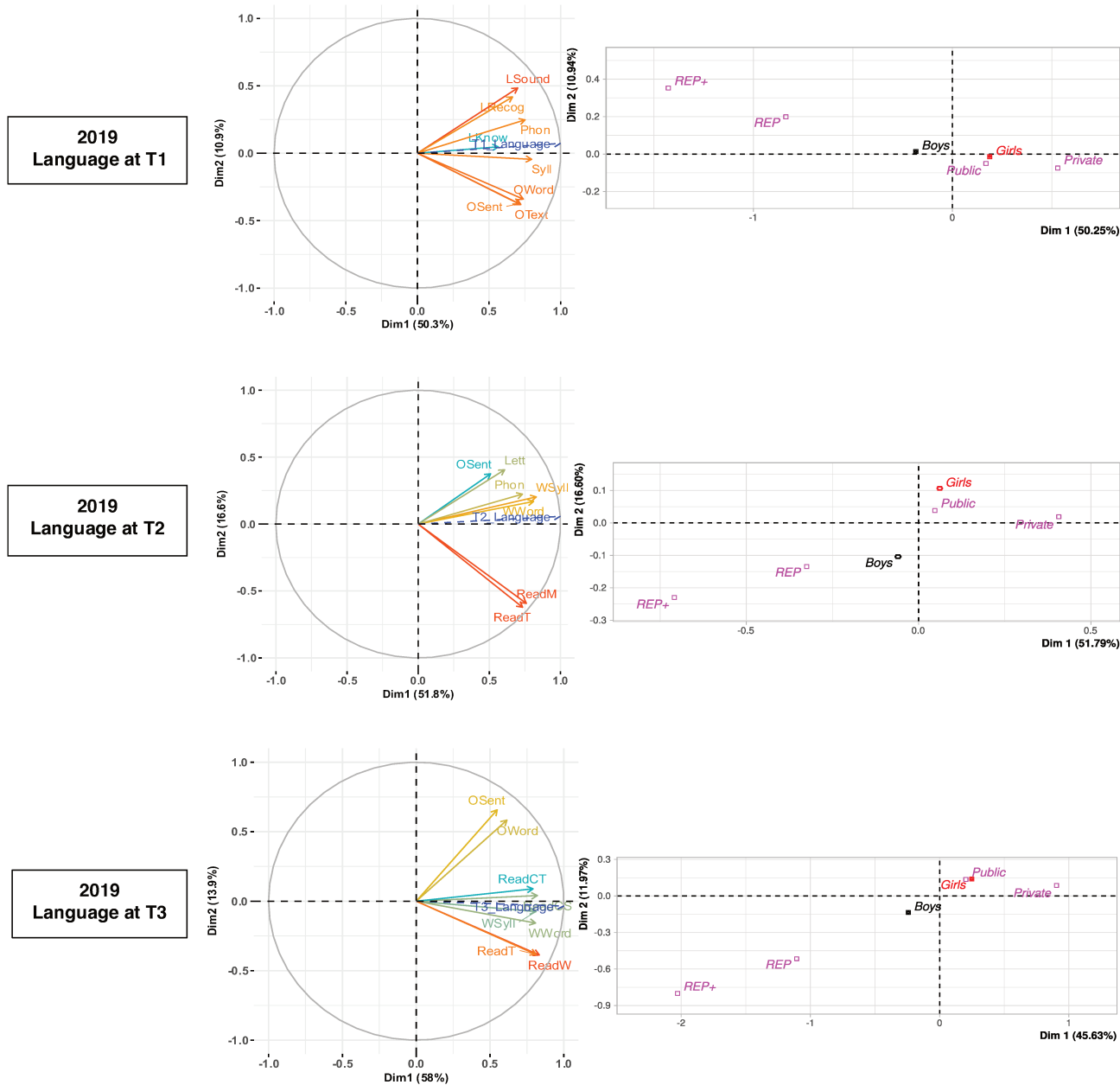


Figure S12. PCA results for language at T1, T2 and T3 in 2020

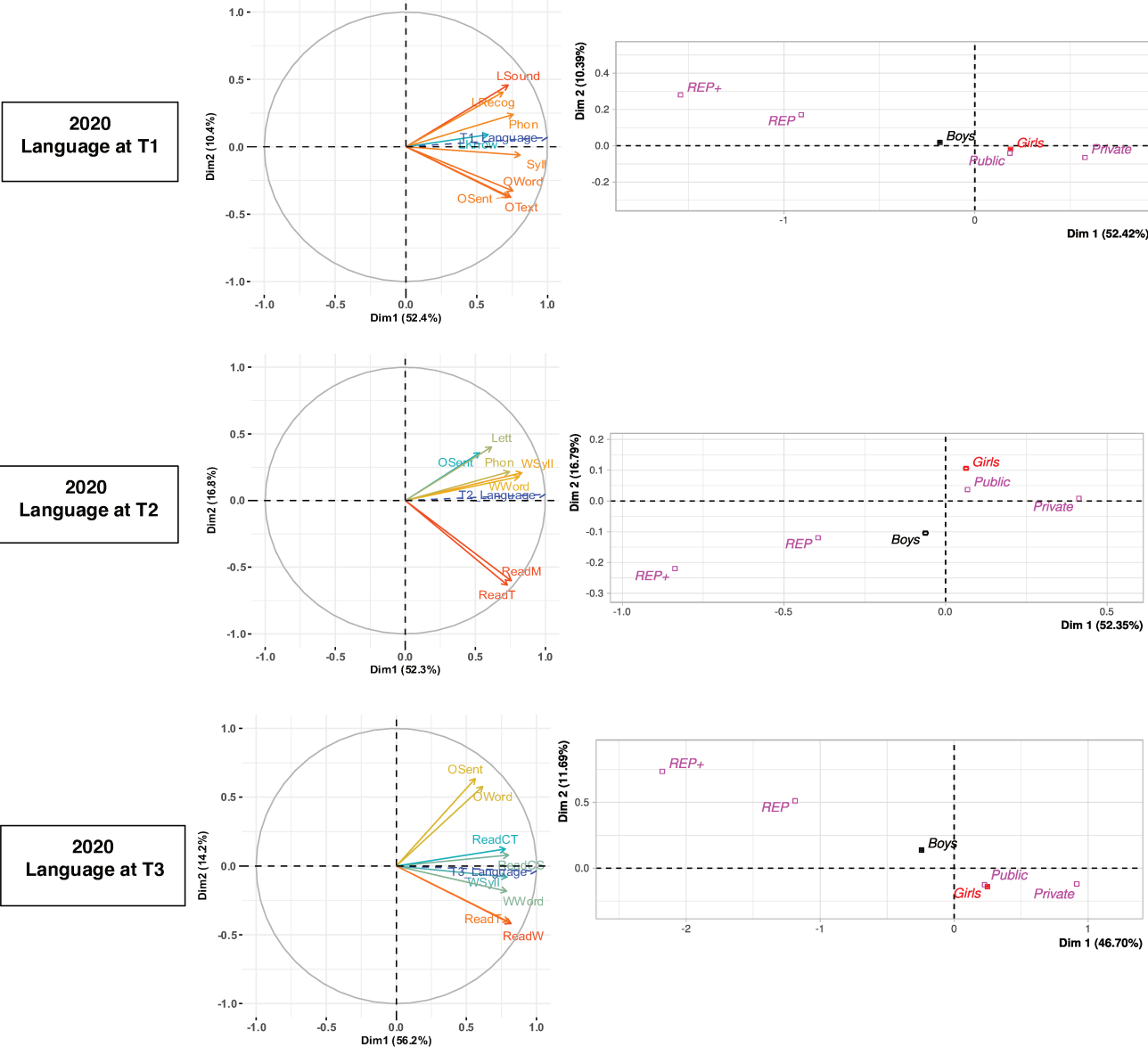


Figure S13. PCA results for language at T1, T2 and T3 in 2021

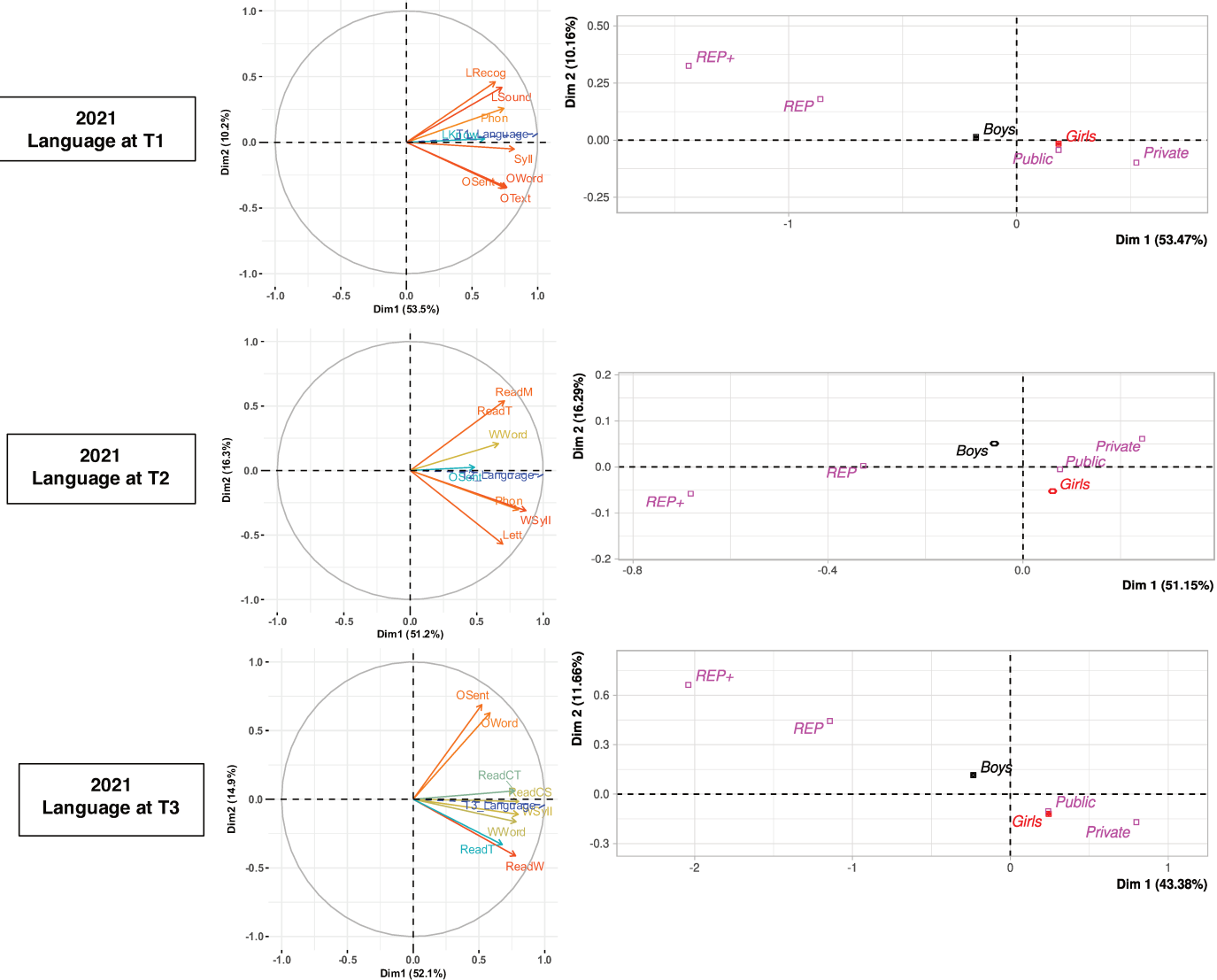


Table S9. Multilevel regression progressive models for reading at T3 in 2018. Reading is composed of the average of reading words and reading texts at T3.

	Model 1 Individual characteristics with language as mean of all tests at T1, Parameter estimate (SD)	p	Model 2 Individual characteristics with 3 latent variables in language at T1, Parameter estimate (SD)	p	Model 3 Individual characteristics with detailed language tests at T1, Parameter estimate (SD)	p	Model 4 Individual characteristics with detailed language and math tests at T1, Parameter estimate (SD)	p
Intercept	-0.0423 (0.0024)	< 0.0001	-0.0359 (0.0024)	< 0.0001	0.0420 (0.0023)	< 0.0001	0.0286 (0.0024)	< 0.0001
Age	-0.0019 (0.0012)	0.0981	0.0099 (0.0012)	< 0.0001	-0.0074 (0.0012)	< 0.0001	-0.0179 (0.0011)	< 0.0001
Gender (Girls < 0; Boys > 0)	0.0837 (0.0023)	< 0.0001	0.0711 (0.0023)	< 0.0001	-0.0862 (0.0021)	< 0.0001	0.0567 (0.0022)	< 0.0001
SES score	0.0801 (0.0024)	< 0.0001	0.0984 (0.0023)	< 0.0001	0.0701 (0.0021)	< 0.0001	0.0872 (0.0023)	< 0.0001
Language T1	0.5184 (0.0013)	< 0.0001	-	-	-	-	-	-
Math at T1	-	-	-	-	-	-	-	-
Oral language at T1	-	-	0.0684 (0.0016)	< 0.0001	-	-	-	-
Meta phono at T1	-	-	-0.2136 (0.0047)	< 0.0001	-	-	-	-
Decoding at T1	-	-	0.6834 (0.0047)	< 0.0001	-	-	-	-
Language	-	-	-	-	-	-	-	-
Oral comprehension of words at T1	-	-	-	-	0.0324 (0.0014)	< 0.0001	0.0228 (0.0015)	< 0.0001
Oral comprehension of sentences at T1	-	-	-	-	0.0160 (0.0013)	< 0.0001	-0.0007 (0.0014)	0.6185
Oral comprehension of texts at T1	-	-	-	-	0.0118 (0.0014)	< 0.0001	0.0057 (0.0015)	0.0001
Phoneme handling at T1	-	-	-	-	0.1589 (0.0016)	< 0.0001	0.1572 (0.0016)	< 0.0001
Syllable handling at T1	-	-	-	-	0.1194 (0.0014)	< 0.0001	0.0962 (0.0015)	< 0.0001

Letter-sound association at T1	-	-	-	-	0.1783 (0.0014)	< 0.0001	0.1657 (0.0015)	< 0.0001
Recognizing letter writing at T1	-	-	-	-	0.1543 (0.0013)	< 0.0001	0.1190 (0.0014)	< 0.0001
Comparing letters at T1	-	-	-	-	0.0820 (0.0013)	< 0.0001	0.0463 (0.0014)	< 0.0001
Math	-	-	-	-	-	-	-	-
Reading numbers at T1	-	-	-	-	-	-	0.0373 (0.0012)	< 0.0001
Writing numbers at T1	-	-	-	-	-	-	0.0523 (0.0013)	< 0.0001
Problem solving at T1	-	-	-	-	-	-	0.0207 (0.0014)	< 0.0001
Number line at T1	-	-	-	-	-	-	0.0319 (0.0013)	< 0.0001
Enumerating quantities at T1	-	-	-	-	-	-	0.0363 (0.0012)	< 0.0001
Comparing numbers at T1	-	-	-	-	-	-	0.0790 (0.0014)	< 0.0001
Variables per class	-	-	-	-	-	-	-	-
First of class in Language is a boy at T1	-0.0001 (0.0021)	0.9525	0.0001 (0.0020)	0.9795	-0.0018 (0.0020)	0.3758	0.0010 (0.0020)	0.6117
Class size	-0.0316 (0.0022)	< 0.0001	-0.0310 (0.0021)	< 0.0001	-0.0071 (0.0021)	0.0007	-0.0223 (0.0021)	< 0.0001
Boys - Girls ratio per class	-0.0002 (0.0021)	0.9086	0.0019 (0.0020)	0.3536	-0.0010 (0.0019)	0.6040	0.0020 (0.0020)	0.3232
Heterogeneity of language at T1	0.0468 (0.0021)	< 0.0001	0.0448 (0.0021)	< 0.0001	-0.0091 (0.0019)	< 0.0001	0.0497 (0.0021)	< 0.0001
Variance (intercept per class)	0.1058	-	0.0953	-	0.1014	-	0.0986	-
Residual	0.6517	-	0.6376	-	0.6051	-	0.5927	-

The best predictors for a higher level of reading at T3, when all other parameters are null, are the letter-sound association at T1 ($\beta = 0.1830^{***}$), phoneme handling ($\beta = 0.1682^{***}$), recognizing letters at T1 ($\beta = 0.1496^{***}$) and syllable handling ($\beta = 0.1177^{***}$). Girls have an advantage in reading abilities at T3 compared to boys. Oral comprehensions (of words, sentences, and texts) do not predict

the reading abilities at T3. A higher heterogeneity of level in language at T1 does not predict a higher reading level at T3. Also, a higher age does not correlate with a higher level in reading abilities at T3. Smaller class sizes are correlated with a higher level of reading at T3. Having a boy being first of class in language at T1 does not correlate with a higher reading level at T3. In math, comparing numbers and reading numbers are higher predictors than comparing letters for a high reading level at T3.

Table S10. Multilevel progressive regression models for Reading comprehension at T3 in 2018.

	Model 1 Individual characteristics with language as mean of all tests at T1, Parameter estimate (SD)	p	Model 2 Individual characteristics with 3 latent variables in language at T1, Parameter estimate (SD)	p	Model 3 Individual characteristics with detailed language tests at T1, Parameter estimate (SD)	p	Model 4 Individual characteristics with detailed language and math tests at T1, Parameter estimate (SD)	p
Intercept	0.0285 (0.0022)	< 0.0001	0.0310 (0.0021)	< 0.0001	-0.0296 (0.0017)	< 0.0001	0.0387 (0.0021)	< 0.0001
Age	0.0080 (0.0011)	< 0.0001	0.0166 (0.0011)	< 0.0001	0.0112 (0.0010)	< 0.0001	-0.0102 (0.0011)	< 0.0001
Gender (Girls < 0; Boys > 0)	-0.0562 (0.0021)	< 0.0001	-0.0612 (0.0021)	< 0.0001	0.0600 (0.0018)	< 0.0001	-0.0763 (0.0021)	< 0.0001
SES score	0.0942 (0.0022)	< 0.0001	0.0855 (0.0020)	< 0.0001	0.0686 (0.0016)	< 0.0001	0.0803 (0.0020)	< 0.0001
Language T1	0.6153 (0.0012)	< 0.0001	-	-	-	-	-	-
Math at T1	-	-	-	-	-	-	-	-
Oral language at T1	-	-	0.3064 (0.0014)	< 0.0001	-	-	-	-
Meta phono at T1	-	-	-0.1347 (0.0043)	< 0.0001	-	-	-	-
Decoding at T1	-	-	0.5032 (0.0043)	< 0.0001	-	-	-	-
Language	-	-	-	-	-	-	-	-
Oral comprehension of words at T1	-	-	-	-	0.1112 (0.0012)	< 0.0001	0.0991 (0.0014)	< 0.0001

Oral comprehension of sentences at T1	-	-	-	-	0.0796 (0.0011)	< 0.0001	0.0766 (0.0013)	< 0.0001
Oral comprehension of texts at T1	-	-	-	-	0.1092 (0.0012)	< 0.0001	0.0912 (0.0014)	< 0.0001
Phoneme handling at T1	-	-	-	-	0.0951 (0.0013)	< 0.0001	0.0769 (0.0015)	< 0.0001
Syllable handling at T1	-	-	-	-	0.1282 (0.0012)	< 0.0001	0.1158 (0.0014)	< 0.0001
Letter-sound association at T1	-	-	-	-	0.0962 (0.0012)	< 0.0001	0.0964 (0.0014)	< 0.0001
Recognizing letter writing at T1	-	-	-	-	0.0880 (0.0011)	< 0.0001	0.0751 (0.0013)	< 0.0001
Comparing letters at T1	-	-	-	-	0.0508 (0.0011)	< 0.0001	0.0307 (0.0013)	< 0.0001
Math	-	-	-	-	-	-	-	-
Reading numbers at T1	-	-	-	-	-	-	0.0538 (0.0011)	< 0.0001
Writing numbers at T1	-	-	-	-	-	-	0.0659 (0.0012)	< 0.0001
Problem solving at T1	-	-	-	-	-	-	0.0541 (0.0013)	< 0.0001
Number line at T1	-	-	-	-	-	-	0.0359 (0.0012)	< 0.0001
Enumerating quantities at T1	-	-	-	-	-	-	0.0658 (0.0011)	< 0.0001
Comparing numbers at T1	-	-	-	-	-	-	0.0773 (0.0013)	< 0.0001
Variables per class	-	-	-	-	-	-	-	-
First of class in Language is a boy at T1	< 0.0001 (0.0019)	0.9982	-0.0001 (0.0018)	0.9611	-0.0011 (0.0015)	NS (0.4429)	0.0009 (0.0018)	0.6195
Class size	-0.0108 (0.0020)	< 0.0001	-0.0111 (0.0019)	< 0.0001	-0.0000 (0.0015)	NS (0.9785)	-0.0035 (0.0019)	0.0613
Boys - Girls ratio per class	0.0016 (0.0019)	0.3930	0.0025 (0.0018)	0.1543	0.0014 (0.0014)	NS (0.3271)	0.0036 (0.0018)	0.0439
Heterogeneity of language at T1	0.0154 (0.0020)	< 0.0001	0.0156 (0.0018)	< 0.0001	-0.0178 (0.0014)	< 0.0001	0.0099 (0.0018)	< 0.0001
Variance (intercept per class)	0.0888	-	0.0752	-	0.0953	-	0.0709	-
Residual	0.5446	-	0.5461	-	0.6376	-	0.5234	-

All language variables at T1 (except for comparing letters) predict pretty similarly a higher level of reading comprehension at T3 when all other parameters are null. Decoding has almost twice the predictive effect on reading comprehension at T3 compared to oral comprehension (see model 2). Surprisingly meta-phonology is negatively correlated with reading comprehension at T3.

Girls have an advantage in reading comprehension compared to boys in every model, when boys used to outperform girls in reading abilities at T2 and at T3. Also, age correlates with a higher level in reading comprehension abilities at T3. Smaller class sizes are correlated with a higher level of reading comprehension level at T3. Having a boy being first of class in language at T1 does not correlate with a higher reading comprehension level at T3. In math, reading numbers, writing numbers, problem solving, comparing numbers, and enumerating quantities are higher predictors than comparing letters for a high reading comprehension level at T3. This can be explained by the need of both the following skills: 1) identifying symbols (i.e., numbers or letters) and 2) Count the number of phonemes in a word to decode it.

Table S11. ANOVA differencing Language at T1, T2 and T3 in 2018, 2019, 2020 and 2021.

	2018	2019	2020	2021	p
n	586,936	686,138	717,326	749,402	
	Mean (sd)	Mean (sd)	Mean (sd)	Mean (sd)	
<i>Language assessments</i>					
T3 Understanding reading a sentence, 0-10	7.78 (2.45)	7.61 (2.64)	7.85 (2.41)	7.91 (2.45)	< 0.0001
T3 Understanding reading a text, 0-8	5.97 (2.03)	5.79 (2.15)	5.98 (2.05)	5.92 (2.08)	< 0.0001
T3 Reading a list of 60 words in 1 min, 0-93	43.49 (19.06)	41.97 (20.06)	44.89 (19.26)	44.68 (19.85)	< 0.0001
T3 Reading 102 words of a text in 1 min, 0-136	50.98 (31.21)	48.79 (32.81)	52.68 (31.92)	45.60 (31.75)	< 0.0001
T3 Reading a list of 60 words in 1 min, 0-60	40.49 (15.16)	39.12 (16.26)	41.69 (15.14)	40.12 (14.87)	< 0.0001
T3 Reading 102 words of a text in 1 min, 0-102	49.04 (27.94)	46.77 (29.44)	50.49 (28.36)	49.54 (28.79)	< 0.0001
<i>Composite variables</i>					
Language at T1	72.70 (15.68)	76.02 (15.39)	74.86 (16.39)	75.78 (16.74)	< 0.0001
Language at T2	64.64 (13.89)	65.34 (13.65)	65.97 (13.74)	66.05 (16.24)	< 0.0001
Language at T3	71.29 (16.08)	69.53 (17.55)	71.78 (16.28)	70.71 (15.99)	< 0.0001

Table S12. ANOVA testing for reading abilities (i.e., decoding) between T2 and T3, and reading comprehension at T3 for all cohorts in 2018, 2019, 2020 and 2021.

Comparing 2018 to	Difference in level of Reading between T3 and T2		Difference in level of Reading comprehension at T3	
	Parameter estimates (sd)	p	Parameter estimates (sd)	p
<i>Intercept</i>	20.1393 (0.0243)	<0.0001	6.9259 (0.0027)	<0.0001
Year 2019	-2.0046 (0.0331)	<0.0001	-0.1756 (0.0036)	<0.0001
Year 2020	0.2836 (0.0328)	<0.0001	0.0377 (0.0036)	<0.0001
Year 2021	7.8386 (0.0325)	<0.0001	0.0401 (0.0036)	<0.0001

When year 2018 was considered as the reference, the reading comprehension level at T3 in 2019 was significantly lower ($\beta = -0.1756 (0.0037) ***$), whereas the reading comprehension level at T3 in 2020 was significantly higher than 2019 ($\beta = 0.2133 (0.0035) ***$) and non-different from the reading comprehension level at T3 in 2021 ($\beta = 0.0024 (0.0033) NS$). Globally, the reading comprehension level at T3 rose up from 2018 to 2021 ($\beta = 0,0401 (0,0035) ***$) and only the year 2019 was affected by a large drop of level. The level in reading words and texts between T2 and T3 significantly diminished in 2019 compared to 2018 ($\beta = -2.0046 (0.0331) ***$). However, the level rose up in 2020 and in 2021 compared to 2018 ($\beta = 7.8386 (0.0325) ***$).

Table S13. Results for reading comprehension of sentences at T3 in 2018, among children belonging to the lowest quintile of level in reading comprehension at T3.

Result	Reading comprehension of sentences among children facing difficulties in reading comprehension at T3 range 0-10 points										
	0	1	2	3	4	5	6	7	8	9	10
Frequency, n	7430	9767	12525	13518	15747	14960	13039	14625	8434	2650	411
Proportion, %	6,57	8,64	11,07	11,95	13,92	13,23	11,53	12,93	7,46	2,34	0,36
Cumulative proportion (X and less) %	6,57	15,20	26,28	38,23	52,15	65,38	76,91	89,84	97,29	99,64	100,00

Table S14. Results for reading comprehension of sentences at T3 in 2018, among children belonging to the other four quintiles of level in reading comprehension at T3.

	Reading comprehension of sentences among the four other quintile in reading comprehension at T3 range 0-10 points										
Result	0	1	2	3	4	5	6	7	8	9	10
Frequency, n	0	0	0	347	2855	7343	20892	38075	85320	143035	81409
Proportion, %	0,00	0,00	0,00	0,09	0,75	1,94	5,51	10,04	22,50	37,71	21,46
Cumulative proportion (X and less) %	0,00	0,00	0,00	0,09	0,84	2,78	8,29	18,33	40,82	78,54	100,00

Table S15. Results for reading comprehension of texts at T3 in 2018, among children belonging to the lowest quintile of level in reading comprehension at T3.

	Reading comprehension of texts among children facing difficulties in reading comprehension at T3 range 0-10 points									
Result	0	1	2	3	4	5	6	7	8	
Frequency, n	4882,00	13869,00	22835	27452	24796,00	10973,00	5480,00	1934,00	615,00	
Proportion, %	4,33	12,29	20,24	24,33	21,98	9,72	4,86	1,71	0,55	
Cumulative proportion (X and less) %	4,33	16,62	36,86	61,18	83,16	92,88	97,74	99,45	100,00	

Table S16. Results for reading comprehension of texts at T3 in 2018, among children belonging to the other four quintiles of level in reading comprehension at T3.

	Reading comprehension of texts among the four other quintile in reading comprehension at T3 range 0-10 points									
Result	0	1	2	3	4	5	6	7	8	
Frequency, n	0	0	875	5792	19717	47178	83412	134205	88097	
Proportion, %	0,00	0,00	0,23	1,53	5,20	12,44	21,99	35,38	23,23	
Cumulative proportion (X and less) %	0,00	0,00	0,23	1,76	6,96	19,40	41,39	76,77	100,00	

Table S17. Describing and comparing results in language subtests at T1, T2 and T3 of children belonging to the latest quintile of level vs. the other four quintiles and vs. children belonging to the best quintile in reading comprehension at T3.

	Latest quintile 20% with most difficulties in reading comprehension at T3	Other 4 quintiles 80% other students in reading comprehension at T3	Best quintile 20% most advanced in reading comprehension at T3
n	112836	379276	71293
Age at T1, month (mean (SD))	73.81 (3.39)	74.50 (3.41)	75.12 (3.36)
Class size, mean (SD)	16.76 (5.65)	17.39 (5.83)	17.69 (5.92)
SES score, mean (SD)	95.17 (18.07)	103.55 (17.28)	108.69 (16.26)
Gender – Boys, n (%)	64232 (56.9)	190564 (50.2)	30300 (42.5)
T1 Oral Comprehension of Words, 0-15 (mean (SD))	9.86 (3.08)	12.27 (2.33)	13.49 (1.65)
T1 Oral Comprehension of Sentences, 0-14 (mean (SD))	10.65 (2.81)	12.45 (1.77)	13.16 (1.21)
T1 Oral Comprehension of Texts, 0-18 (mean (SD))	10.64 (3.97)	13.74 (3.20)	15.48 (2.39)
T1 Phoneme handling, 0-15 (mean (SD))	6.29 (3.36)	9.30 (3.52)	11.30 (3.15)
T1 Syllable handling, 0-15 (mean (SD))	9.08 (3.34)	11.96 (2.71)	13.35 (1.98)
T1 Letter-sound association, 0-10 (mean (SD))	5.61 (2.76)	7.77 (2.30)	8.80 (1.72)
T1 Decoding, letter writings recognition, 0-7 (mean (SD))	3.52 (1.99)	5.01 (1.65)	5.67 (1.39)
T1 Comparing letters, visuo-attentional abilities, 0-24 (mean (SD))	12.19 (6.83)	15.78 (6.46)	17.47 (5.91)
T2 Number of words read out of a list of 30 words in 1 minute, range 0-100 (mean (SD))	13.53 (9.11)	26.36 (16.37)	37.06 (21.13)
T2 Number of words read out of a list of 29 words of a text in 1 minute, 0-195 (mean (SD))	13.72 (11.92)	31.48 (24.04)	48.35 (33.76)
T3 Number of words read out of a list of 60 words in 1 min at T3, 0-93 (mean (SD))	24.92 (14.53)	46.97 (16.21)	57.82 (15.93)
T3 Number of words read out of a list of 102 words in a text in 1 min at T3, 0-136 (mean (SD))	21.45 (19.12)	55.62 (27.43)	77.85 (28.03)
T3 Understanding reading a sentence, 0-10 (mean (SD))	5.32 (2.44)	6.50 (1.27)	10.00 (0.00)
T3 Understanding reading a text, 0-8 (mean (SD))	4.08 (1.59)	5.53 (1.23)	8.00 (0.00)
T3 Reading comprehension, 0-100, mean (SD)	40.85 (14.74)	83.36 (9.73)	100.00 (0.00)

Chapter 4. Schooling triggers a gender gap in math: evidence from three million children.

This chapter has been submitted as part of the following scientific article:

- Martinot P., Colnet B., Huguet P., Spelke E., Bressoux P., Dehaene-Lambertz G., Dehaene S. (*submitted to Nature*) “Schooling induces a gender gap in math: evidence from three million children”.

I) Introduction

Why are women underrepresented in Science, Technology, Engineering, and Mathematics (STEM) domains (OECD, 2015; Wang & Degol, 2017)? Biologically, all humans start life with core knowledge of objects, space, and number that serves as a foundation for mathematical development (Amalric & Dehaene, 2016; Dehaene, 1999; Spelke, 2005) and number sense, the ability to distinguish sets of objects based on their numerosity, is identical in male and female infants (Kersey et al., 2018).

In young children, most math-related cognitive tasks exhibit near-zero sex differences in overall performance, and distributions of inter-individual variability overlap massively across both genders (Hutchison et al., 2019; Hyde et al., 2008; Miller & Halpern, 2014; Spelke, 2005). The male advantage for mental rotation and spatial navigation skills which is occasionally reported in infancy (Levine et al., 2016; Miller & Halpern, 2014) is small, disputed, and may not occur until age five (Enge et al., 2023; Lauer et al., 2019). Furthermore, such disparities vary across cultures and testing conditions (Nosek et al., 2009).

For instance, gender differences favoring males in 3-D mental rotation tasks diminish when time pressure is removed (Voyer, 2011). Furthermore, where they exist, early gender differences are present at both ends of the scale: boys exhibit a higher rate of cognitive developmental disorders compared to girls (Zablotsky et al., 2019), and an excess of boys is found amongst both the best and worst performers in primary school (Hyde et al., 2008; Penner & Paret, 2008).

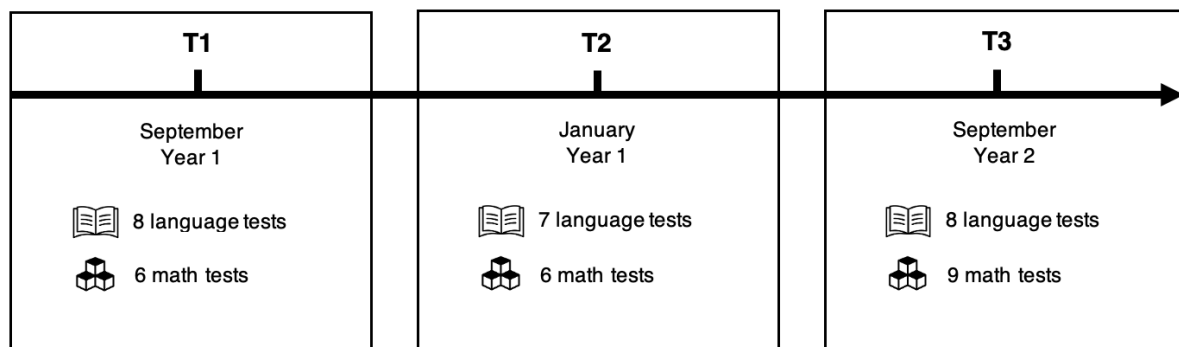
For these reasons, instead of a biological origin, young children's attitudes, perceptions, and interest and competence for math are thought to be primarily shaped by a sociocultural belief that girls exhibit lesser proficiency in mathematics relative to boys (Breda et al., 2020; Cimpian et al., 2016; Gunderson et al., 2012; Nollenberger et al., 2016). Indeed, both the size and the direction of the math gender gap, as well as attitudinal variables such as confidence in mathematics, valuing mathematics, and math anxiety, can change rapidly with affirmative political interventions (OECD, 2015, 2018a).

Adults' beliefs and stereotypes, including teacher's techniques and ratings, may interfere with the neutral estimation of students' performance and reinforce or reduce gender disparities in math achievement (Carlana, 2019; Cimpian et al., 2016; Miller & Halpern, 2014; Robinson-Cimpian et al., 2014). In particular, math anxiety in female teachers has been found to decrease girls' math performance, while boys remained unaffected (Beilock et al., 2010). Girls also suffer more from anxiety than boys, particularly on competitive or time-limited math tests, an effect that emerges as early as second grade worldwide (Van Mier et al., 2019). Parents and teachers may also be biased in the time spent challenging children of either gender in math and reading (Bharadwaj et al., 2016; Miller et al., 2018). Despite extensive research, the dynamics and associated factors of the math gender gap remain unclear.

Two broad hypotheses may be cited. First, sociocultural stereotypes may be slowly internalized by children as a function of age, as they gather increasing testimony that math is a male activity. Second, alternatively, specific trigger events such as schooling may play a causal role: when first confronted with math at school, female students, regardless of age, may invest less effort in this difficult discipline than boys because they quickly become aware, through teacher, parent and peer feedbacks as well as their own observations, that people like themselves are unlikely to enjoy and succeed in it (note that this explanation may account for gaps related to ethnicity and socio-economic status as well as gender (Gershenson et al., 2022)). In other words, school would be an "accelerant" for sociocultural learning about math and gender stereotypes, where exposure to formal schooling would create an explicitly competitive environment, in which more boys (for various reasons) find motivating and more girls find suppressing.

Resolving this question is essential, as specific mechanisms suggest radically different remediation strategies. Here, we shed light on these critical issues by taking advantage of the country-wide French national evaluation program *EvalAide* (“*évaluer pour mieux aider*”: assess to better help; **Figure 4** in Chapter 2).

Figure 4 (Chapter 2). Design of a country-wide longitudinal assessment of cognitive skills. EvalAide is a nation-wide longitudinal assessment of language and math abilities among all French 1st and 2nd graders, comprising three measurement periods (T1, T2 and T3). We present EvalAide data from four consecutive years (2018, 2019, 2020 and 2021), for a total of 2,871,080 children.



This battery of language and math tests was designed by scientists and educators to provide French teachers with a detailed picture of the needs, achievements, and progress of every child in their classroom, thus supporting focused pedagogical interventions and the setting of national standards. Every year, all French children received longitudinal tests at the beginning of first grade (T1), after 4 months of school (T2), and at the beginning of second grade (T3). Math tests included digit identification, counting, number comparison, number-line knowledge, problem solving, calculation, and geometry, while language tests cover letter knowledge, letter-sound correspondences, phonological awareness, reading aloud, vocabulary, oral comprehension and reading comprehension (see **Table 4** in chapter 2). Here, we analyzed four consecutive cohorts of 5-to-7-year-old first graders (2018, 2019, 2020 and 2021), for a total of ~2.8 million children.

II) Materials and methods

A. Materials

Data quality was high, and a reproducible data-management pipeline was implemented for the few missing values and outliers (~ 1.2% of all data; no bias was induced by these methods, see Chapter 2 – Material and methods). Because test difficulty increased, raw scores could not be directly compared across sessions. We therefore used normalized and gaussianized results (z-scores) on one hand, and Cohen's d as a gender gap effect size measurement on the other hand. The resulting scores were stable and sensitive, for instance exhibiting a strictly monotonic effect of every additional month of age on math performance, together with a large and increasing lead for children who were 1 year ahead, and an equally large and increasing lag for those who were 1 year behind (see Chapter 2).

As expected, school category and socioeconomical status (SES) had a major impact (see Chapter 2). Children in low-income school districts initially lagged behind but caught up to some extent in the course of 1st grade, in part due to a nationwide policy that halved classroom sizes in these districts (Bressoux et al., 2019). All these effects were stable across all years (2018, 2019, 2020 and 2021) and found in both language and math (see Chapter 2).

B. Methods

1) Statistical analyses

Whenever quantitative variables were compared, Cohen's d (see below) and Student's t tests were used, using *rstatix* package in *R software*, whereas when categorial variables were compared, Chi2 tests were implemented, both using the packages *tidyverse*, *dplyr* and *table1* packages in *R software*.

Gender gap effect size (in Cohen's d) in 2018, 2019, 2020 and 2021, and its stability.

Table 21 showed that gender gaps effects, measured in Cohen's d, were remarkably stable for math, language, problem-solving and number line both for T1, T2 and T3 and, in 2018, 2019, 2020 and 2021 and found for all socio-economical categories and ages. The results confirmed the rapid emergence of a gender gap favoring boys for number line and problem-solving assessments (positive values).

Table 21. Cohen's D effect size for gender gaps in 2018, 2019, 2020 and 2021 among children of normal age at T1

Variables	2018	2019	2020	2021
	N boys = 288,587 N girls = 281,184	N boys = 336,978 N girls = 328,654	N boys = 350,960 N girls = 344,489	N boys = 367,099 N girls = 355,131
Math at T1	-0.0166	0.0127	0.0082	0.0066
Math at T2	0.0468	0.0895	0.0832	0.0698
Math at T3	0.2230	0.1938	0.1974	0.2036
Problem solving at T1	-0.0546	-0.0212	-0.0224	-0.0239
Problem solving at T2	0.0296	0.0362	0.0362	0.0189
Problem solving at T3	0.1040	0.1030	0.1096	0.1364
Number line at T1	0.0271	0.0453	0.0487	0.0364
Number line at T2	0.0915	0.1085	0.1102	0.1105
Number line at T3	0.2588	0.2595	0.2731	0.1729
Language at T1	-0.1935	-0.1818	-0.1770	-0.1720
Language at T2	-0.0845	-0.0768	-0.0756	-0.0707
Language at T3	-0.1371	-0.1283	-0.1319	-0.0985

Gender gap measures for each specific subtest in math and language, comparing boys and girls.

To explore if the gender gaps identified in math and language were consistent in all subtests, we present the results for each subtest, each gender, and each cohort (2018, 2019, 2020 and 2021) (see **Table S18**). The results indicated that only a few subtests exhibited a reversal of the average gender gap. We can only offer here a few hypotheses about the origins of those reversals. Mental calculation was slightly superior in girls, perhaps because of its greater dependency on verbal automatisms (Dehaene & Cohen, 1997). Geometry was also slightly superior in girls, perhaps because it involved selecting among drawings and thus could be interpreted as a visual test that offered a break from all the other symbolic arithmetic tests (Huguet et al., 2001). Conversely, in speeded reading, boys tended to outperform girls, perhaps because they reacted better to the attentional challenge posed by speeded tests or because they reacted to what may look like a performance and competitive test when girls might exhibit more anxiety with time-limited exercises. Those reversed differences, however, were always small relative to the main gender gap effect reported in the main text.

Multilevel multivariate mixed regression models to measure predictive weight of variables on their outcome.

Before multilevel modelling, all quasi-continuous variables that previously underwent normalization (ranging from 0 to 100), were then gaussianized (centered and reduced with a mean = 0, and a standard deviation = 1). Among independent variables, only gender remained non-scaled.

In this study, children were taught within classes, all nested within schools. Due to these different environments, data contained natural groupings which impacted on individual children's performance. These multiple levels also implied that individual observations were not independently sampled. Multilevel linear mixed models allowed to overcome these two limitations of conventional models by accounting for nested

sources of variation in the data and avoiding assuming independently sampled data. Mathematically, nested patterns were introduced in the intercept and in the slope at the class level. Using stepwise multilevel models allowed to consider class effects (i.e., gender and math at T1) as random effects.

Corresponding to the class effect, the 2nd level random part of the multilevel model was specified step by step, following a stepwise multilevel model (see **Table S19**): intercept, gender, and math at T1 variances as well as their respective covariances proved significance. Math at T1 was introduced as a random variable as it was a strong predictor of math at T3. Gender was also introduced as a random variable because it represented our variable of interest. We explored progressively more complex linear regression models, starting with the simplest, eventually adding individual, contextual and interaction terms, as presented in **Table S19**. The decrease in deviance represented the model's significance (i.e., deviance model 1 = 1564075.7 and deviance model 10 = 1158166.8).

Multilevel Linear mixed modelling, fitted by maximum likelihood, was performed using the R package *lmerTest* and allowed to estimate both several individual and environmental parameters regarding the gender gap (see **Table 22**). Language and math' individual levels at T1, as well as gender, presented with the highest predictive coefficients for math level at T3. SES score's coefficient was more than 10 times smaller than the three previous variables and, compared to the other years, SES score predictive coefficient was more important in 2019 (i.e., year with less school exposure due to Covid-19). Age had a very small positive influence on math at T3. First of class being a boy was associated with a small but higher math level at T3. The boys-girls ratio per class did not have any significant association with math at T3. A wider class heterogeneity of level in math at T1 was associated with a lower level in math at T3. In addition, we focused on interactions with gender to analyze which factors were associated with a gender gap raise or a diminution in math at T3: Significant gender-related effects are highlighted in bold. Higher math level at T1, SES score and first of class being a boy, were associated with a gender gap raise in favor of boys, whereas a higher language level, age and heterogeneity of level in class were associated with

a smaller gender gap at T3. The model significance was estimated with the decrease of the model's deviance, showed in **Table S19** in SOM.

In addition, the overall effectiveness of a class, specifically its capacity to enhance the mean math scores at T3, as denoted by the 'intercept between-class variance' in **Table 22**, demonstrated a dynamic trend. Initially, in 2018 and 2019, there was a significant positive correlation. This positive correlation indicated that when boys exhibited greater improvements in their math performance compared to girls, the class as a whole demonstrated a higher efficiency in elevating their average math level. However, in contrast, during 2020 and 2021, we observed a subtle yet significant negative correlation between class efficiency and the gender gap in math performance.

This negative correlation in 2020 and 2021 suggested that when boys outperformed girls within a given class, the overall effectiveness of that class in improving students' math performance diminished. In simpler terms, prioritizing boys' progress did not contribute to an overall improvement in the class's math performance during these years. In fact, classes that excelled in teaching math during 2020 and 2021 tended to have a narrower gender gap. Notably, many classes implemented teaching strategies that not only raised the overall math proficiency but also narrowed the gender gap. Identifying these effective strategies should be a top priority for future research in this field.

Furthermore, across all cohorts, we consistently observed a negative correlation between the random effects of gender and the random effect of initial math levels within each class. For instance, the coefficient for 2018 was -0.31, as detailed in Table 1. This finding suggested that math scores at T3 were less dependent on the initial math scores at T1 in classes with a strong gender effect. In simpler terms, in classrooms where a noticeable bias favored boys over girls, with girls often achieving lower scores, the performance levels of both boys and girls at T3 showed a reduced dependency on their initial scores at T1.

The results obtained from our multilevel models were consistent and reproducible across multiple years, spanning 2019, 2020, and 2021, for math performance at T1, T2, and T3 (refer to **Table 22**). However, it's important to note that the outcomes in

language proficiency at T3 differed significantly from those in math proficiency at T3 for all cohorts (as shown in **Table S20**).

Analysis and figures of variables per class.

As this study measured the gender gap in various tests at school, and as it included both individual and class-level variables, we had the opportunity to test the following class-level associations with the gender gap in math (see **Figure S14** in SOM).

As we wanted to measure the effect of per-class-variables on the gender gap, we selected classes with a sufficient number of both genders (i.e., we selected at least 30% of boys and 30% of girls per class). Selections were indicated as “Step 5” in the **Table S21** and were applied for data management in 2018, 2019, 2020 and 2021. In addition, as we focused all of our analysis on the population of typical age in first grade (i.e., 69- to 80-year-old children) in this study, and as explained in materials, we firstly defined all class-level variables among children of all ages (i.e., advance-, typical-, late-in-age) in order to be representative of the variables’s class-effect. Then, as we wanted to explore the impact of class-level variables on the gender gaps between boys and girls of typical age in first grade, we focus the following graph on these children and did not include the extreme aged children (i.e., advance-in-age nor late-in-age children) in the following class-level analyses (see **Figure S14**). Therefore, our results were representative of the class-variables effects on gender gaps among children of typical age in first grade only.

Firstly, we measured the average math gender gap density over classrooms: The distribution was centered on zero at T1, but many classrooms showed a bias (i.e., a shift in favor of boys) at T2 and a more pronounced bias in favor of boys at T3. Results, presented as the average mean difference of boys’ class percentage of success in math minus girls’, were similar for 2018, 2019, 2020 and 2021 (see **Figure S14 - A**). Secondly, we measured the class’s gender gap of typical age children at T1, T2 and T3, in function of (1) class size, (2) class initial level in Math, (3) boys-girls ratio per class, (4) heterogeneity of level in math in the class and found a robustness in the

gender gaps (averaged within each classroom) in 2018, 2019, 2020 and 2021 (see **Figure S14 - B**).

Finally, we measured the impact of the role model in the class in math. For this latter, the graph was divided into two: the graph on the left represented classes with boys being first of class in math at T1, the graph on the right represented classes with girls being first of class in math at T1. For both graphs, we removed the math mean of the first of class and represented the class's gender gap of the rest of the class in order to visualize the impact the role model had on the other's progress in math and gender gap. Also, to be coherent with all the analysis of this paper, we focused the gender gap measurement on the population of typical age in first grade only and did not consider advance-in age nor late-in-age children. Having in mind that advance-in-age children had higher results in math, this selection on typical-age children explained the gender gaps results on **Figure S14 – C**. Having a girl or a boy as the first of class in math at T1 had a small but significant and positive association with the gender gap in math at T3 (see **Table 22**). In classes with a boy as a role model in math, boys had a larger advantage in math from T1 to T3.

Having in mind that in average, boys were more numerous among the extreme parts of the distribution (see **Figure 30-C**), when we withdrew the mean of the best boys in math from our analysis, it gave a small apparent advantage to girls in math at T1. Regarding the classes with girls as role model in math at T1, as less boys belong to the highest ranks of the distributions in these classes, and as more boys belong to the extreme low of levels in math compared to girls (see **Figure 30-C**), it gave a small apparent gender gap in favor of girls at T1. In addition, the global gender gap in math was less in favor of boys in classes where girls were the role model of the class in math at T1 (see **Figure S14 - C**).

Modelling math at T1 using a multi-level regression model to estimate association weight of variables with math at T1.

Most predictors for math at T1 were of the same significance, magnitude, and direction in 2018, 2019, 2020 and 2021 (see **Table S22**). The results indicated a massive effect

of expected predictors such as age and SES score. More crucially for the aims of this paper, the gender gap in math at T1 was non-significant in 2018 ($\beta = -0.0044 (\pm 0.0050)$, NS), whereas it was small and in favor of boys in 2019, 2020 and 2021. The highly significant Gender * SES score interaction indicated, however, that children with a higher SES were already affected by a gender effect. The negative effect of class size, at the beginning of the year, was somewhat surprising but might reflect a genuine early influence of class size on test results (see **Table S22**).

Modelling math at T2 using a multi-level regression model to measure predictive weight of variables on later math level at T2.

As **Table S23** showed, for math at T2 the main effects related to gender were smaller, but very similar to those at T3: there was already a large and significant advantage for boys, which was larger for children with a higher level in math at T1 and with a higher SES, and smaller for children with a higher level in language at T1. Contrary to T3, however, having a boy as first of class in math at T1 was not yet influential; in fact it was slightly but significantly negatively correlated with the gender gap in favor of boys at T2 (only in 2018 and in 2021 but not in 2019 nor in 2020).

Modelling language at T3 using a multi-level regression model to measure predictive weight of variables on later language level at T3.

In language, a negative effect of genders indicated that, everything else being equal, including language and math performance at T1, girls showed better performance than boys at T3 (see **Table S20**). The gender effect coefficient in language, however, was 10 times smaller than the gender effect on math (comparison with **Table 22**). Furthermore, variables such as initial level in language or in math, boys-girls ratio per class, or age did not contribute to significantly modulate the change in the gender gap in language from T1 to T3. The role model effect of having a boy as first of class in language did have a small influence in 2 out of 3 cohorts. A larger class size and a higher SES score also favored boys.

Matching and causal inference methods to estimate the influence of gender on math results.

Matching techniques and six causal inference methods (i.e., Average weighting using G-computation, propensity weighted regression, inverse propensity weighting (IPW), doubly-robust estimation (AIPW) with various nuisance components estimation techniques such as Ordinary Least Square (OLS), logistic regression (logit), random forest approaches and target maximum likelihood estimation (TMLE) for causal inference) were implemented to test and confirm all previous analyses by ‘estimating the causal effect of gender’ on the math’s level. Compared to matching, the latter techniques present the advantage that no data is dropped. Within the several nuisance components estimation techniques, the random-forests approaches cancel the parametric assumptions inherent to logistic regression or OLS, and are supposed to be more reliable, at least in large samples (which is the case in the present study).

Firstly, we started with the most intuitive causal inference model: matching. The concept of matching relies on the emulation of randomized controlled trials using massive observational data. The idea is to identify pairs of individuals who are matched according to initial characteristics, and differ in only one parameter (i.e., gender). This method allows to ‘estimate the causal effect’ of one variable (i.e., gender) on an outcome (i.e., Math performance at T3). We considered two matching scenarii (see **Figure 32**):

- **Matching only at T1** - Pairs were matched on school type (Private/ regular public vs. PE/HPE), on deciles of SES score, on age in first grade (+/- 4 months), on the 6 tests in Math at T1 (+/- 5 points over 100), as well as on their mean in language at T1 (+/- 5 points over 100). Results were shown in **Table S24**.
- **Matching at T1 and at T2** - This scenario is a variant of the previous scenario where both level of math at T2 and level of reading at T2 were added and pairs were matched when results corresponded to +/- 5 points.

Those scenarii, the number of pairs found and their means were detailed in **Figure 32** and **Table S24**. For those analyses, we used the package *MatchIt* in R. Once the

matching was performed, a simple test for a difference in means between boys and girls was enough when using exact matching, but to adjust for any potential remaining imbalance, we used linear regression to estimate the effect. Results were detailed in **Table S24**.

As matching analysis focused on matched pairs (i.e., running the risk of being unrepresentative of the cohort, more particularly both scenarii led to an over-representation of children in the top results), we then implemented six additional causal estimation techniques. The same covariates than matching were used for adjustment, only the statistical methodology changed. Intuitively, those approaches allowed all models to decide how to best capture any initial imbalance in the so-called nuisance functions, rather than pairing individuals on them.

Contrary to matching, these methods did not drop any data, but either weigh them differentially, or relied on an outcome model to infer the average effect of being perceived as a girl or a boy, all other characteristics remaining equal. The first four methods used parametric nuisance functions, while random forest and TMLE relied on a non-parametric approach. Results were presented in **Table S25**. No matter which causal inference techniques was implemented, all results tended towards same conclusions: a large emerging gender gap (i.e., + 4 to 5 points over 100 points) favoring boys in math. Note that instead of using z-scores here, we implemented all causal inference models using the mean in math and language in percentage of success (i.e., explained in the material section above), ranging from 0 to 100 points. These variables were both used in the matching selection process, as well as on **Figure S14** for estimating the gender gaps per class and both in the sensitivity analysis presented in **Table 7** in Chapter 2 and in **Table 21** and **Table S18** to measure boys-girls differences per test at T1, T2 and T3.

Student T-test to compare gender gap differences between the four cohorts.

Statistics related to **Figure 31** were presented in **Table S26** where, on one hand both gender gaps from T1 to T2 and from T2 to T3 were measured in 2018, 2019, 2020 and

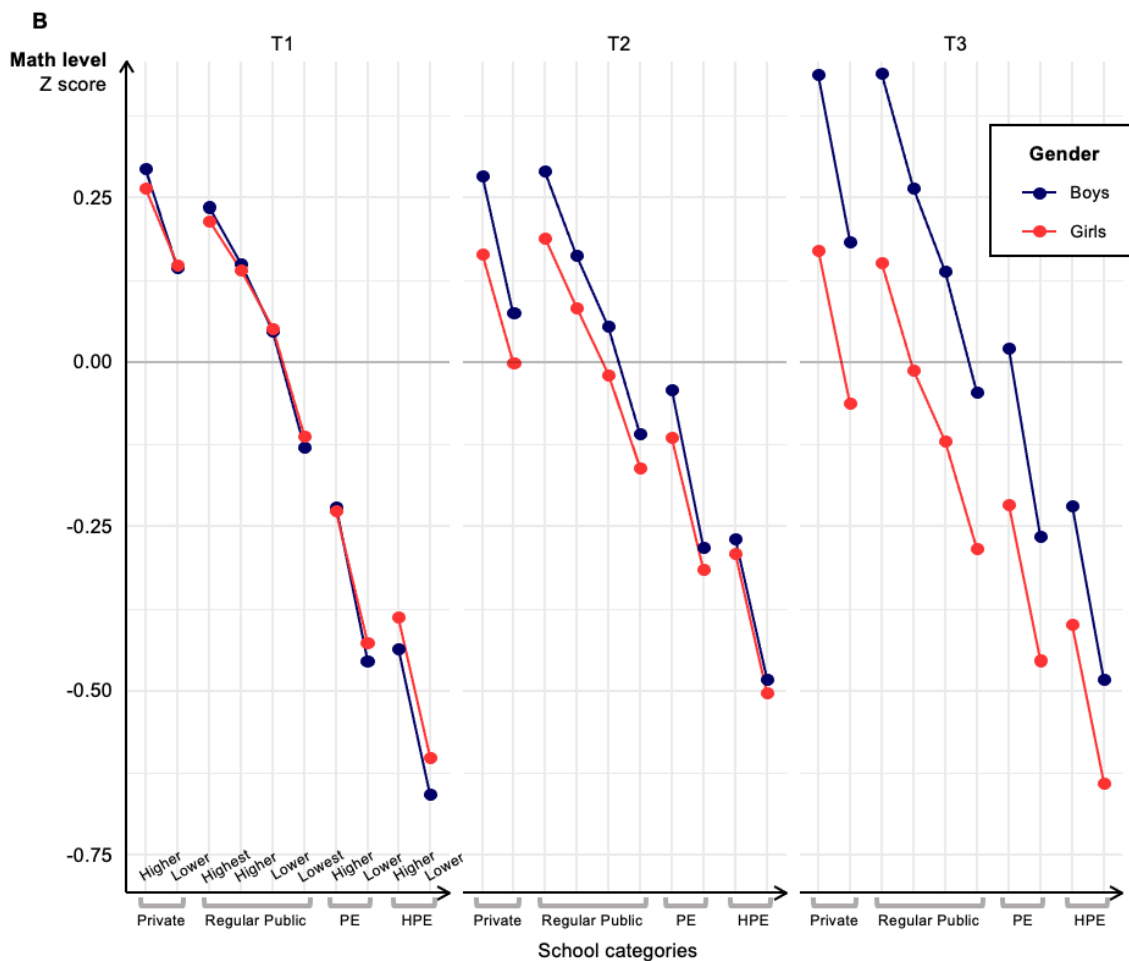
2021, and on the other hand, comparison between 2018-2019, 2019-2020, 2020-2021 and 2018-2021 were presented. **Table S26** confirmed the drop in the gender gap found in 2019 between T2 and T3 (i.e., difference = -1.23234 ***), the almost null difference between 2019 and 2020 (i.e., difference = 0.00354***) and the gender gap raise in 2021 compared to 2020 (i.e., difference = 0.48306***). These large variations in differences were not found for gender gaps between T1 and T2 in math, nor in language. Regarding the difference in gender gaps between T2 and T3 in language, gender gaps' magnitudes were smaller compared to math's. We found a significant larger gender gap in favor of girls in 2019 and 2020 compared to 2018 and 2021 (2019-2018 difference = -0,18056***; 2021-2020 difference = 0,66929***). Altogether, these results indicated that the gender gap's difference between T2 and T3 in math was significantly larger in 2018 and in 2021 compared to years 2019 and 2020 in favor of boys, and the gender gap's difference in language between T2 and T3 was significantly smaller in 2018 and in 2021 compared to years 2019 and 2020, illustrating an advantage for boys in language in 2019 and 2020 compared to 2018 and 2021.

III) Results

A. Rapid emergence of a math gender gap in first grade

Strikingly, the data revealed the rapid emergence of a math gender gap (**Figure 30**). There was a close-to-zero gender gap on math performance at school entry (Cohen's $d = -0.0166$), but a highly significant one favoring boys after 4 months of schooling (Cohen's $d = 0.0468$) and a massive one at the beginning of 2nd grade (Cohen's $d = 0.2230$; i.e., Boys presented with ~ 5 points over 100 more than girls, which in a class of ~20 pupils, corresponded to boys gaining ~1 position relative to girls, see **Table 21** and **Table S18** in SOM). Such a rapid emergence was replicated in every cohort, within every school category and socio-economical level (**Figure S15** in SOM), and in most math tests, including the problem-solving and number line subtests that were repeatedly probed at each time point (**Figure S16** in SOM).

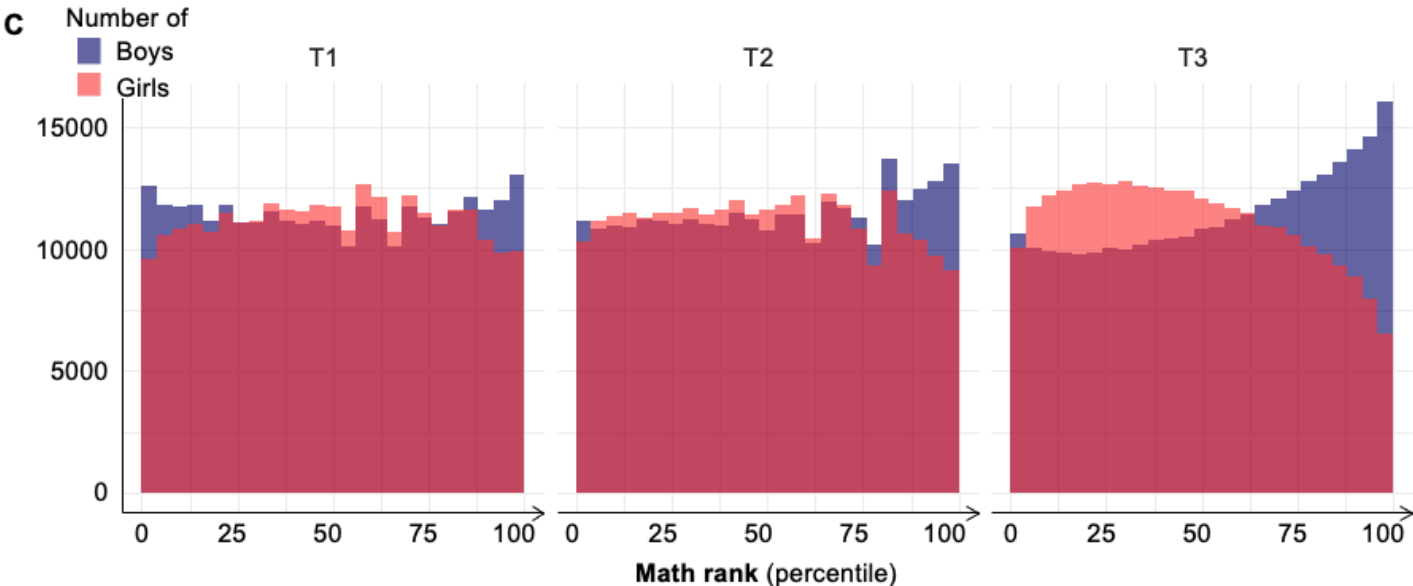
Fig. 30. Rapid emergence of the math gender gap found in the national program Evalaide. Overall performance of boys (blue) and girls (red) in mathematics.



For each of four school categories (private schools, regular, priority education [PE], and higher-priority education [HPE] public schools), a median split or quartile split (for regular public schools only) was implemented based on the school average socio-economic status (SES) score, with – for each school category - higher SES scores on the left and lower SES scores on the right of the x-axis. Within each school category, the gender gap is absent at school start (T1), detectable after 4 months (T2), and large after one year of schooling (T3) and, reproducible in 2019, 2020 and 2021 (see supplementary material).

Examination of the distribution of math scores over children clarified how the gender gap emerged (**Figure 30 - C**). At school onset (T1), although boys and girls had the same mean, boys were over-represented at both ends of the distribution (worst and best deciles), as previously described in older children (Cimpian et al., 2016; Hyde et al., 2008; Penner & Paret, 2008). By one year of schooling, however, the distribution shifted massively, with the upper percentile ranks in math at T3 comprising approximately twice as many boys as girls. Similar results were found when the gap was computed within each class as the difference in mean performance between boys and girls (**Figure S14** in SOM).

Figure. 30C. Distribution of ranks in math among boys and girls, showing an initially higher density of boys in both high- and low-performers, quickly shifting to a large advantage in favor of boys and, reproducible in 2019, 2020 and 2021 (see supplementary material (SOM)).



Elder children presented a *smaller* gender gap at T2 and T3, contrary to the hypothesis of a slowly accruing sociocultural bias (i.e., $\beta_{\text{gender*age}} = -0.0025 (\pm 0.0005)$ *** in **Table 22** and negative slope on Figure 2A). More particularly, and regardless of age, there was a sudden and rapid increase in the gender gap from T1 to T2 and from T2 to T3 (Figure 2A and Table S4-S5). This result was significant when age at test was kept constant, within every age slice in months. Thus, the gender gap does not increase slowly with age, but is rather triggered and amplified by schooling.

Table 22. Analysis of fixed and random factors associated with children’s math scores at T3, using multilevel mixed regression models on normal-aged children ($n_{\text{Total}} = 1.783.666$ children). Note: Significant gender-related effects are highlighted in bold. The model significance was estimated with the decrease of the model’s deviance, showed in Table S7 in SOM.

Variables	Math individual level at T3							
	2018		2019		2020		2021	
N individuals	569,771		665,632		695,449		722,230	
N groups (classes)	39,573		46,671		49,010		49,701	
Fixed effects	Estimate	p	Estimate	p	Estimate	p	Estimate	p
Intercept	-0.0000 (0.0026)	NS (0.9956)	-0.1408 (0.0028)	< 0.0001	-0.1693 (0.0028)	< 0.0001	-0.1739 (0.0028)	< 0.0001
Language individual level at T1	0.4078 (0.0013)	< 0.0001	0.4164 (0.0017)	< 0.0001	0.3946 (0.0017)	< 0.0001	0.3861 (0.0017)	< 0.0001
Math individual level at T1	0.3810 (0.0013)	< 0.0001	0.3306 (0.0017)	< 0.0001	0.3506 (0.0018)	< 0.0001	0.3738 (0.0017)	< 0.0001
Gender (Boys)	0.3453 (0.0038)	< 0.0001	0.2975 (0.0035)	< 0.0001	0.3099 (0.0035)	< 0.0001	0.3025 (0.0034)	< 0.0001
SES score at T1	0.0277 (0.0020)	< 0.0001	0.0646 (0.0019)	< 0.0001	0.0055 (0.0019)	0.0047	0.0001 (0.0019)	NS (0.9547)
Age at T1 (month)	0.0016 (0.0003)	< 0.0001	0.0008 (0.0003)	0.0139	0.0049 (0.0003)	< 0.0001	0.0062 (0.0003)	< 0.0001
Heterogeneity in math at T1	-0.0305 (0.0018)	< 0.0001	-0.0187 (0.0017)	< 0.0001	-0.0259 (0.0017)	< 0.0001	-0.0242 (0.0017)	< 0.0001
Boys-Girls ratio per class	0.0005 (0.0018)	NS (0.7740)	-0.0024 (0.0018)	NS (0.1677)	-0.0057 (0.0018)	0.0012	0.0001 (0.0018)	NS (0.9480)
First of class is a boy in math at T1	0.0063 (0.0019)	0.0008	0.0068 (0.0018)	0.0002	0.0043 (0.0018)	0.0172	0.0052 (0.0018)	0.0036
Class size	0.0095 (0.0020)	< 0.0001	0.0065 (0.0019)	0.0007	0.0069 (0.0019)	0.0003	0.0115 (0.0019)	< 0.0001
Gender * Language individual level at T1	-0.0065 (0.0024)	0.0077	-0.0138 (0.0023)	< 0.0001	-0.0206 (0.0023)	< 0.0001	-0.0310 (0.0023)	< 0.0001
Gender * Math individual level at T1	0.0644 (0.0024)	< 0.0001	0.0655 (0.0023)	< 0.0001	0.0707 (0.0023)	< 0.0001	0.0552 (0.0022)	< 0.0001
Gender * SES score at T1	0.0049 (0.0020)	0.0148	-0.0014 (0.0019)	NS (0.4486)	0.0060 (0.0019)	0.0016	0.0052 (0.0018)	0.0039
Gender * Age at T1	-0.0025 (0.0005)	< 0.0001	-0.0013 (0.0005)	0.0071	-0.0023 (0.0005)	< 0.0001	-0.0026 (0.0005)	< 0.0001

Gender * Heterogeneity of level at T1	-0.0048 (0.0018)	0.0081	-0.0040 (0.0017)	0.0160	-0.0013 (0.0017)	NS (0.4236)	-0.0049 (0.0016)	0.0029
Gender * Boys-Girls ratio per class	-0.0010 (0.0020)	NS (0.6127)	0.0037 (0.0018)	0.0363	-0.0005 (0.0018)	NS (0.7770)	-0.0045 (0.0018)	0.0110
Gender * First of class is a boy in math at T1	0.0064 (0.0019)	0.0006	0.0030 (0.0017)	NS (0.0799)	0.0066 (0.0017)	0.0001	0.0062 (0.0017)	0.0002
Gender * Class size	0.0043 (0.0020)	0.0276	0.0047 (0.0018)	0.0095	0.0010 (0.0018)	NS (0.5958)	0.0030 (0.0017)	NS (0.0821)
Random effects								
Between-class variance (Level 2)								
Intercept between-class variance	0.1003		0.0798		0.0839		0.0845	
Gender variance	0.0091		0.0049		0.0086		0.0071	
Math at T1 variance	0.0046		0.0033		0.0042		0.0032	
Correlation Intercept Gender	0.13		0.02		-0.06		-0.22	
Correlation Intercept Math at T1	0.32		0.32		0.22		0.36	
Correlation Gender Math at T1	-0.31		-0.22		-0.23		-0.51	
Within-class variance (Level 1)	0.3982		0.3986		0.4082		0.4195	
Deviance (-2 log L)	1158166.8		1344146.4		1423858.3		1493119.3	

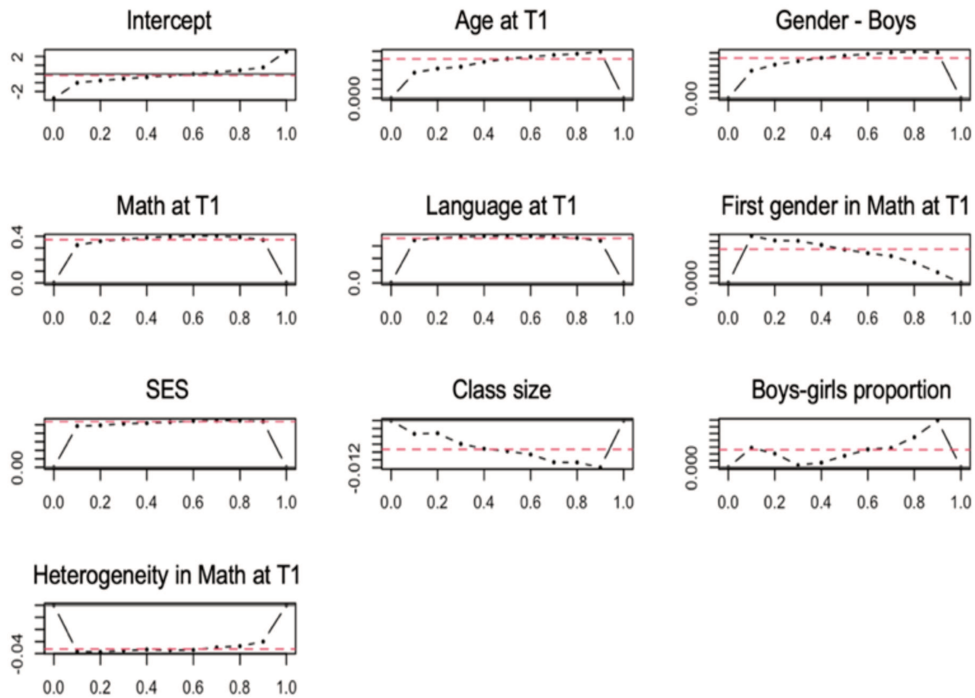
To estimate the distributional differences noted in the histograms (see **Figure 30-C**), we implemented a quantile regression. The **Figure 33** below showed the conditional quantile functions as a linear combination of the different predictors. The red line corresponded to the coefficients of the linear regression model, and the black dots represented the quantile regression weights for each variable of the model, where the effect was measured by decile of level in Math at T3, 0 being the lowest and 1 being the highest decile of level. The results showed that the highest the level in math at T3, the greater the gender gap was, in favor of boys. This result was in line with our previous findings and confirmed the growing advantage for boys at the highest math scores. With schooling, the highest positions were increasingly occupied by boys, to the detriment of girls. There was nevertheless a slight drop in the effect for the last decile, but this might be due to a slightly compressed tail of the distribution, as there was a ceiling effect for the distribution of T3 scores.

In addition, we implemented a growth model, using a growth framework with time nested in students nested in classes. First, we added a “time” variable (0, 4 and 12 months, respectively corresponding to T1, T2 and T3). We then entered this variable into several new multilevel models, capturing the z-scored gaussianized test scores in math following **model 1**: $\text{math} \sim \text{time} + \text{Age at T1} + \text{Gender} + \text{Gender} * \text{time} + \text{Gender} * \text{Age} + (1 + \text{time} \mid \text{Student level}) + (1 + \text{time} + \text{Gender} \mid \text{Class level})$ (see **Table S27** in SOM).

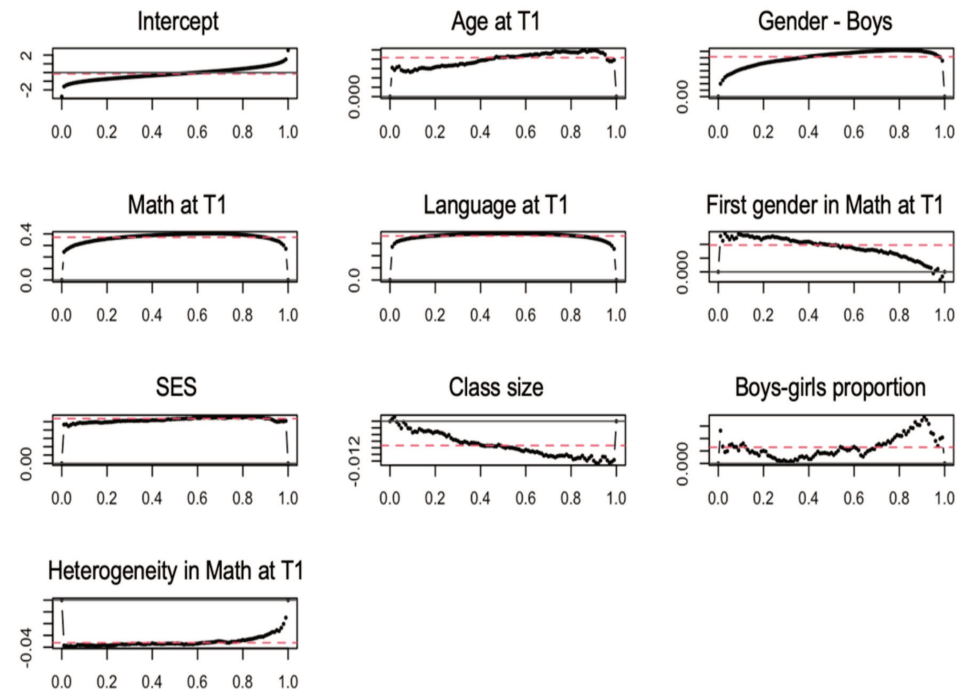
In this model, all variables were centered and reduced (mean = 0 and standard deviation = 1) and the time variable was added. It was quite understandable that the variable time has very little or no effect (Estimate -0.0102 , $p = 0.0001$), as the mean score in math, once expressed as a z-score of the entire population, did not evolve much with time. However, the interaction between time and gender was highly significant in this model (Estimate = 0.0202 , $p < 0.0001$), indicating that the gender gap was widening with time in school.

Figure 33. Quantile regression of math at T3 using math deciles (left) and math percentiles (right).

Quantile regression – Math level deciles at T3



Quantile regression – Math level in percentiles at T3



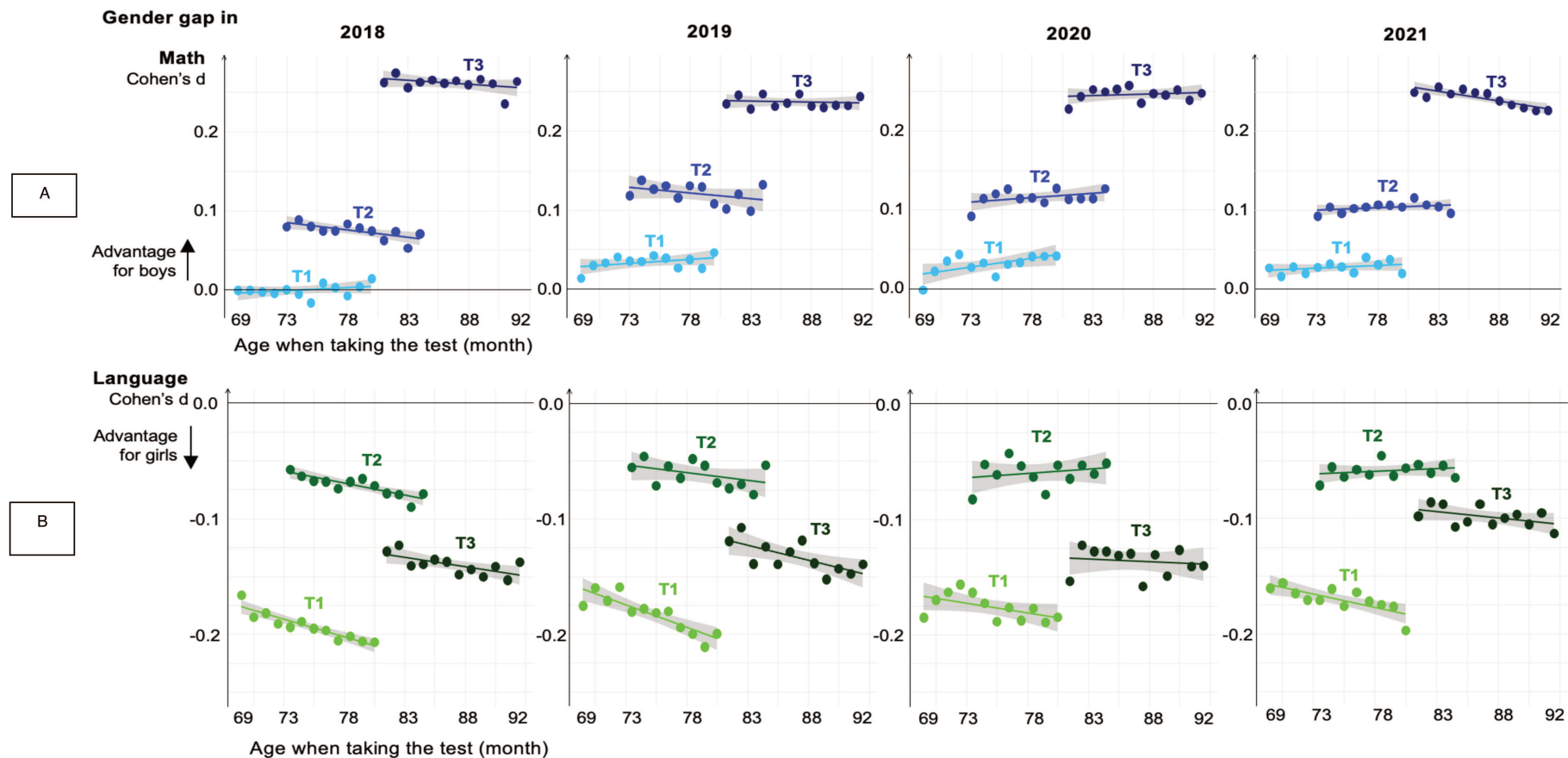
We also implemented a more complete multilevel model (i.e., **model 2**), including all variables assessed in the main regression model of Table 1 (except for Language at T1, as all girls and boys were not at an equal level in language). In addition, we assessed interactions between time with gender. Even with the inclusion of the later variables, we noted similar results regarding the interaction of time and gender in math compared to the previous model. All results converged towards a significant interaction between time and the gender effect ($p < 0.0001$) (see **Table S28** in SOM).

As time progresses, a difference between boys and girls emerged: the time \times gender interaction was positive and significant, meaning that the gender gaps grew with time. This interaction was modulated by several triple interactions: both the proportion of boys per class and a higher social position index increased the interaction between time and the gender gap, while having a boy at the top of the class at T1 tended to attenuate it.

B. A new element: age is not a modulator of the gender gap, school is.

Did schooling merely enhance pre-existing differences between children? Girls were ahead in all language tests at T1, a linguistic advantage that has been proposed to drive girls away from math (Breda & Napp, 2019) (**Figure 31-B** and **Table 21 and S18**). Even within math, there was some heterogeneity among tests, with girls performing slightly better at T1 in enumerating quantities, reading, and writing numbers, and problem solving, while boys performed better in comparing numbers and placing them on a number line (Table S6 in SOM).

Fig. 31. A new element: Age is not a modulator of the gender gap in math. (A) Decorrelation of age and schooling. This analysis took advantage of the fact that, in France, a strict cutoff on birth date determines school entry. Thus, children can have the same age (x axis) while varying in their level of schooling (colors and regression lines). Regardless of age when taking the test, children showed minimal or no gender gap in math at T1, but a growing effect after 4 (T2) and 12 months (T3) of schooling. Thus, mathematics gaps appear to grow rapidly after the exposure of formal math schooling and are not varying with age. Note that in 2019, between T2 and T3, schools were closed for an average of 52 days due to Covid, and the gap was correspondingly reduced while the overall performance increased (see ANOVA in Table S13-14 in **SOM**). The year 2020 was also disrupted, though to a lesser extent. **(B)** Distinct dynamics of the gender gap for language: girls are already in advance of boys at T1, an effect that widens with age and is only transiently reduced during the school year (T2), but largely restored after the summertime school break (T3).

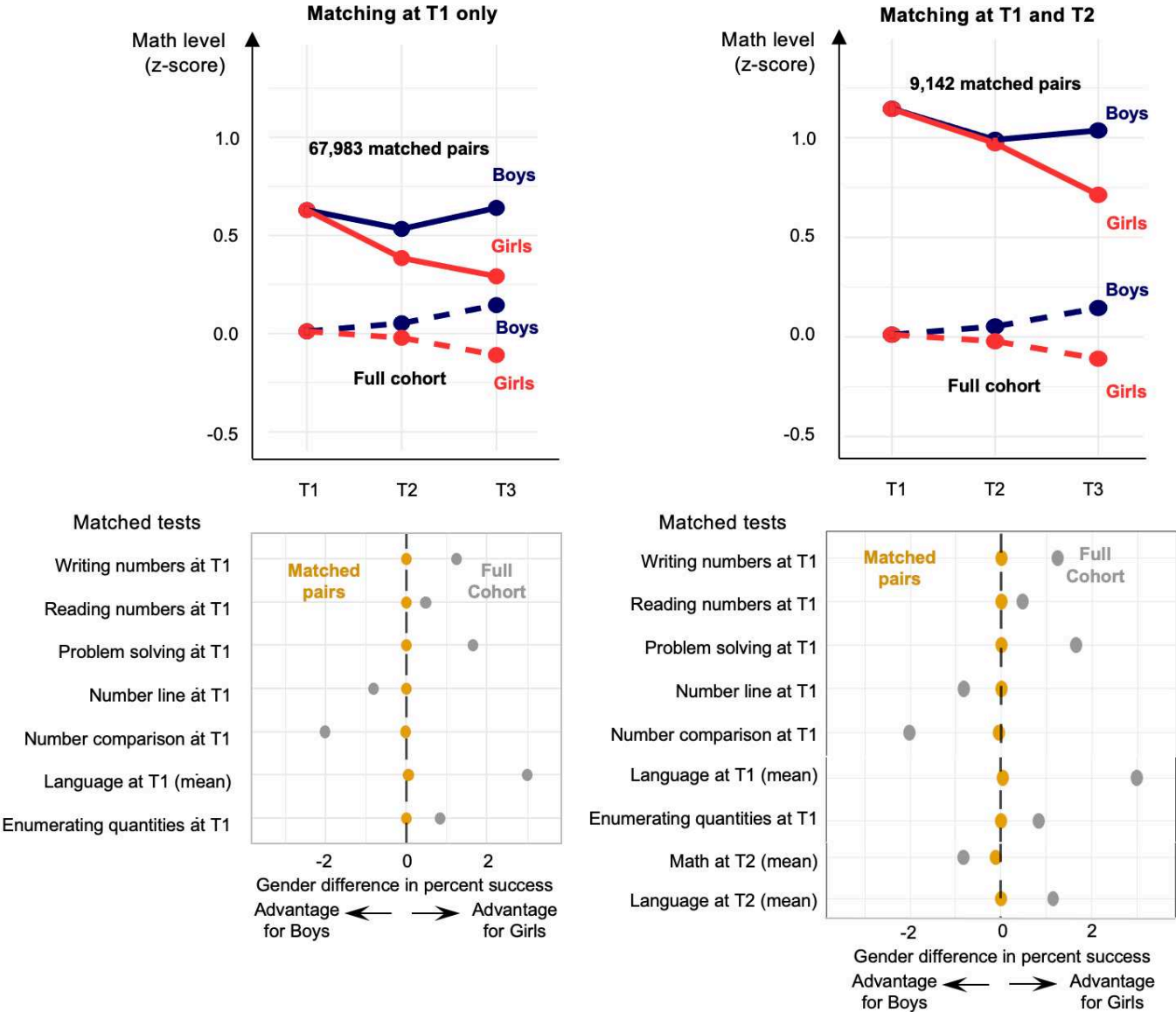


What you can see on **Figure 31** was that age has zero effect, or even a small protective effect, against the emergence of gender bias in math. Rather, for equal age (and we equalized many other variables in the paper), it is schooling (even just for 4 months [T2] or 12 months [T3]) that determined whether boys are ahead of girls in math. This is the crucial finding of this chapter. Schooling does not act like a rapid acceleration of the social stereotyping but rather, schooling *determines* the effect. These data provided a radical refutation of the classic idea that stereotypes were slowly infused. Schooling, not time, is the relevant variable.

C. Matching experiments: a massive gender gap remains after matching processes

To control for such differences, we implemented two matching scenarios: in the first scenario, 67,943 boys and girls were matched in pairs according to their results in every math test at T1, their mean language performance at T1, their SES score, their age and school category (**Figure 32, Table S24** in SOM). The second scenario matched 9,142 pairs according to the previous variables and the addition of their language level at T2 and their math level at T2. While both matching procedures eliminated any gender difference at T1, a massive gender gap remained at T2 and T3 (**Figure 32, Table S24**). Even when students were further matched on their T2 scores, a gender gap emerged at T3, indicating a cumulative influence of the entire school year. Six additional causal inference methods confirmed that gender was associated to a gap of around 5 points in math between T1 and T3, a very stable value over the different methods used and over time in 2018, 2019, 2020 and 2021 (**Table S25** in SOM).

Figure 32. A massive gender gap emerges and remains at T2 and T3 after the elimination of any gender difference at T1 when implementing matching procedures. Boys (in blue) and girls (in red) of typical age (69 to 80 months) were paired based on almost exact results in math assessments at T1, SES score, age, language at T1 and categories of schools (scenario 1 on the left) and additionally based on language at T2 and math at T2 (scenario 2 on the right). Even when students were matched on their T1 and T2 scores, the emergence of a gender gap at T2 and its widening at T3 is in line with a cumulative influence of the school exposure. Results were presented in 2018, and replicated in 2019, 2020 and 2021 in SOM.



Matching allows us to answer a precise question: do initial differences at T1 (or T1 and T2) suffice to explain the growing gender gap? By matching at T1 and T2, we can further exclude that the gender gap corresponds to a latent dimension that is revealed by schooling. Rather, even when no gap was observed at T2, a gap still emerged at T3, in fitting with our conclusion that it is the amount of exposure to school that drives it.

While these matching analyses indicated that schooling did not merely amplify preexisting differences between boys and girls, we also found that several variables modulated the math gender gap. T3 data were entered into a mixed-effect multilevel linear model with gender and its interactions with several potential modulators: SES, class size, T1 performance in language and math, class heterogeneity in math, boy-girl ratio, and gender of the top student in the class (**Table 22**). The gender gap emerged at all levels of these variables (**Figure S14** in SOM) and was larger for younger pupils with a higher initial math level and a lower initial language level. At the class level, the gender gap tended to be larger when the first-of-class was a boy (i.e., the role model effect – **Table 22** and **Figure S14**), SES was higher, and class size was larger (although the latter effects were not significant in every year, see Table 1). Taken together, these findings were compatible with our hypothesis: at school, girls and boys must decide how much effort to invest in the difficult domain of mathematics, and they do so more readily if they are already advanced in math or can identify with the first of class, and less so if they are already more advanced in language and reading (Breda & Napp, 2019).

D. Language gender gaps present different dynamics than math

Importantly, language performance followed strikingly different dynamics than math (**Figure 31** and **Figure S17**). In language, a gender gap favoring girls was already present and important at T1 (Cohen's $d = -0.1935$), was reduced at T2 (Cohen's $d = -0.0845$) and widened again at T3 (Cohen's $d = -0.1371$) (**Table 18** and **S26**). Controlling for differences at T1, the gender gap effect on language at T3 was ~10 times smaller than on math at T3 ($\beta_{\text{gender gap language T3}} = -0.0328 (\pm 0.0039) ***$, $\beta_{\text{gender gap math T3}} = 0.3453 (\pm 0.0038) ***$, **Table 22**, **Table S20**). Thus, in language, an early, sustained female advantage existed which, unlike math, was not drastically altered by

schooling but rather boys' disadvantage in language diminished with school exposure (Breda & Napp, 2019; OECD, 2015) (**Figure 31**). All in all, school is more beneficial to boys – who progressed both in math and language – compared to girls.

E. Reproducibility of the results from 2018 to 2022

The effects described above were largely reproducible across four consecutive years (see SOM) and afforded one more test of the exposure to school thanks to the natural experiment of Covid-19. Between T2 and T3 of the 2019 cohort, the pandemic-induced disruption caused French elementary school to close for 52 consecutive days, followed by the 2-month summer vacation. Interestingly, the gender gap in math was significantly lower during this period compared to the previous years (**Figure 31** and **Table S26**). A similar but smaller reduction was seen in 2020, where some schools again were closed. No such reduction was seen for the language gender gap confirming that it is driven by other factors (**Figure 31** and **Table S26**). All in all, these results were consistent with the hypothesis that a reduced exposure to school was associated with a gender gap decrease in math.

F. Covid-19: A natural experiment of a lower exposure to school

Another difference between cohorts was that, in 2019, 2020 and 2021, but not in 2018, a small but significant math gender gap favoring boys already existed at T1 (**Figure 31**; only in math, not language, see Table S26). During those years, some kindergarten teachers started to introduce formal exercises in math and language to prepare their children for the first-grade national assessments, which were first introduced in 2018. This pedagogical change, which made kindergarten more similar to elementary school, may have had the undesirable effect of inducing gender stereotypes at an earlier age, as shown elsewhere (Cimpian et al., 2016).

IV) Discussion

In summary, while previous cross-sectional international studies concluded that the math gender gap appeared around age 8-9 or 4th grade (Bharadwaj et al., 2016; Campbell et al., 2021), the present longitudinal results indicate a much earlier emergence, in agreement with earlier results on smaller samples (Bharadwaj et al., 2016; Campbell et al., 2021).

Thanks to massive data, we discovered a rapidly induced and entrenched gender gap in math favoring boys, after only 4 months of schooling in 1st grade, not related to children's age and emerging earlier among higher SES backgrounds. Math thus differs drastically from language, where large differences favoring girls exist prior to schooling, develop homogeneously among SES backgrounds and are linear with children's age.

Crucially, the present study elucidates the conditions for the emergence of the math gender gap, which does not reflect pre-existing gender differences, nor does it require a lengthy period of internalization. Gender stereotypes are not internalized slowly nor as a function of age (i.e., younger children tend to internalize gender stereotypes faster, particularly in math, see gender gaps negative slopes for math on **Figure 31**). Rather, the math gender gap significantly emerges and deepens after exposure to the formal math teaching and a higher schooling exposure duration. Teachers' attitudes and formal math education may play an important role, if they interact differently with boys and girls, transmit their math anxiety to girls (Beilock et al., 2010; Bharadwaj et al., 2016; Cimpian et al., 2016; Contini et al., 2017; Fischer & Thierry, 2021; Gunderson et al., 2012; Nollenberger et al., 2016; Robinson & Lubienski, 2011) or encourage girls' efforts at reading more than at math (Cimpian et al., 2016).

However, the onset of schooling also may prompt a change in the attitudes of children themselves, parents, family members and other professionals (del Río et al., 2019; OECD, 2015; Voyer & Voyer, 2014). The simple belief that boys and girls have different interests and abilities can reinforce gender disparities (Gunderson et al., 2012; Nollenberger et al., 2016). Last but not least, girls may exhibit greater math anxiety

and therefore avoid competition, a behavior that may explain why, among all math and language exercises, the male advantage is more pronounced for the more challenging, novel or complex tests tapping executive functions (Buser et al., 2014; Van Mier et al., 2019).

In addition, as gender gaps in math sped up and deepened in favor of boys and as gender gaps were reduced in language in favor of girls, the French school benefits more to boys than to girls by improving the math level of boys and by improving the level of boys in language from T1 to T3.

Our work goes way beyond the most impressive study to date, Cimpian et al. 2016 (Cimpian et al., 2016) which bears on 20,000 American students from 1999 and 2011, but is not at all a random sample of the US population as they state explicitly, for instance, that “The ECLS-K includes an oversample of Asian/Pacific Islander children” and offer various weighing and subsampling schemes to counter such problems (see (National Center for Education Statistics, 2011)). In addition, regarding the possible country specificity of our “French-centered” results, the closest findings are by Cimpian et al (2016), which suggest that an equally fast infusion of biases may occur in the USA, although unlike ours, they do not dissociate age from schooling. However, there is substantial evidence that country-specific factors may play an important role indeed.

In the latest PISA study, 21 countries managed to reduce the math gender gap between 2009 and 2018, and in 5 countries this was achieved thanks to improvements in girls’ level in math (OECD, 2018a). The present findings suggest that interventions should come early in the curriculum. Cross-national sociocultural, political, and educational equality in adults does not necessarily predict a reduced math gender gap (Bharadwaj et al., 2016; Campbell et al., 2021; OECD, 2018a).

Are gender gaps in math similar (in strength and direction) worldwide?

The answer is no, both the international assessments of students' math and science abilities, namely TIMSS (which evaluates 10- and 14-year-olds) and PISA (focused on

15-year-olds' language and math skills), did not find consistent advantage for boys over girls in math performance across all countries in the Organisation for Economic Co-operation and Development (OECD) (Miller et al., 2018; Miller & Halpern, 2014; Mullis et al., 2009, 2012).

In the 2018 PISA assessments, it was observed that on the one hand, boys outperformed girls in math in 32 countries, on the other hand, girls outperformed boys in 14 countries, and there was no significant difference in mathematical achievement between boys and girls in the other OECD countries (OECD, 2019c). Furthermore, the extent and direction of these gender disparities (whether they favor boys or girls) varied across countries worldwide and were often influenced by political decisions, such as policies aimed at enhancing mathematical and language abilities in either boys or girls (OECD, 2019c). This overall pattern suggested that there was no inherent or unavoidable gender disparity in mathematical achievements (OECD, 2019c).

Comparing the exact same tests between 2009 and 2018 in PISA allowed to identify that 21 out of 64 countries managed to reduce the math gender gap between 2009 and 2018: in 5 countries this was achieved thanks to improvements in girls' level in math (OECD, 2019c). Similar trends were also identified in TIMSS data of 2015 (i.e., 4th and 8th graders are assessed in math), which examined gender differences in various aspects of mathematics, including preferences for mathematics, confidence in mathematics, and the importance placed on mathematics. These findings revealed, on the one hand, that the variations of disparities in gender gaps in math between countries could not be supported by biological explanations between boys and girls and on the other hand, their findings identified that there was a shift between 4th and 8th grade in the mathematical affects of children, where girls were losing interest in math (Ghasemi & Burley, 2019).

At what time do gender gaps in math in favor of boys emerge in different countries?

Studies showed that the gender gap more than triples from age 9-10 to age 15-16 (Bharadwaj et al., 2016; Borgonovi et al., 2021; Contini et al., 2017). In UK, using a data set from 1969 to 2003, a gender gap emergence in favor of boys in math has

been identified by the end of primary school, at age 11, with a magnitude of ~ 0.8 to 2.7 percentile points on average, and raising up to ~ 3 to 7 percentile points at age 16 (Machin & McNally, 2005). From 2003 to 2013, boys were 4 percentile points more likely to excel in math compared to girls, and 8 percentile points to achieve the higher standards at age 11 in the UK (Cavaglia et al., 2020).

When analyzing the US dataset “Early Childhood Longitudinal Study, Kindergarten Class of 1998– 1999” (ECLS-K), the math gender gap was identified as early as the end of kindergarten and increased with age (Cimpian et al., 2016). While the gender gap did not exist at the beginning of primary school, boys performed more than 2-tenths of a standard deviation better than girls in math by the end of the 6th year of primary school (Fryer Jr. & Levitt, 2010). The math gender gap was higher for top performing students. Initially boys appeared to do better than girls among well performers and worse at the bottom of the distribution; However, by third grade, the gender gap, while still larger at the top, appeared throughout the distribution (Penner & Paret, 2008). Gender gaps at the top of the distribution were substantial: in the fall of kindergarten, girls made up only 20% of students above the 99th percentile in math (Robinson & Lubienski, 2011).

On an exhaustive population survey of 4th and 8th graders in Chile (i.e., SIMCE national scores), the gaps nearly doubled by 8th grade (i.e., 0.08 and 0.2 of a standard deviation, respectively for 4th and 8th grade). Furtherly, a larger gap in favor of boys was found among the top 5% performers, with ratios of boys to girls of ~ 2 for both grades 4 and 8 (Bharadwaj et al., 2016). All in all, the gender gap in favor of boys in math is accelerated at the beginning of first grade when exposed to formal math teaching, and kindergarten and preschool focused research should explore the factors associated with this raise of gender gap in mathematical concepts to have more precise information about both when this gender gap emerges and why.

Furthermore, the latest PISA assessments of 2018 showed that 15-year-old girls’ performance in literacy exceeded that of boys in every country, and the gap was considerable. For science and math, the gender gap was much smaller and varied from country to country (OECD, 2015, 2018a). In UK, using a data set from 1969 to

2003, a gender gap emergence in favor of boys in math has been identified by the end of primary school, at age 11, with a magnitude of .79 to 2.71 percentile points on average, and raising up to 3 to 6.8 percentile points at age 16 (Machin & McNally, 2005). At age 11 in the UK, boys were 4 percentile points more likely to excel in math compared to girls, and 8 percentile points to achieve the higher standards (Cavaglia et al., 2020).

From a policy perspective, tackling the gender gap in mathematics at the earliest stage (kindergarten or 1st grade) may be more cost-efficient and effective, as it comes before girls lose confidence in their math abilities and become resistant to counter-stereotypic information (Huguet & Régner, 2009). Which factors should be targeted? Our findings suggest that class-level variables such as class size, gender ratio, heterogeneity in math level, or gender of the student at the top of class only have a small influence. Single-sex schools or classes also are ineffective (Lee et al., 2014; Miller & Halpern, 2014).

The most important intervention may be to convince all children that math is worth the effort for both genders. Past research suggests that the following actions may be effective: supporting parents, informing them, and encouraging the development of a stimulating home learning environment (Gunderson et al., 2012; Melhuish et al., 2008; Miller & Halpern, 2014); encouraging both genders to play similar games for spatial thinking (Levine et al., 2016); encouraging teachers' gender-fair ratings and practices (Cimpian et al., 2016), such as questioning girls and boys equally often during math and science courses (Miller & Halpern, 2014); exposing children to both male and female role models with whom they can identify (Stout et al., 2011); providing girls with means to cope with competitive stress (Buser et al., 2014, 2021) and math anxiety (D'Agostino et al., 2022); and informing them about the possible impact of stereotype threats in math (Johns et al., 2005; Miller & Halpern, 2014); emphasizing the role of effort, perseverance, and an incremental view of intelligence in efficient learning (Alan & Ertac, 2019; Yeager et al., 2019); and implementing self-affirmation tasks to protect girls from stereotype threat (Johns et al., 2005; Miyake et al., 2010). More generally, the present findings should enhance societal awareness of the absence of gender disparities in mathematical ability prior to the onset of school math learning and their

rapid emergence when formal teaching of mathematics begins, non-correlated with age. Such awareness is a prerequisite to efforts, by parents as well as teachers, to encourage their children equally to build on their aptitude for learning.

V) Supplementary materials

Table S18. Percent success in each test for all four cohorts (2018, 2019, 2020 and 2021), separately for each gender, among children of normal age at T1. For each subtest, the results of the gender with superior performance were highlighted in bold. Results were normalized and presented in percentage of success in the domain (range 0-100).

	2018			2019			2020			2021		
	Boys	Girls	p	Boys	Girls	p	Boys	Girls	p	Boys	Girls	p
n	288,587	281,184		336,978	328,654		350,960	344,489		367,099	355,131	
Age in first grade (month)	74.45 (3.42)	74.43 (3.43)	0.011	74.42 (3.42)	74.40 (3.43)	0.023	74.46 (3.43)	74.44 (3.43)	0.007	74.43 (3.42)	74.42 (3.42)	0.331
Class size	17.24 (5.86)	17.20 (5.86)	0.010	17.02 (5.77)	17.00 (5.77)	0.078	16.93 (5.71)	16.91 (5.72)	0.138	17.43 (5.84)	17.43 (5.85)	0.672
SES score	102.53 (17.76)	102.42 (17.79)	0.023	102.77 (17.83)	102.59 (17.85)	< 0.0001	102.78 (17.86)	102.64 (17.91)	0.002	103.51 (18.42)	103.42 (18.49)	0.039
Oral comprehension of words at T1	78.82 (18.40)	80.24 (17.40)	< 0.0001	77.11 (19.56)	78.50 (18.70)	< 0.0001	76.34 (20.25)	77.72 (19.34)	< 0.0001	76.75 (20.81)	78.26 (19.83)	< 0.0001
Oral comprehension of sentences at T1	84.99 (16.20)	88.88 (14.04)	< 0.0001	86.13 (15.51)	89.63 (13.48)	< 0.0001	85.21 (16.51)	89.01 (14.39)	< 0.0001	84.87 (17.19)	88.69 (14.87)	< 0.0001
Oral comprehension of texts at T1	72.71 (20.45)	75.36 (19.55)	< 0.0001	73.88 (20.85)	76.59 (19.89)	< 0.0001	73.43 (21.70)	76.36 (20.64)	< 0.0001	73.65 (21.99)	76.43 (20.93)	< 0.0001
Phoneme manipulation at T1	58.36 (25.50)	60.87 (24.50)	< 0.0001	60.24 (23.40)	62.64 (22.48)	< 0.0001	59.14 (23.97)	61.63 (23.02)	< 0.0001	61.53 (24.60)	63.82 (23.59)	< 0.0001
Syllable manipulation at T1	75.41 (21.22)	78.61 (19.70)	< 0.0001	77.19 (22.84)	80.19 (21.17)	< 0.0001	75.41 (24.07)	78.52 (22.41)	< 0.0001	76.38 (24.08)	79.40 (22.21)	< 0.0001
Letter-sound association at T1	72.79 (26.47)	76.31 (24.29)	< 0.0001	76.49 (25.94)	79.53 (23.57)	< 0.0001	74.62 (27.25)	77.57 (25.05)	< 0.0001	76.10 (27.36)	78.93 (25.19)	< 0.0001
Recognizing letters at T1	66.59 (27.23)	70.17 (24.90)	< 0.0001	65.36 (27.05)	68.78 (24.89)	< 0.0001	63.83 (27.98)	67.28 (25.97)	< 0.0001	67.50 (27.33)	70.82 (25.14)	< 0.0001
Comparing letters at T1	62.01 (27.92)	65.19 (27.69)	< 0.0001	82.72 (23.17)	85.33 (21.77)	< 0.0001	81.50 (24.08)	84.34 (22.63)	< 0.0001	81.35 (24.30)	84.44 (22.36)	< 0.0001
Oral comprehension of sentences at T2	86.48 (13.87)	88.62 (12.59)	< 0.0001	86.47 (14.04)	88.53 (12.72)	< 0.0001	86.16 (14.32)	88.35 (12.88)	< 0.0001	86.63 (14.64)	88.80 (13.18)	< 0.0001

Reading words at T2	25.79 (18.00)	24.32 (16.52)	< 0.0001	26.57 (18.28)	24.89 (16.66)	< 0.0001	27.55 (18.73)	25.91 (17.24)	< 0.0001	31.10 (21.91)	29.54 (20.39)	< 0.0001
Reading texts at T2	15.58 (13.34)	15.15 (12.97)	< 0.0001	15.63 (13.10)	15.01 (12.52)	< 0.0001	16.31 (13.56)	15.64 (12.94)	< 0.0001	19.04 (15.96)	18.44 (15.39)	< 0.0001
Writing syllables at T2	78.75 (24.25)	81.09 (22.34)	< 0.0001	80.18 (24.01)	82.34 (22.07)	< 0.0001	81.09 (23.66)	83.13 (21.80)	< 0.0001	79.45 (29.35)	81.64 (27.81)	< 0.0001
Writing words at T2	73.04 (28.17)	75.77 (26.54)	< 0.0001	74.01 (28.43)	76.71 (26.71)	< 0.0001	75.51 (27.98)	77.97 (26.39)	< 0.0001	77.63 (27.69)	80.29 (25.70)	< 0.0001
Phoneme manipulation at T2	77.07 (23.17)	78.84 (22.19)	< 0.0001	78.24 (21.46)	80.04 (20.28)	< 0.0001	78.93 (21.34)	80.83 (19.97)	< 0.0001	77.24 (25.07)	79.21 (23.95)	< 0.0001
Letter-sound association at T2	93.62 (13.27)	94.62 (11.71)	< 0.0001	94.79 (12.23)	95.57 (10.52)	< 0.0001	94.86 (12.36)	95.70 (10.56)	< 0.0001	90.59 (23.02)	91.63 (21.96)	< 0.0001
Oral comprehension of words at T3	89.04 (13.69)	90.21 (12.64)	< 0.0001	88.39 (14.49)	89.62 (13.47)	< 0.0001	88.89 (14.13)	90.12 (13.10)	< 0.0001	88.46 (14.63)	89.69 (13.48)	< 0.0001
Oral comprehension of sentences at T3	91.22 (11.33)	93.14 (9.94)	< 0.0001	91.07 (11.78)	93.05 (10.43)	< 0.0001	91.02 (11.71)	92.97 (10.33)	< 0.0001	90.53 (12.36)	92.64 (10.79)	< 0.0001
Writing syllables at T3	79.83 (24.14)	82.30 (22.36)	< 0.0001	77.24 (26.33)	79.80 (24.83)	< 0.0001	80.62 (23.83)	83.07 (22.20)	< 0.0001	80.55 (23.96)	82.66 (22.45)	< 0.0001
Writing words at T3	70.66 (26.38)	74.25 (24.36)	< 0.0001	66.99 (28.19)	70.85 (26.50)	< 0.0001	70.60 (26.52)	74.22 (24.66)	< 0.0001	69.29 (26.74)	71.95 (25.30)	< 0.0001
Reading comprehension of sentences at T3	76.93 (25.03)	79.82 (22.96)	< 0.0001	75.22 (26.91)	78.11 (25.08)	< 0.0001	77.67 (24.60)	80.39 (22.62)	< 0.0001	78.65 (24.81)	80.63 (23.17)	< 0.0001
Reading comprehension of texts at T3	72.77 (25.73)	77.59 (24.08)	< 0.0001	70.49 (27.18)	75.37 (25.73)	< 0.0001	72.95 (26.02)	77.65 (24.34)	< 0.0001	72.62 (26.34)	76.57 (24.79)	< 0.0001
Reading words at T3	47.50 (20.57)	46.74 (20.05)	< 0.0001	45.98 (21.65)	45.30 (21.23)	< 0.0001	49.03 (20.72)	48.28 (20.28)	< 0.0001	49.25 (21.38)	47.96 (20.96)	< 0.0001
Reading texts at T3	37.31 (22.63)	38.39 (23.00)	< 0.0001	35.80 (23.80)	36.70 (24.20)	< 0.0001	38.71 (23.15)	39.55 (23.51)	< 0.0001	34.18 (23.32)	33.70 (23.13)	< 0.0001
Writing numbers at T1	92.92 (15.58)	94.16 (13.87)	< 0.0001	95.32 (12.82)	96.13 (11.38)	< 0.0001	95.01 (13.48)	95.76 (12.10)	< 0.0001	95.84 (12.43)	96.59 (10.94)	< 0.0001
Reading numbers at T1	96.79 (10.02)	97.26 (9.03)	< 0.0001	96.79 (10.41)	97.25 (9.29)	< 0.0001	96.23 (11.60)	96.69 (10.62)	< 0.0001	96.87 (10.54)	97.21 (9.71)	< 0.0001
Problem solving at T1	62.85 (30.48)	64.49 (29.50)	< 0.0001	68.13 (28.23)	68.72 (27.35)	< 0.0001	66.75 (28.96)	67.39 (27.92)	< 0.0001	68.09 (28.62)	68.77 (27.64)	< 0.0001
Enumerating quantities at T1	93.27 (14.82)	94.11 (13.64)	< 0.0001	92.11 (16.87)	93.15 (15.48)	< 0.0001	91.42 (17.93)	92.50 (16.58)	< 0.0001	92.37 (16.90)	93.28 (15.74)	< 0.0001

Comparing numbers at T1	39.44 (24.25)	37.42 (23.45)	< 0.0001	56.41 (28.50)	53.86 (27.79)	< 0.0001	55.96 (29.17)	53.85 (28.49)	< 0.0001	57.17 (29.03)	55.12 (28.39)	< 0.0001
Number line at T1	51.65 (31.25)	50.82 (29.91)	< 0.0001	53.80 (31.61)	52.40 (30.55)	< 0.0001	53.56 (32.22)	52.02 (31.07)	< 0.0001	52.82 (33.39)	51.63 (32.13)	< 0.0001
Comparing numbers at T2	92.06 (18.12)	91.39 (18.26)	< 0.0001	79.11 (24.33)	77.51 (24.16)	< 0.0001	79.08 (24.31)	77.71 (24.08)	< 0.0001	79.08 (24.26)	77.83 (24.15)	< 0.0001
Number line at T2	55.55 (25.21)	53.29 (24.04)	< 0.0001	58.56 (25.79)	55.82 (24.71)	< 0.0001	59.35 (25.89)	56.55 (24.89)	< 0.0001	62.25 (26.53)	59.36 (25.72)	< 0.0001
Writing numbers at T2	90.62 (18.25)	89.24 (18.58)	< 0.0001	92.19 (17.06)	90.74 (17.48)	< 0.0001	92.83 (16.48)	91.58 (16.84)	< 0.0001	93.70 (16.14)	92.73 (16.51)	< 0.0001
Problem solving at T2	69.41 (28.15)	68.58 (27.81)	< 0.0001	69.57 (27.98)	68.56 (27.55)	< 0.0001	70.64 (27.79)	69.64 (27.37)	< 0.0001	70.18 (30.27)	69.61 (29.77)	< 0.0001
Addition at T2	82.55 (23.90)	82.12 (23.38)	< 0.0001	79.99 (25.05)	78.61 (24.82)	< 0.0001	80.81 (24.76)	79.74 (24.43)	< 0.0001	83.79 (23.71)	83.23 (23.18)	< 0.0001
Subtraction at T2	72.81 (35.09)	73.45 (34.83)	< 0.0001	70.67 (33.15)	68.88 (32.65)	< 0.0001	72.28 (32.36)	70.59 (31.90)	< 0.0001	71.81 (33.85)	70.38 (33.52)	< 0.0001
Geometry at T3	72.46 (21.35)	73.16 (19.27)	< 0.0001	74.47 (21.38)	75.24 (19.17)	< 0.0001	74.69 (21.07)	75.37 (18.91)	< 0.0001	74.76 (21.34)	75.33 (19.25)	< 0.0001
Number line at T3	50.45 (25.25)	44.25 (22.57)	< 0.0001	50.15 (25.79)	43.80 (23.11)	< 0.0001	51.73 (25.48)	45.11 (22.93)	< 0.0001	67.43 (31.09)	62.07 (30.89)	< 0.0001
Addition at T3	67.97 (30.99)	58.20 (30.90)	< 0.0001	73.71 (29.17)	65.96 (29.45)	< 0.0001	76.22 (27.89)	68.86 (28.42)	< 0.0001	75.23 (27.91)	68.49 (28.10)	< 0.0001
Subtraction at T3	51.46 (34.20)	43.66 (31.15)	< 0.0001	55.66 (35.23)	49.20 (32.97)	< 0.0001	59.44 (34.57)	53.11 (32.57)	< 0.0001	57.28 (35.78)	51.14 (33.27)	< 0.0001
Mental calculation at T3	83.88 (20.61)	85.50 (19.29)	< 0.0001	83.80 (20.91)	85.40 (19.73)	< 0.0001	84.33 (20.37)	86.04 (18.99)	< 0.0001	84.67 (20.11)	86.12 (18.93)	< 0.0001
Writing numbers at T3	85.10 (23.39)	81.80 (24.39)	< 0.0001	84.45 (24.31)	81.00 (25.50)	< 0.0001	86.77 (22.11)	83.81 (23.28)	< 0.0001	85.60 (22.82)	80.32 (26.45)	< 0.0001
Reading numbers at T3	87.69 (19.15)	83.62 (20.45)	< 0.0001	86.96 (20.31)	82.97 (21.69)	< 0.0001	88.70 (18.32)	85.05 (19.75)	< 0.0001	88.55 (18.47)	84.41 (20.21)	< 0.0001
Problem solving at T3	69.82 (27.24)	67.02 (26.67)	< 0.0001	69.54 (27.73)	66.70 (27.26)	< 0.0001	71.12 (27.07)	68.17 (26.66)	< 0.0001	70.91 (26.96)	67.24 (26.86)	< 0.0001
Associating numbers and quantities at T3	61.46 (24.75)	60.26 (24.12)	< 0.0001	60.29 (25.57)	59.65 (24.79)	< 0.0001	61.64 (25.28)	60.94 (24.46)	< 0.0001	-	-	-
Language level at T1	71.46 (16.10)	74.45 (14.85)	< 0.0001	74.89 (15.82)	77.65 (14.52)	< 0.0001	73.69 (16.87)	76.55 (15.49)	< 0.0001	74.77 (17.18)	77.60 (15.69)	< 0.0001

Language level at T2	64.33 (14.21)	65.49 (13.07)	< 0.0001	65.13 (13.97)	66.15 (12.76)	< 0.0001	65.77 (14.07)	66.79 (12.83)	< 0.0001	65.95 (16.45)	67.08 (15.33)	< 0.0001
Language level at T3	70.66 (16.16)	72.81 (15.19)	< 0.0001	68.90 (17.57)	71.10 (16.77)	< 0.0001	71.19 (16.32)	73.28 (15.43)	< 0.0001	70.44 (15.97)	71.98 (15.16)	< 0.0001
Math level at T1	72.82 (13.83)	73.04 (12.88)	< 0.0001	77.09 (14.33)	76.92 (13.46)	< 0.0001	76.49 (15.14)	76.37 (14.21)	0.001	77.19 (14.61)	77.10 (13.77)	0.005
Math level at T2	77.17 (17.78)	76.35 (17.32)	< 0.0001	75.02 (18.83)	73.35 (18.31)	< 0.0001	75.83 (18.65)	74.30 (18.11)	< 0.0001	76.80 (18.56)	75.52 (18.04)	< 0.0001
Math level at T3	71.10 (18.24)	67.15 (17.19)	< 0.0001	72.34 (18.74)	68.78 (18.00)	< 0.0001	74.12 (17.75)	70.69 (17.02)	< 0.0001	75.55 (18.17)	71.89 (17.82)	< 0.0001

Gender * Heterogeneity of level at T1	-	-	-	-	-	-	-	-	-	-	-	-
Gender * Class size	-	-	-	-	-	-	-	-	-	-	-	-
Gender * Boys-Girls ratio per class	-	-	-	-	-	-	-	-	-	-	-	-
Gender * First of class is a boy in math at T1	-	-	-	-	-	-	-	-	-	-	-	-
Gender * SPI	-	-	-	-	-	-	-	-	-	-	-	-
Random effects												
Between-class variance												
Intercept variance	0.1553	0.1011	0.1561	0.1052	0.1012	0.1052						
Gender variance	-	-	0.0164	-	-	0.0107						
Math at T1 variance	-	0.0079	-	-	0.0080	-						
Correlation Intercept Gender	-	-	0.39	-	0.39	0.14						
Correlation Intercept Math T1	-	0.43	-	-	-	-						
Correlation Gender T1 Math	-	-	-	-	-	-						
Within-class variance	0.8351	0.4910	0.8143	0.4799	0.4740	0.4772						
Deviance (-2 log L)	1564075.7	1270071.2	1552911.3	1253150.3	1251340.8	1252985.6						

Models	Model 7 Math T3 ~ Math T1*Gender + (1 + Math T1 + Gender ID class)		Model 8 Math T3 ~ individual variables + (1 + Math T1 + Gender ID class)		Model 9 Math T3 ~ individual variables + collective variables + (1 + Math T1 + Gender ID class)		Model 10 Math T3 ~ individual variables + collective variables + interactions + (1 + Math T1 + Gender ID class)	
	Parameter estimate (Sd)	p	Parameter estimate (Sd)	p	Parameter estimate (Sd)	p	Parameter estimate (Sd)	p
Intercept	0.0091 (0.0019)	< 0.0001 ***	0.0106 (0.0018)	< 0.0001 ***	0.0182 (0.0019)	< 0.0001 ***	0.0107 (0.0019)	< 0.0001 ***
Math level at T1	0.6505 (0.0011)	< 0.0001 ***	0.6455 (0.0012)	< 0.0001 ***	0.6358 (0.0012)	< 0.0001 ***	0.3810 (0.0013)	< 0.0001 ***
Gender (Boys)	0.2523 (0.0019)	< 0.0001 ***	0.2553 (0.0019)	< 0.0001 ***	0.2532 (0.0020)	< 0.0001 ***	0.3285 (0.0018)	< 0.0001 ***
Language level at T1	-	-	-	-	-	-	0.4078 (0.0013)	< 0.0001 ***

Age at T1 (month)	-	-	0.0343 (0.0001)	< 0.0001 ***	0.0361 (0.0002)	< 0.0001 ***	0.0062 (0.0009)	< 0.0001 ***
SES score	-	-	-	-	0.0883 (0.0019)	< 0.0001 ***	0.0277 (0.0020)	< 0.0001 ***
First of class is a boy in math at T1	-	-	-	-	0.0099 (0.0018)	< 0.0001 ***	0.0063 (0.0019)	0.0008 **
Boys-Girls ratio per class	-	-	-	-	-0.0023 (0.0017)	NS (0.192)	0.0005 (0.0018)	NS (0.7742)
Class size	-	-	-	-	0.0001 (0.0019)	NS (0.940)	0.0095 (0.0020)	< 0.0001 ***
Heterogeneity of level in math at T1	-	-	-	-	-0.0522 (0.0017)	< 0.0001 ***	-0.0305 (0.0018)	< 0.0001 ***
Gender * Age at T1	-	-	-	-	-	-	-0.0094 (0.0019)	< 0.0001 ***
Gender * Math level at T1	0.0536 (0.0019)	< 0.0001 ***	-	-	-	-	0.0644 (0.0024)	< 0.0001 ***
Gender * Language level at T1							-0.0065 (0.0024)	0.0075 **
Gender * Heterogeneity of level at T1	-	-	-	-	-	-	-0.0048 (0.0018)	0.0081 **
Gender * Class size	-	-	-	-	-	-	0.0043 (0.0020)	0.0275 *
Gender * Boys-Girls ratio per class	-	-	-	-	-	-	-0.0010 (0.0020)	NS (0.6124)
Gender * First of class is a boy in math at T1	-	-	-	-	-	-	0.0064 (0.0019)	0.0006 **
Gender * SES score	-	-	-	-	-	-	0.0049 (0.0020)	0.0146 *
Random effects								
Between-class variance								
Intercept variance	0.1012		0.1008		0.0878		0.1003	
Gender variance	0.0109		0.0106		0.0107		0.0091	
Math at T1 variance	0.0079		0.0079		0.0080		0.0046	
Correlation Intercept Gender	0.15		0.11		0.06		0.13	
Correlation Intercept Math T1	0.40		0.39		0.36		0.32	
Correlation Gender T1 Math	-0.27		-0.40		-0.39		-0.31	
Within-class variance	0.4807		0.4711		0.4713		0.3982	
Deviance (-2 log L)	1251112.5		1250806.8		1247356.8		1158166.8	

Table S20. Multilevel regression model for Language at T3 among children of normal age at T1.

The formula implemented was as follow: *Language at T3 ~ Age at T1 + Gender + Math level at T1 + language level at T1 + First of class being a boy in language + SES score at T1 + Class size + Boys-Girls ratio per class + Heterogeneity of level in language in the class + 8 interactions between each variable and Gender + (1 + Gender + Language level at T1 | class)*

Variables	Language at T3							
	2018		2019		2020		2021	
N	569,771		665,632		695,449		722,230	
N group (classes)	39,573		46,671		49,010		49,701	
Fixed effects	Parameter estimates (sd)	p	Parameter estimates (sd)	p	Parameter estimates (sd)	p	Parameter estimates (sd)	p
Intercept	0.0459 (0.0026)	< 0.0001	0.0606 (0.0029)	< 0.0001	0.0454 (0.0028)	< 0.0001	0.0272 (0.0029)	< 0.0001
Language individual level at T1	0.5759 (0.0015)	< 0.0001	0.5764 (0.0018)	< 0.0001	0.5684 (0.0018)	< 0.0001	0.5175 (0.0019)	< 0.0001
Math individual level at T1	0.1622 (0.0013)	< 0.0001	0.1451 (0.0017)	< 0.0001	0.1559 (0.0017)	< 0.0001	0.1672 (0.0018)	< 0.0001
Gender (Boys)	-0.0328 (0.0039)	< 0.0001	-0.0337 (0.0036)	< 0.0001	-0.0386 (0.0035)	< 0.0001	-0.0168 (0.0037)	< 0.0001
SES score at T1	0.0667 (0.0020)	< 0.0001	0.1137 (0.0020)	< 0.0001	0.0709 (0.0020)	< 0.0001	0.0716 (0.0019)	< 0.0001
Age at T1 (month)	-0.0043 (0.0003)	< 0.0001	-0.0045 (0.0004)	< 0.0001	-0.0014 (0.0003)	0.0001	-0.0002 (0.0004)	0.4927
Class size	-0.0029 (0.0020)	0.1464	-0.0047 (0.0019)	0.0157	-0.0076 (0.0019)	0.0001	-0.0014 (0.0019)	0.4583
First of class is a boy in language at T1	-0.0006 (0.0019)	0.7532	0.0004 (0.0018)	0.8440	-0.0001 (0.0018)	0.9503	-0.0006 (0.0018)	0.7293
Boys – Girls ratio per class at T1	0.0011 (0.0018)	0.5429	-0.0027 (0.0018)	0.1390	-0.0053 (0.0018)	0.0029	0.0024 (0.0018)	0.1665

Heterogeneity of level in language at T1	-0.0234 (0.0018)	< 0.0001	-0.0125 (0.0017)	< 0.0001	-0.0120 (0.0017)	< 0.0001	-0.0104 (0.0017)	< 0.0001
Gender * Language at T1	-0.0070 (0.0025)	0.0059	-0.0030 (0.0024)	0.2011	-0.0054 (0.0024)	0.0229	-0.0076 (0.0024)	0.0019
Gender * Math at T1	0.0021 (0.0025)	0.3926	-0.0015 (0.0023)	0.5075	0.0048 (0.0023)	0.0402	0.0026 (0.0024)	0.2798
Gender * SES score	-0.0061 (0.0021)	0.0031	-0.0102 (0.0019)	< 0.0001	-0.0121 (0.0019)	< 0.0001	-0.0109 (0.0019)	< 0.0001
Gender * Age at T1	0.0006 (0.0005)	0.2764	0.0001 (0.0005)	0.9137	0.0001 (0.0005)	0.8688	0.0001 (0.0005)	0.7818
Gender * Class size	0.0046 (0.0020)	0.0221	0.0032 (0.0018)	0.0779	0.0033 (0.0018)	0.0767	0.0051 (0.0019)	0.0059
Gender * First of class is a boy in Language at T1	0.0107 (0.0019)	< 0.0001	0.0003 (0.0017)	0.8509	0.0047 (0.0017)	0.0061	0.0059 (0.0018)	0.0010
Gender * Boys-Girls ratio per class	-0.0041 (0.0020)	0.0392	0.0022 (0.0018)	0.2195	-0.0031 (0.0018)	0.0811	-0.0062 (0.0019)	0.0009
Gender * Heterogeneity of level in language at T1	0.0051 (0.0018)	0.0058	0.0033 (0.0017)	0.0505	-0.0008 (0.0017)	0.6403	0.0034 (0.0017)	0.0492
Random effects								
Between-class variance (Level 2)								
Intercept variance	0.0990		0.0803		0.0803		0.0749	
Gender variance	0.0060		0.0044		0.0057		0.0061	
Language at T1 variance	0.0083		0.0055		0.0085		0.0136	
Correlation Intercept Gender	-0.11		-0.25		-0.22		-0.18	
Correlation Intercept Language at T1	0.24		0.18		0.08		-0.08	
Correlation Gender Language at T1	0.09		0.18		0.21		0.02	
Within-class variance (Level 1)	0.4268		0.4229		0.4242		0.4775	
Deviance (-2 log L)	1196878.7		1381396.8		1449135.7		1587511.5	

Table S22. Multilevel regression model for Math at T1 among children of normal age at T1.

The formula implemented was as follow: *Math at T1 ~ Age at T1 + Gender + SES score at T1 + Class size + Boys-Girls ratio per class + Gender * Age at T1 + Gender * SES score + Gender * Class size + Gender * Boys-Girls ratio per class + (1 | class)*

Cohort	Math at T1							
	2018		2019		2020		2021	
N	569,771		665,632		695,449		722,230	
Classes	39,573		46,671		49,010		49,701	
Fixed effects	Estimate (sd)	p	Estimate (sd)	p	Estimate (sd)	p	Estimate (sd)	p
Intercept	-0.3449 (0.0033)	< 0.0001	-0.3519 (0.0038)	< 0.0001	-0.3403 (0.0037)	< 0.0001	-0.3330 (0.0037)	< 0.0001
Age at T1 (month)	0.0551 (0.0003)	< 0.0001	0.0541 (0.0005)	< 0.0001	0.0522 (0.0004)	< 0.0001	0.0520 (0.0004)	< 0.0001
Gender (Boys)	-0.0044 (0.0050)	0.3832	0.0190 (0.0047)	0.0001	0.0135 (0.0046)	0.0031	0.0216 (0.0045)	< 0.0001
SES score at T1	0.2666 (0.0024)	< 0.0001	0.2674 (0.0024)	< 0.0001	0.2905 (0.0024)	< 0.0001	0.2552 (0.0023)	< 0.0001
Class size	-0.0779 (0.0025)	< 0.0001	-0.0746 (0.0025)	< 0.0001	-0.0767 (0.0024)	< 0.0001	-0.0642 (0.0024)	< 0.0001
Boys-Girls ratio per class	-0.0084 (0.0022)	0.0001	-0.0019 (0.0023)	0.3906	-0.0009 (0.0022)	0.6736	-0.0048 (0.0022)	0.0313
Gender * Age at T1	0.0010 (0.0007)	0.1577	0.0018 (0.0006)	0.0056	0.0022 (0.0006)	0.0005	0.0008 (0.0006)	0.1749
Gender * SES score	0.0257 (0.0026)	< 0.0001	0.0172 (0.0024)	< 0.0001	0.0171 (0.0024)	< 0.0001	0.0135 (0.0023)	< 0.0001
Gender * Class size	-0.0088 (0.0026)	0.0007	-0.0014 (0.0024)	0.5701	-0.0047 (0.0024)	0.0481	-0.0003 (0.0023)	0.9013
Gender * Boys-Girls ratio per class	0.0008 (0.0025)	0.7573	0.0030 (0.0024)	0.2094	0.0030 (0.0023)	0.1861	0.0023 (0.0023)	0.3125
Between-class variance (Level 2)								
Intercept variance	0.1512		0.1266		0.1278		0.1349	
Within-class variance (Level 1)	0.7530		0.7629		0.7540		0.7603	
Deviance (-2 log L)	1508033.3		1766520.4		1838940		1916132.3	

Table S23. Multilevel regression model for Math at T2 among children of normal age at T1.

The formula implemented was as follow: *Math at T2 ~ Age at T1 + Gender + Math level at T1 + language level at T1 + First of class being a boy in math + SES score at T1 + Class size + Boys-Girls ratio per class + Heterogeneity of level in math in the class + 8 interactions between each variable and Gender + (1 + Gender + Math level at T1 | class)*

Cohort	Math at T2							
	2018		2019		2020		2021	
N	569,771		665,632		695,449		722,230	
N group (classes)	39,573		46,671		49,010		49,701	
Fixed effects	Estimate (sd)	p	Estimate (sd)	p	Estimate (sd)	p	Estimate (sd)	p
Intercept	-0.0155 (0.0026)	< 0.0001	-0.1280 (0.0028)	< 0.0001	-0.1192 (0.0027)	< 0.0001	-0.0993 (0.0028)	< 0.0001
Language level at T1	0.3937 (0.0013)	< 0.0001	0.3807 (0.0016)	< 0.0001	0.3726 (0.0016)	< 0.0001	0.3782 (0.0017)	< 0.0001
Math level at T1	0.3838 (0.0014)	< 0.0001	0.4090 (0.0017)	< 0.0001	0.4216 (0.0017)	< 0.0001	0.3882 (0.0017)	< 0.0001
Gender (Boys)	0.1530 (0.0038)	< 0.0001	0.1792 (0.0034)	< 0.0001	0.1625 (0.0033)	< 0.0001	0.1480 (0.0034)	< 0.0001
Age at T1 (month)	0.0037 (0.0003)	< 0.0001	0.0072 (0.0003)	< 0.0001	0.0066 (0.0003)	< 0.0001	0.0049 (0.0003)	< 0.0001
First of class is a boy in math at T1	0.0062 (0.0019)	0.0012	0.0033 (0.0019)	0.0740	0.0023 (0.0018)	0.1966	0.0045 (0.0018)	0.0147
Boys-Girls ratio per class	-0.0033 (0.0018)	0.0662	-0.0064 (0.0018)	0.0004	-0.0049 (0.0017)	0.0045	-0.0028 (0.0018)	0.1236
Class size	-0.0014 (0.0020)	0.4908	0.0015 (0.0020)	0.4386	0.0045 (0.0019)	0.0173	0.0038 (0.0019)	0.0443
SES score	-0.0028 (0.0020)	0.1602	-0.0196 (0.0020)	< 0.0001	-0.0377 (0.0019)	< 0.0001	-0.0220 (0.0019)	< 0.0001
Heterogeneity of level in math at T1	-0.0346 (0.0018)	< 0.0001	-0.0098 (0.0018)	< 0.0001	-0.0093 (0.0017)	< 0.0001	-0.0070 (0.0017)	0.0001
Gender * Language level at T1	-0.0179 (0.0025)	< 0.0001	-0.0214 (0.0022)	< 0.0001	-0.0252 (0.0022)	< 0.0001	-0.0251 (0.0022)	< 0.0001
Gender * Math level at T1	0.0211 (0.0025)	< 0.0001	0.0381 (0.0022)	< 0.0001	0.0346 (0.0022)	< 0.0001	0.0367 (0.0022)	< 0.0001
Gender * Age at T1	-0.0006 (0.0005)	0.2464	-0.0010 (0.0005)	0.0247	0.0004 (0.0005)	0.3632	0.0006 (0.0005)	0.1740
Gender* First of class is a boy in math at T1	-0.0053 (0.0018)	0.0041	-0.0027 (0.0016)	0.0980	-0.0024 (0.0016)	0.1351	-0.0034 (0.0016)	0.0384
Gender * Boys-Girls ratio per class	-0.0001 (0.0020)	0.9491	0.0008 (0.0017)	0.6328	-0.0015 (0.0017)	0.3731	-0.0024 (0.0017)	0.1708
Gender * Class size	-0.0021 (0.0020)	0.2948	-0.0014 (0.0017)	0.4177	0.0001 (0.0017)	0.9494	0.0012 (0.0017)	0.4684
Gender * SES score	0.0148 (0.0020)	< 0.0001	0.0089 (0.0018)	< 0.0001	0.0135 (0.0018)	< 0.0001	0.0055 (0.0018)	0.0018
Gender * Heterogeneity in math at T1	-0.0024 (0.0018)	0.1918	-0.0029 (0.0016)	0.0651	-0.0030 (0.0016)	0.0549	-0.0010 (0.0016)	0.5493

Random effects				
Between-class variance Intercept	0.1026	0.0910	0.0833	0.0956
Class level I Gender	0.0041	0.0032	0.0036	0.0053
Class level I T1 Math	0.0068	0.0034	0.0041	0.0094
Correlation Class level intercept I Gender	-0.02	-0.23	-0.20	-0.14
Correlation Class level intercept I T1 Math	0.18	-0.01	0.01	0.25
Correlation Gender I T1 Math	0.15	-0.05	-0.13	0.04
In-between-class variance (residuals)	0.4121	0.3751	0.3759	0.4012
Deviance (-2 log L)	1177958.9	1308348.5	1366670.4	1474527.4

Table S24. Matching experiments and scenarii

	Matching at T1 only	Matching at T1 and T2
School category, private, public, PE, HPE	Exact	Exact
SES score, 50-150	Same decile	Same decile
Age at T1, 69-80 (months)	+/- 4 months	+/- 4 months
6 tests in math at T1, 0-100	+/- 5 points	+/- 5 points
Language at T1, mean, 0-100	+/- 5 points	+/- 5 points
Math at T2, mean, 0-100	-	+/- 5 points
Language at T2, mean, 0-100	-	+/- 5 points
Number of matched pairs in 2018	67,983 pairs	9,142 pairs
Estimate of the gender effect at T3, Percent (SE) in 2018	5.156 (0.059) ***	4.296 (0.128) ***
Number of matched pairs in 2019	94,279 pairs	17,338 pairs
Estimate of the gender effect at T3, Percent (SE) in 2019	4.456 (0.049) ***	3.257 (0.089) ***
Number of matched pairs in 2020	96,777 pairs	19,448 pairs
Estimate of the gender effect at T3, Percent (SE) in 2020	4,209 (0.046) ***	3.195 (0.079) ***
Number of matched pairs in 2021	106,878 pairs	21,350 pairs
Estimate of the gender effect at T3, Percent (SE) in 2021	4.263 (0.046) ***	3.099 (0.077) ***

Table S25. Results of causal inference methods applied to the gender gap in Math between T1 and T3. CI = 95% confidence interval.

	2018			2019			2020			2021		
	Average gender effect in math between T3 and T1 (Percent of success, 0-100)	Lower CI	Upper CI	Average gender effect in math between T3 and T1 (Percent of success, 0-100)	Lower CI	Upper CI	Average gender effect in math between T3 and T1 (Percent of success, 0-100)	Lower CI	Upper CI	Average gender effect in math between T3 and T1 (Percent of success, 0-100)	Lower CI	Upper CI
G-computation (OLS)	5.088	4.888	5.287	4.729	4.551	4.907	4.419	4.252	4.587	4.634	4.464	4.805
Propensity weighted regression	5.496	5.400	5.593	4.892	4.799	4.985	4.616	4.530	4.701	4.753	4.667	4.839
IPW (logit)	5.106	4.729	5.483	4.597	4.241	4.952	4.294	3.939	4.650	4.513	4.158	4.868
AIPW (OLS & logit)	5.330	5.262	5.398	4.712	4.648	4.775	4.469	4.409	4.529	4.614	4.553	4.675
Random forest	5.293	5.227	5.360	4.801	4.739	4.864	4.563	4.504	4.622	4.779	4.718	4.839
TMLE (Target Maximum Likelihood Estimation) for causal inference	5.334	5.266	5.402	4.713	4.649	4.777	4.468	4.408	4.528	4.617	4.556	4.678

Table S26. T-tests measuring the differences of gender gaps magnitude between 2018, 2019, 2020 and 2021.

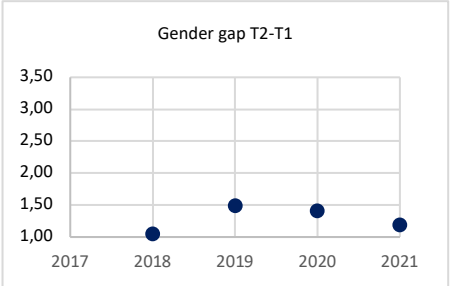
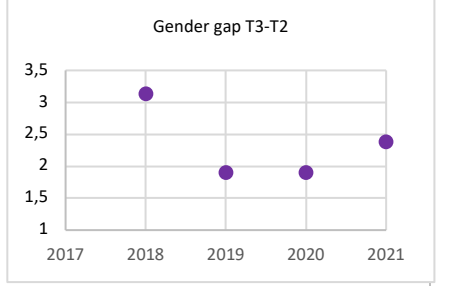
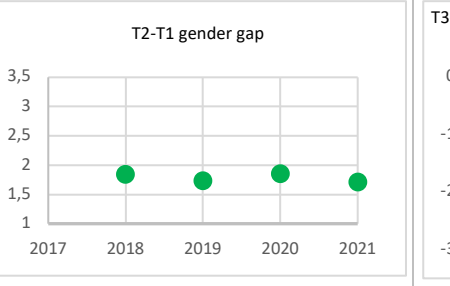
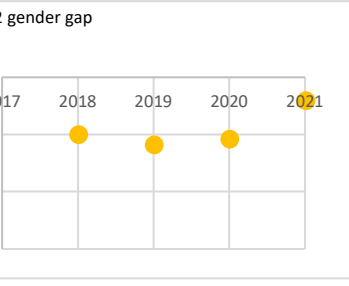
	Math				Language			
	Difference T2-T1 gender gap		Difference T3-T2 gender gap		Difference T2-T1 gender gap		Difference T3-T2 gender gap	
2018	1.04259		3.13147		1.84115		-0.99544	
2019	1.48566		1.89913		1.73366		-1.17600	
2020	1.40989		1.90267		1.84957		-1.07814	
2021	1.18361		2.38573		1.70562		-0.40885	
								
	Difference	p	Difference	p	Difference	p	Difference	p
2019 vs. 2018	0.44307	< 0.0001	-1.23234	< 0.0001	-0,10749	< 0.0001	-0,18056	< 0.0001
2020 vs. 2019	-0.07577	< 0.0001	0.00354	< 0.0001	0,11591	< 0.0001	0,09786	< 0.0001
2021 vs. 2020	-0.22628	< 0.0001	0.48306	< 0.0001	-0,14395	< 0.0001	0,66929	< 0.0001
2021 vs. 2018	0.14102	< 0.0001	-0.74574	< 0.0001	-0,13553	< 0.0001	0,58659	< 0.0001

Table S27. Fixed and random effects for model 1 in quantile regression of Math at T3 in 2018.

Fixed effects for Model 1

Y = Math	Estimate	Std. Error	p
<i>Intercept</i>	-0,0052	0,0027	0,0547
Time	-0,0102	0,0002	0,0000***
Age at T1	0,1676	0,0011	0,0000***
Gender	-0,0019	0,0027	0,4833
Gender * Time	0,0202	0,0003	0,0000***
Gender * Age	-0,0002	0,0015	0,8801

Random effects for Model 1

Groups	Name	Variance	Std.Dev.	
Student level	Intercept	4.692e-01	0.684970	
	time	1.627e-05	0.004034	
Groups	Name	Variance	Std.Dev.	Corr.
Class level	Intercept.	1.696e-01	0.411820	
	Gender Boy	6.952e-03	0.083377	0.31
	time	9.465e-04	0.030766	-0.53
0.15				
Residual		3.096e-01	0.556410	

Table S28. Fixed and random effects for model 2 in quantile regression of Math at T3 in 2018.

Fixed effects

Y = Math	Estimate	Std. Error	p
<i>Intercept</i>	0.0001	0.0026	0.9681
Time	-0.0101	0.0002	0.0000
Age at T1	0.1785	0.0010	0.0000
Gender	0.0010	0.0024	0.6813
First of class is a boy in Math at T1	-0.0873	0.0026	0.0000
SES	0.2429	0.0027	0.0000
Class size	-0.0712	0.0028	0.0000

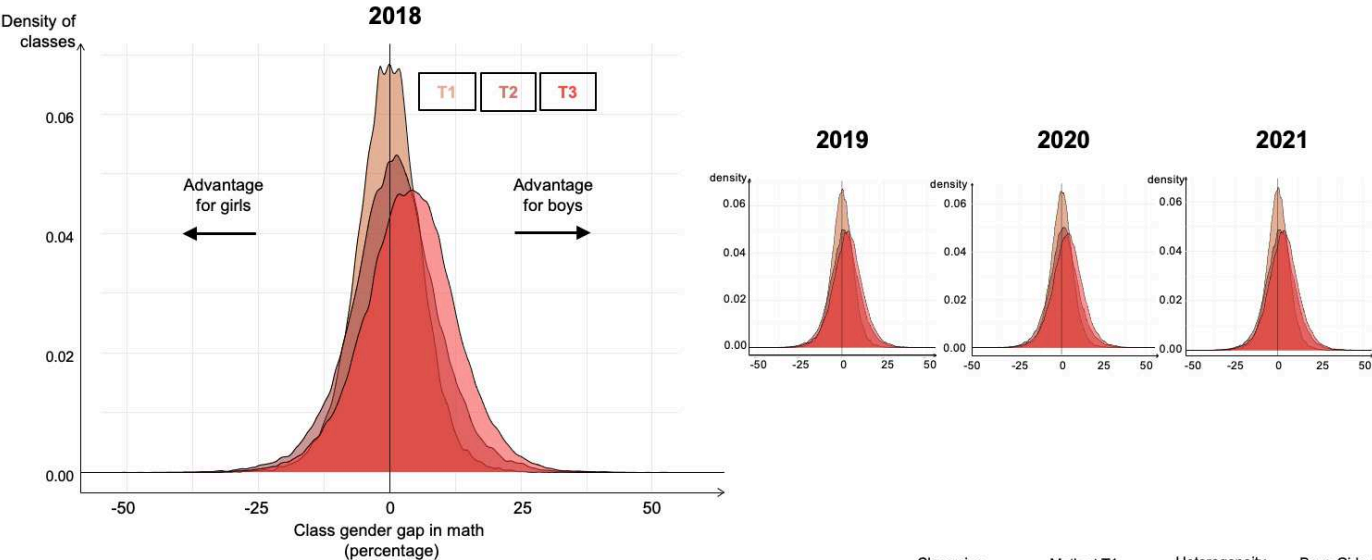
Boy proportion in class	0.0098	0.0025	0.0001
Heterogeneity of level in math at T1	0.0016	0.0024	0.5034
Time * Age	-0.0022	0.0001	0.0000
Time * Gender	0.0200	0.0003	0.0000
Time * First Boy	0.0032	0.0002	0.0000
Time * SES	-0.0003	0.0003	0.2804
Time * Class size	0.0012	0.0003	0.0000
Time * Heterogeneity	-0.0046	0.0002	0.0000
Gender * First Boy	0.1902	0.0024	0.0000
Gender * SES	0.0197	0.0026	0.0000
Gender * Class size	-0.0116	0.0026	0.0000
Gender * Boy proportion	-0.0373	0.0024	0.0000
Gender * Heterogeneity	-0.0067	0.0020	0.0010
Time * Boy Proportion	-0.0002	0.0002	0.3596
Time * Gender * First Boy	-0.0051	0.0003	0.0000
Time * Gender * SES	0.0009	0.0003	0.0041
Time * Gender * Class size	0.0004	0.0003	0.2034
Time * Gender * Boy proportion	0.0014	0.0003	0.0000

Random effects

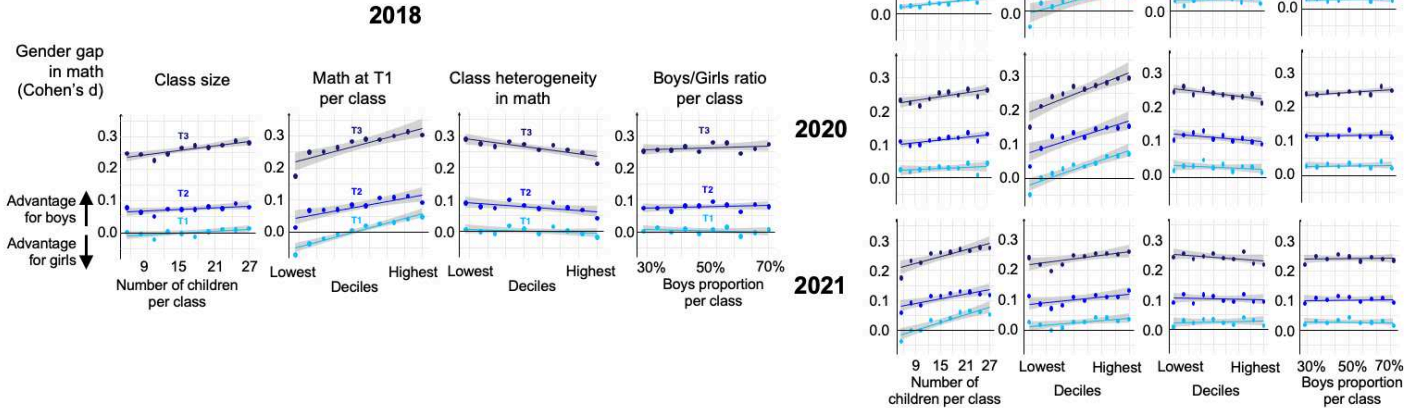
Groups	Name	Variance	Std.Dev.	Corr
Student level	time	0.0030030	0.05480	
Class level	Intercept	0.1604464	0.40056	
	Gender	0.0714369	0.26728	-0.26
	time	0.0004477	0.02116	-0.63
0.04				
Residual		0.6174566	0.78578	

Figure S14. Gender gap explanatory class-level covariates on children with typical age in first grade and among classes with at least 30% of boys and 30% of girls. (A) Density over classrooms of the average math gender gap. The distribution was centered on zero at T1, but many classrooms showed a bias at T2 and especially at T3. Results were similar for 2018, 2019, 2020 and 2021. **(B) Robustness of the gender gap (averaged within each classroom) to variations in class size, class's initial level in math, class's heterogeneity of level in math, and boys-girls ratio per class.** A higher heterogeneity of level in math was associated with a lower gender gap in favor of boys whereas a higher-class level in math was associated with a higher gender gap in favor of boys. Boys and girls ratio per class did not effect on gender gaps in math. **(C) Role model effect on the gender gap.** Having a girl or a boy as the first of class in math at T1 had a small impact on the gender gap. For this analysis, means for boys and girls within a class were computed while excluding the data from the best student(s) at T1, who are supposed to be role models and while focusing the analysis on typical-age children only. Classes with girls as first of class in math had a lower gender gap in favor of boys. Results were similar in 2018, 2019, 2020 and 2021.

A. Classes' gender gaps (Density)



B. Class size, math at T1 per class, heterogeneity, boys-girls ratio effects on gender gaps in math



C. Gender gaps in math when boys or girls are first of class

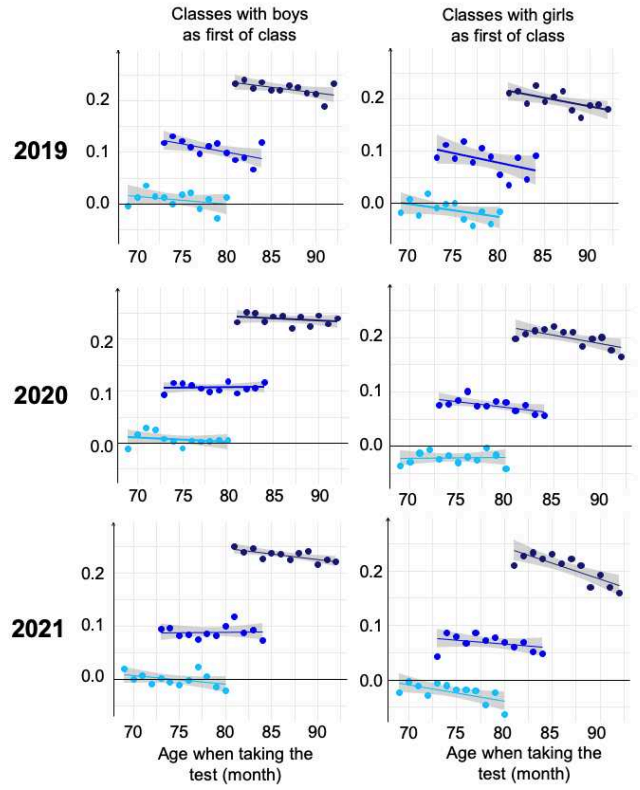
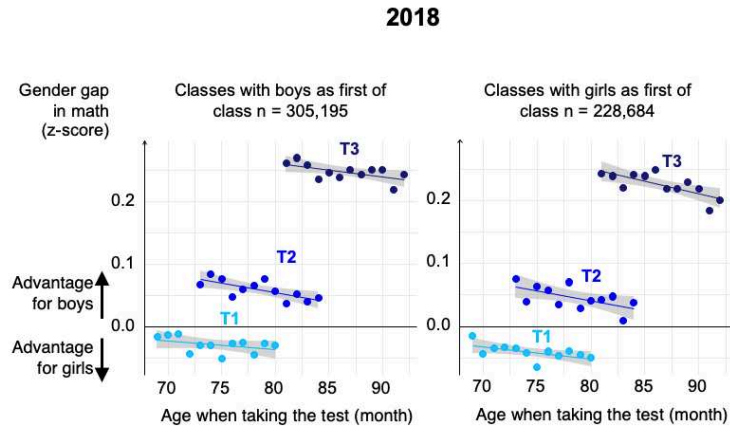
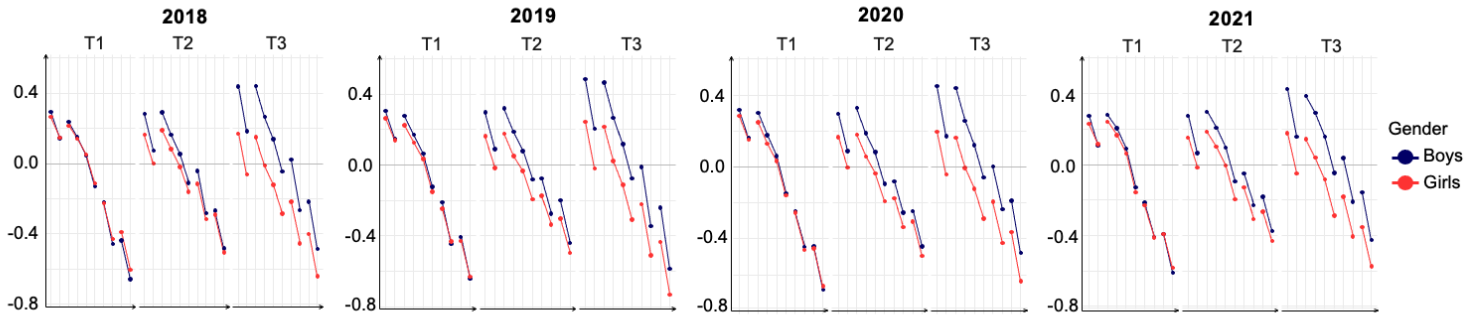
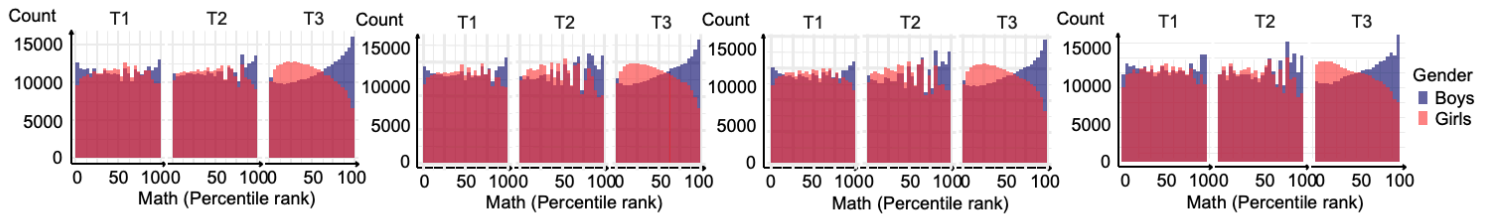


Figure S15. Math panels in 2018, 2019, 2020 and 2021 for figures 30 and 32. (A) Gender gap in math in function of school categories. (B) Boys and girls distribution in math in percentile ranks. (C) Matching in math at T1 only (left) and in math at T1 and T2 (right). Results were similar in 2018, 2019, 2020 and 2021.

A. Gender gap in math in function of school categories



B. Boys and girls distribution in math (percentile rank)



C. Matching in math at T1 only (left) and in math at T1 and T2 (right) for every year

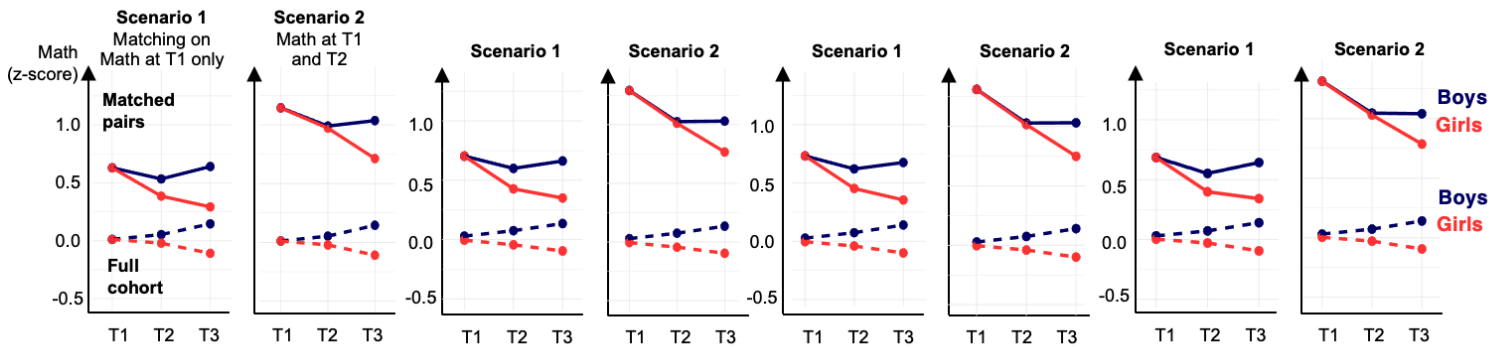
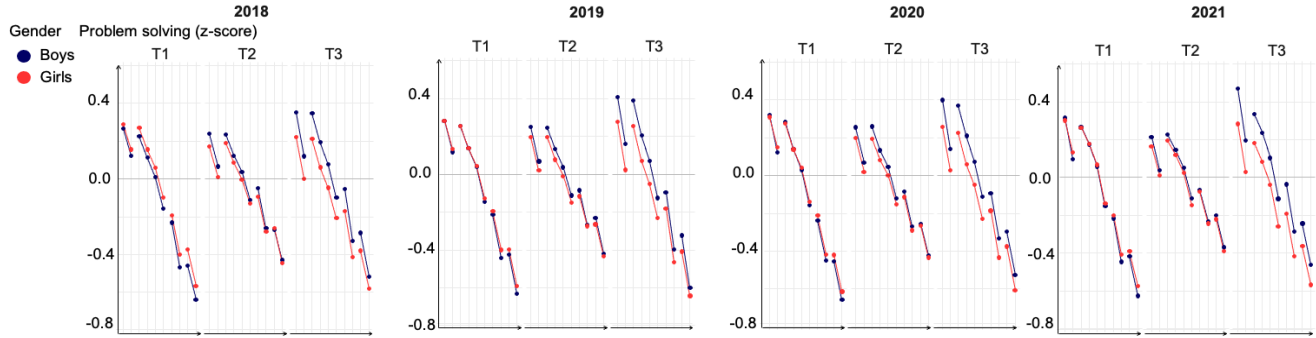
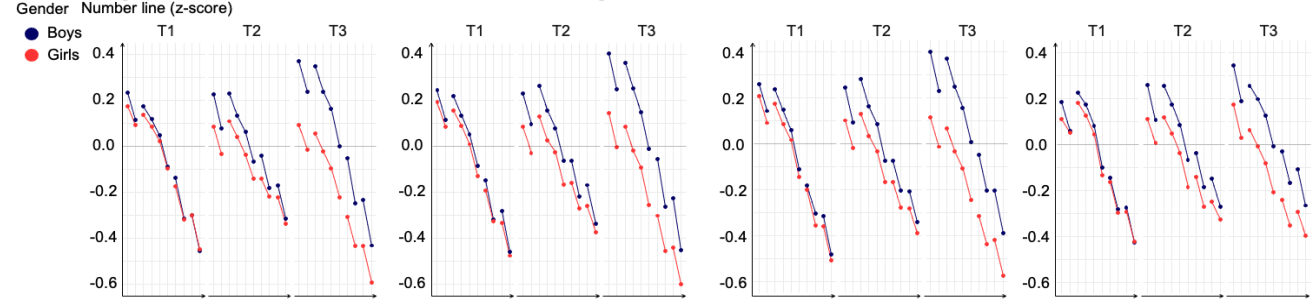


Figure S16. Panels for number line and problem solving in 2018, 2019, 2020 and 2021. Performance of boys (blue) and girls (red) in mathematics, in (A) problem-solving and (B) number-line assessments. Within each school category, the gender gap was almost null at school start (T1), detectable after 4 months (T2), and large after one year of schooling (T3), except for higher SES score school categories, where the gender gap was already in favor of boys. Gender gaps in function of age were presented for (C) problem-solving and (D) number line.

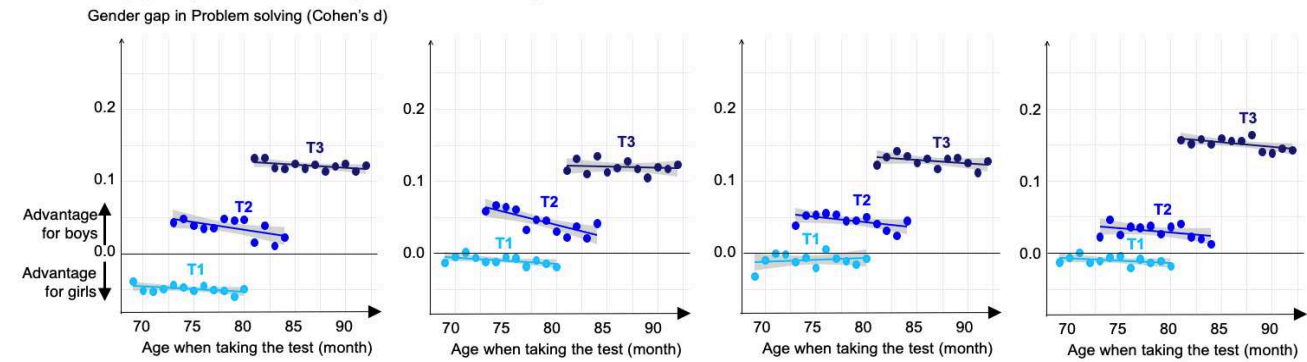
A. Gender gap in problem solving in function of school categories



B. Gender gap in number line in function of school categories



C. Gender gap in problem solving in function of age



D. Gender gap in number line in function of age

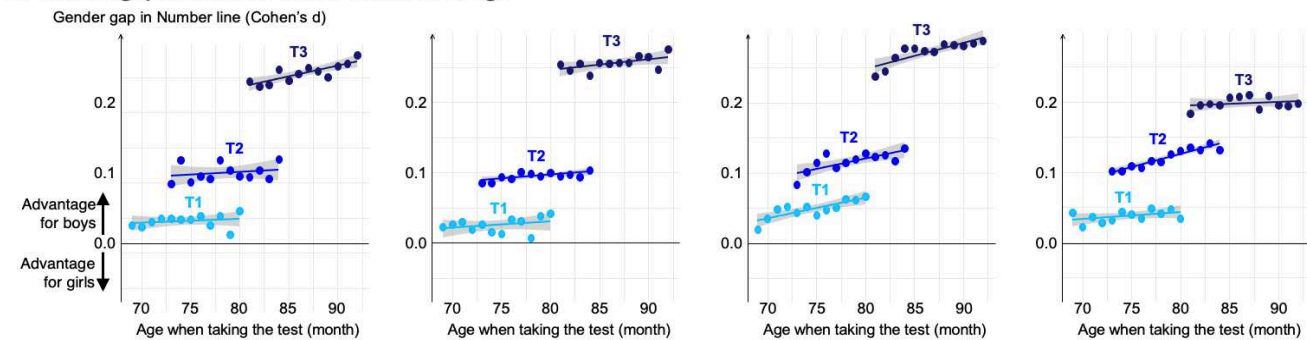
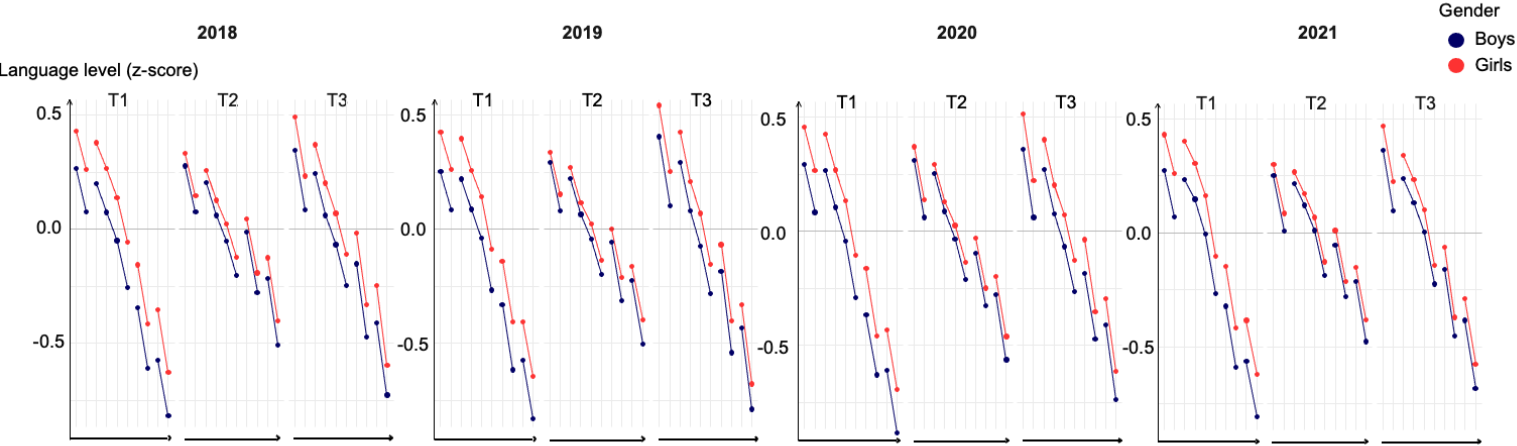
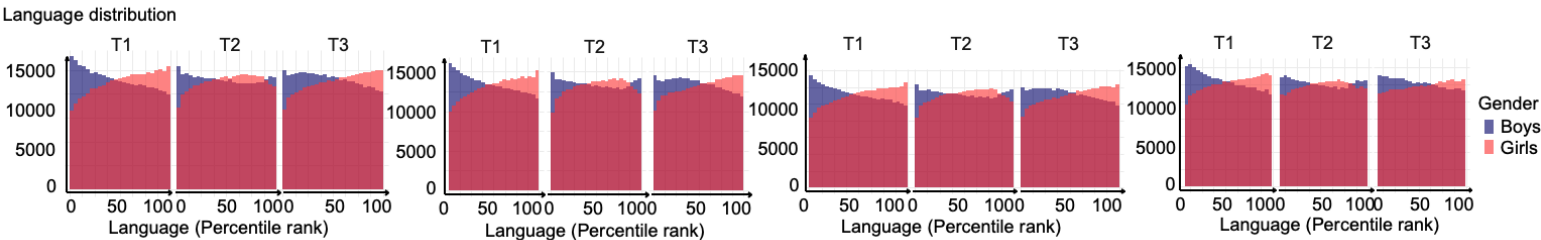


Figure S17. Language panels in 2018, 2019, 2020 and 2021. Distinct dynamics of the gender gap in language: girls were already in advance of boys at T1, an effect that widened with age and was only transiently reduced during the school year (T2), but largely restored after the school break (T3). Data results replicated in 2018, 2019, 2020 and 2021. In 2019, a slight gender gap reduction happened between T2 and T3 and, in 2020, a significant gender gap reduction happened between T2 and T3. This situation was restored to the initial gender gap dynamic in 2021 compared to 2018 (see **Table S26**).

A. Gender gap in language in function of school categories



B. Boys' and girls' distribution in language (percentile rank)



Chapter 5. Perspectives and Conclusion

In recent years, research in the field of developmental psychology and education has shed light on the challenges many children face in these areas during their initial years of formal schooling around learning to read and math. To our knowledge, this study is quite unique in the quality of its dataset, and representativity. In sharp contrast to previous educational studies that test only a small, often unrepresentative sample of a country's school population, the analyses of individual language and math assessments, class-level, and school level information of ~ 3 million first graders in France for four years in a row, allowed us to answer the following specific questions.

First, we were able to detect the fine correlations between specific cognitive language assessments, notably, refining the learning pathways of typical-age children in first grade. We were able to enrich the reading model of Gough (i.e., decoding x oral vocabulary comprehension = reading comprehension) with precise predictive weights of factors in first grade and sharing hints about children' learning needs to perform better in second grade and reduced socioeconomical gaps in reading acquisition. Among the fundamental skills measured in first grade, we now know a wider range of language assessments predicted reading comprehension: syllable handling in first grade was paramount, followed by oral comprehension of words, phoneme handling and letter-sound association. In addition, we confirmed that reading words in one minute (and not reading comprehension) was predicted by different skills in first grade, principally by letter-sound association, and followed by phoneme handling, letter knowledge and syllable handling.

Second, taking the example of reading comprehension, we showed how precise and specific the analyses could be for isolating and defining subgroups of children experiencing similar difficulties and needed similar learning reinforcement (i.e., depending on their age, their gender, their type of school, and their strengths or weaknesses in language domains). In this process of identifying the learning needs of children, we showed one way to use these massive data inside a classroom, by

educators, for adapting their teaching strategy to the needs of specific subgroups inside their classroom. In parallel, this process also allows teachers to detect the early signs of difficulties in language and therefore to intervene the earliest in the children academic pathways. Further exploration of every type of difficulty as an outcome would be informing in order to plan learning reinforcement sessions adapted to every child's needs, both in language and in math subdomains (e.g., exploring the characteristics of children with difficulties in phoneme awareness).

Third, our study confirmed that lower SES children presented more needs in oral language comprehension compared to the other children, and also presented with different specific groups of learning needs (i.e., some were facing large difficulties in oral language comprehension and performed well in other domains, others performed worse in meta phonology but well in oral language comprehension, some presented with both difficulties). For instance, among PE and HPE public schools, even though children were similar in SES and school types, age or gender, they represented a heterogeneous group with diverse language reinforcement needs (i.e., three distinct groups were found when analyzing all language assessments at T1) rather than forming one homogeneous group with "difficulties in language".

Also, these analyses allowed to identify that in function of age, the learning needs are very different if you are advance-in-age (younger than 6-year-old when beginning first grade), typical-in-age or late-in-age (older than 6-year-old in first grade), and both advance and late did not present with a linear association between their age and their level in reading comprehension in second grade, contrary to typical-age children. Notably, all advanced-in-age children performed better than the oldest population of typical-in-age children. This finding implies that we should explore the learning skills and needs of advance-in-age children as they might need a specific training (i.e., more advanced and challenging compared to typical-in-age children) with a school programme adapted to their cognitive development. On the other hand, we would need more in-depth studies about the specificities that comprises the subgroup "late-in-age", as they might include students with identified handicap, students with an under-

stimulating environment around them, students with specific needs in health (detections of vision, audition, motricity, nutrition, sleep, ...).

As reading comprehension is the final goal of learning to read and as it is closely related to future academic success, we focused our study on reading comprehension in second grade as our main outcome (see Chapter 3). However, here we only presented and focused the analyses on typical-age children in first grade, when defining the learning needs of both advance-in-age children and late-in-age children would be necessary as they appeared to present very different responses among the three-levels data assessed (i.e., individual, class and school level). In addition, a deeper focus on children with lower SES score, either belonging to regular public schools or to PE and HPE public schools, would be necessary in future work, notably for matching Evalaide national data with additional individual and environmental data, as we only depicted here how related SES factors were to learning reading comprehension and would rather identify specific predictors to improve their language and math performance at the classroom and home level (e.g., Number of books at home, parental level of education, language stimulation daily activities, etc...) as found elsewhere (Chen et al., 2018; Demir-Lira et al., 2019). Finally, as we explored gender and their performances, and thanks to a situation similar to an “interrupted time series analysis” (i.e., our study is similar to Angrist’s in its designs that exploit a discontinuity (Angrist & Lavy, 1999) – here due to the French law of age for beginning first grade, within what should otherwise be a continuous curve as a function of age), we were able to identify that school, not age, is a trigger of the gender gap in math at school in France in favor of boys.

Amidst these academic challenges, the paramount imperative is early specific language identification. Recognizing specific predictors of oral comprehension difficulties, phonological deficits, and reading comprehension challenges in their nascent stages is the linchpin of effective intervention. By identifying and addressing these predictors, we can proactively steer children towards smoother learning pathways, fostering their academic growth and minimizing the enduring impact of these difficulties.

Back to Evalaide, not only having national evaluations allowed to develop and affine national education standards but also, this three-time assessments completed the classroom evaluation system by making it possible to determine, for each of the students in difficulty, what type of reinforcement intervention would be necessary for the child in a specific language domain, whether it should be redirected, or whether it should be stopped because the student would have developed the expected skills. In other words, these massive data represent a revolution in the knowledge to be acquired by children and Evalaide is a tool that makes it possible to know what every child needs to work on, to assess his progress and new learning needs at every period, in a precise way. According to the identified needs per child, subgroups of similar needs could be drawn per class with Evalaide data, helping teachers to organize and plan their reinforcement instructions. Reinforcement sessions and adapted learning solutions have been identified in the world to correct learning difficulties without excluding the child from his class, and lead to more equality regarding learning processes and for instance, reducing social inequalities in learning.

To promote the development of solid reading comprehension skills in first and second graders and to address any existing difficulties, earlier interventional educational program at school have been proven to be effective and to present with long-lasting effects on cognitive, social and schooling outcomes (Barnett, 2011). Indeed, some early interventions showed efficiency in tackling non-phonological language skills (i.e., vocabulary knowledge and syntactic skills) among 3-to-5-year-old children (Fricke et al., 2017a).

In 2012, the United Kingdom introduced a compulsory decoding assessment (i.e., called the phonics check) for all first graders, which emphasized the relationship between sounds and letters, and which consistently demonstrated positive outcomes in improving decoding skills and, subsequently, reading comprehension (Department of education, 2020; Duff et al., 2015; Ehri et al., 2001). This test involved reading aloud 20 frequent words and 20 invented words, alone with the teacher. The aim was to identify children who, at the end of the first school year, were unable to read a large proportion of these words correctly, in order to offer them individualized exercises. The

children identified in this way took the same test again a year later to ensure that they have made up for their shortcomings. If not, they were offered more targeted exercises. As shown by the international PIRLS study of reading comprehension in 10-year-olds (corresponding to CM1 in France), the scores of children educated in England improved between 2011 and 2016, while those of children educated in France deteriorated (PIRLS and TIMSS, 2016b). While it's not possible to attribute England's success solely to the introduction of phonics check, it was worth noting that English children's levels on this test had also risen year on year. Other similar literacy programs had been implemented in the United States (Gilbert et al., 2013; Kamps et al., 2008; Linan-Thompson et al., 2006; Lonigan et al., 2013; Shanahan & Lonigan, 2010) and Australia (Ken Rowe, 2005), and have demonstrated the effectiveness of the phonics check.

Another example of a study based on the Response to educational Intervention (RTI) protocol was the one in which 318 schools in Florida participated in which children from disadvantaged backgrounds were massively enrolled (72%), 14% of whom did not have English as their first language. This protocol provided high quality teaching, adapted to the needs of the children with a follow-up and a screening of those with difficulties, carried out using standardized tests administered four times a year to all children. Teachers were trained to use the results of these assessments to guide their pedagogical decisions. The main goal of this program was to enable most children to improve their reading skills and reach a level corresponding to their academic level. Over the four years of the program's implementation, the percentage of children reported as potentially struggling decreased, as has the percentage of those actually failing (Torgesen, 2009; Torgesen & Davis, 1996).

Another finding among NAPLAN Australian national results regarding gender gaps in reading, pointed the gender differences in using their metacognitive strategies and reading-related attitudes: the study suggested to teach children about these reading-related attitudes in order to reduce the gender gap reading achievement (Thomas et al., 2022). In addition, a study from Denmark revealed the advance level in reading for girls was associated with larger hours of training out of school compared to boys – this

later could encourage the parental implication as well in order to help boys raising their reading levels (E. Smith & Reimer, 2023).

By focusing on these proven strategies and addressing the foundational components of reading comprehension, first and second graders were empowered to develop strong reading comprehension skills, positioning them for academic success and lifelong learning.

In France, some evidence encouraged the specific training for early identified children with difficulties in subspecific domain of language, especially phonology skills in kindergarten (Bianco et al., 2010, 2012). In addition, efficient interventions in priority education kindergarten and primary schools of France were implemented and efficient on children's language development and found out that the most promising pedagogies are structured pedagogies, which frequently assess children's skills and properly target their needs. Also, frequent testing of pedagogy, measurement of progress, and adapting the classroom organization to small working groups are the ingredients for effective reading instruction for children with difficulties (Ecalte et al., 2019; Zorman † et al., 2015). A recent study identified a literacy interventional program for teachers on language training in kindergarten in France, where content was adapted to each child's needs, and which was three times more efficient than reducing class size on later reading skills (Bouguen, 2016).

We have the opportunity to curb the social inequalities in learning languages in France. These results amplify the need to target both lower SES children and identified-at-risk-of-difficulties children in language training, earlier than grade 1, based on evidence-based-education sources. Understanding the predictors of reading comprehension difficulties in kindergarten and first grade is crucial for early intervention efforts. Identifying children at risk and providing targeted support, such as phonological awareness training and vocabulary enrichment programs, may help mitigate these challenges and improve overall academic outcomes (Fricke et al., 2017b; Shanahan & Lonigan, 2010). Furthermore, ongoing research should explore the interplay of these predictors and the development of effective, evidence-based interventions that

address oral comprehension difficulties comprehensively. By addressing these predictors and implementing early interventions, educators and policymakers can work together to foster language development and academic success in young learners. France was one of the PISA-participating countries where children perceived some of the lowest levels of support and feedback from their teachers. Fewer than two in five children in France – compared to almost one in two children on average across OECD countries – reported that they think that their teacher usually helps them improve (OECD, 2018b). Although the teacher cannot change anything in the child’s past, he or she intervenes in the child’s present and can therefore influence the child’s future. The child’s destiny is not engraved in his or her genes, nor in the family or social environment. Teachers have the power to make all children progress, and to modify learning trajectories, at least to some extent (OECD, 2019c, 2019d).

Furthermore, regarding the study of gender gap in math favoring boys over girls at an early stage in school in France, several efficient remediation strategies have been showed as efficient. Which factors should be targeted? Our findings suggested that class-level variables such as class size, gender ratio, heterogeneity in math level, or gender of the student at the top of class only had a small influence. Single-sex schools or classes also were ineffective (Lee et al., 2014; Miller & Halpern, 2014). The most important intervention might be to convince all children that math is worth the effort for both genders. Other studies suggested that the following actions may be effective: supporting parents, informing them, and encouraging the development of a stimulating home learning environment (Gunderson et al., 2012; Melhuish et al., 2008; Miller & Halpern, 2014); encouraging both genders to play similar games for spatial thinking (Levine et al., 2016); encouraging teachers’ gender-fair ratings and practices (Cimpian et al., 2016), such as questioning girls and boys equally often during math and science courses (Miller & Halpern, 2014); exposing children to both male and female role models with whom they can identify (Stout et al., 2011); providing girls with means to cope with competitive stress (Buser et al., 2014, 2021) and math anxiety (D’Agostino et al., 2022); and informing them about the possible impact of stereotype threats in math (Johns et al., 2005; Miller & Halpern, 2014); emphasizing the role of effort, perseverance, and an incremental view of intelligence in efficient learning (Alan &

Ertac, 2019; Yeager et al., 2019); and implementing self-affirmation tasks to protect girls from stereotype threat (Régner et al., 2019). More generally, the present findings (in chapter 4) should enhance societal awareness of the absence of gender disparities in mathematical ability prior to the onset of school math learning and their rapid emergence when formal teaching of mathematics begins, non-correlated with age. Such awareness is a prerequisite to efforts, by parents as well as teachers, to encourage their children equally to build on their aptitude for learning school mathematics (Régner et al., 2019)

All these learning trajectories in both written and oral language and math abilities would not be possible for any children without the following fundamental health and well-being elements their body and brain rely on: Firstly, the need for proper cognitive functions, which can be altered or slowed down by birth factors such as prematurity (Vandormael et al., 2019), low birth weight (Byrne et al., 1993) and their resulted attentional problems (Ribeiro et al., 2011); exposure to toxics: alcohol during pregnancy (Mamluk et al., 2020), that present with attention-deficit and neuropsychological long-term effect (Lees et al., 2020); tobacco during pregnancy (Banderali et al., 2015; Zhou et al., 2014); other drugs during pregnancy (Ross et al., 2015; Thompson et al., 2009); the type of food during pregnancy and first years of life (i.e., lack of specific vitamins (Benton, 2012) such as the neurodevelopmental thiamine (vitamine B1) following the Remedias scandal in Israel in 2003 (Harel et al., 2017), deficiencies in iron (Radlowski & Johnson, 2013), in long chain poly unsaturated fat acids (Martinot et al., 2022)). Specific deficiencies such as auditory dysfunctions (Benasich et al., 2002) and visual dysfunctions (Ferretti et al., 2008) must be identified the earliest and taken care of with the proper care and help. The lack of sleep, both in quantity and in quality, is an underestimated factor associated with deficiencies in learning abilities (Knowland et al., 2019), as well as with a worst mental health (Blok et al., 2022; Cook et al., 2020). A lack of physical activity during the day contributes to less attention and affects children' cognitive functions (Watson et al., 2017). Finally, an abusing exposure to screen, depriving children from an interaction of quality with adults, results in poorer language abilities. In the meantime, every factor associated with a higher rate of interaction with adults that speak properly their language and

interact with a proper language with children will result in a higher level of language among concerned children (Martinot et al., 2021). Health outcomes associated with a better learning at school can be improved and, public policies should be encouraged to work on these fundamental steps to provide children with basis needed to learn properly.

Overall, throughout this work, several suggestions have emerged for an optimized use of these data and for supporting future work. For instance, (1) facilitating access and understanding of massive data results in education would be one way to go towards a personalized education (i.e., adapt the learnings to the children' needs) while maintaining children in a collective environment (i.e., the classroom and the school) in addition to help detect the children-at-risk of developing later difficulties in language or math, and to be a helpful tool to adapt the reinforcement learning intervention to the children' progresses; (2) evaluating the impact of learning and teaching intervention in the time would be possible and an objective measurement of children's progress to adapt the intervention to his needs ; (3) exploring other domains such as math specific subdomains and problem-solving that used both oral language comprehension for the instructions and the use of math skills to answer ; (4) exploring the interconnexions of math, problem-solving and language items, and subgroups of needs in when difficulties were noted.

After the cognitive sciences revolution thanks to the brain imaging and psychological experiments from the 90's, scientific results about learnings and educational practices need to be shared using useful, concrete and evidence-based information to help children perform in their academic journey. Imagining further perspectives following this work, the introduction of national assessments represented a step towards a more effective management of our educational system, at different levels: The individual level, the class level, the national level. For instance, these models facilitated the detection of children in need and at risk, streamlining the identification of their learning requirements while considering individual learning pace variations. This critical information supports the implementation of tailored preventive interventions in the classroom. Furthermore, this research serves as added incentive for educators to

employ personalized assessments from kindergarten to third grade, offering a dependable and practical foundation for monitoring each student's progress and selecting the most suitable teaching tools tailored to individual needs. Just as medicine and health have benefited from scientific evidence and experiments, science can enhance the effectiveness of education. Similar to the concept of personalized medicine tailoring prevention individualized programmes and treatment to individual patient needs, data and scientific approaches enable us to model, anticipate, and tailor educational requirements to the specific needs of children.

In the pursuit of constructing an optimal environment to support the development of every child and the expression of his full potential, examining the school environment's influence on children's learning experiences is one facet of this endeavor, identifying factors that may enhance or hinder the educational process. This multifaceted approach extends its reach to benefit not only children but also teachers and the classroom atmosphere, necessitating different organizational and determinative factors to meet the varied learning needs. Nonetheless, it is important to acknowledge that resistance to the adoption of these assessments persists among teachers and trade unions. A prevalent concern is that such assessments may pave the way for an individualized performance evaluation system for each teacher, potentially leading to differentiated remuneration and promotion criteria. Overcoming this resistance necessitates ongoing pedagogical efforts to underscore that these assessments primarily serve the interests of students' progress and aim to establish a prescriptive "diagnosis-action" framework for teachers.

At a broader level, decision-makers at the national and international levels can utilize scientific insights to inform policy decisions in education: using national data as a basis for coherent school programs; planning national strategy based on data and scientific proves (e.g., class size, teachers explicit teaching, parental implications, school rhythms, classes frequencies and age when starting school, stimulating girls and adapt teaching gesture especially in math ...); recognizing the profound impact of the child's environment, which encompasses family dynamics and the quality and quantity of interactions with adults, ultimately, fostering a wishful education for every child aims to

unlock and enhance their potential, fostering high-level, high-quality learning experiences. As Nobel laureate J. Heckman emphasized, “society bears a substantial cost when children fail to realize their full potential” (Heckman, 2008).

The imperative to establish strong interconnections between decision-makers (in the realm of public policies concerning education and the well-being of children) and research cannot be overstated. In a rapidly evolving world, where the needs of our youth and the dynamics of education are in constant flux, evidence-based interventions are the linchpin to effective policymaking. Research, backed by empirical data and rigorous scientific methodologies, provides invaluable insights into what works and what doesn't. It is through these insights that decision-makers can craft policies that are not only well-informed but also responsive to the real-world needs and challenges faced by children and educators. The interplay between research and policy is a reciprocal relationship, with research offering guidance to policymakers and policymakers, in turn, facilitating the application of evidence-based solutions. This synergy is essential for fostering innovation, enhancing the quality of education, and ultimately, ensuring the well-being and future success of our children. The union of decision-makers with research and evidence-based interventions forms the bedrock upon which we can build a brighter and more equitable educational landscape for the next generation.

Conclusion

In this study, the massive data of 2.9 million children followed at 3 times between grade 1 and grade 2 confirmed and identified precise predictors of reading comprehension in grade 2 and of the overall level in language domains in first and second grades. Also, we were able to identify patterns of difficulties to implement a specific training to prevent the development and long-term implementation of reading delays. In addition, our results measured the direct effect of schooling, of class size and of other classroom parameters for the development of reading abilities in second grade. Through a comprehensive analysis of potential early predictors, this thesis aimed at illuminating the way forward for educators, researchers, and policymakers, offering a roadmap

towards enhancing the educational journeys of young learners and shared hints of learning domains to implement effective evidence-based interventions.

Bibliography

1. Adlof, S. M., Catts, H. W., & Lee, J. (2010). Kindergarten Predictors of Second vs. Eighth Grade Reading Comprehension Impairments. *Journal of Learning Disabilities, 43*(4), 332–345. <https://doi.org/10.1177/0022219410369067>
2. Adlof, S. M., Scoggins, J., Brazendale, A., Babb, S., & Petscher, Y. (2017). Identifying Children at Risk for Language Impairment or Dyslexia With Group-Administered Measures. *Journal of Speech, Language, and Hearing Research: JSLHR, 60*(12), 3507–3522. https://doi.org/10.1044/2017_JSLHR-L-16-0473
3. Al Otaiba, S., Folsom, J. S., Schatschneider, C., Wanzek, J., Greulich, L., Meadows, J., Li, Z., & Connor, C. M. (2011). Predicting First-Grade Reading Performance from Kindergarten Response to Tier 1 Instruction. *Exceptional Children, 77*(4), 453–470. <https://doi.org/10.1177/001440291107700405>
4. Alan, S., & Ertac, S. (2019). Mitigating the Gender Gap in the Willingness to Compete: Evidence from a Randomized Field Experiment. *Journal of the European Economic Association, 17*(4), 1147–1185.
5. Amalric, M., & Dehaene, S. (2016). Origins of the brain networks for advanced mathematics in expert mathematicians. *Proceedings of the National Academy of Sciences, 113*(18), 4909–4917. <https://doi.org/10.1073/pnas.1603205113>
6. Angrist, J., & Lavy, V. (1999). Using Maimonides' Rule To Estimate The Effect Of Class Size On Scholastic Achievement. *The Quarterly Journal of Economics, 114*, 533–575. <https://doi.org/10.1162/003355399556061>
7. Banderali, G., Martelli, A., Landi, M., Moretti, F., Betti, F., Radaelli, G., Lassandro, C., & Verduci, E. (2015). Short and long term health effects of parental tobacco smoking during pregnancy and lactation: A descriptive review. *Journal of Translational Medicine, 13*(1), 327. <https://doi.org/10.1186/s12967->

8. Barnett, W. S. (2011). Effectiveness of Early Educational Intervention. *Science*, 333(6045), 975–978. <https://doi.org/10.1126/science.1204534>
9. Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, 107(5), 1860–1863. <https://doi.org/10.1073/pnas.0910967107>
10. Bellon, E., Fias, W., & De Smedt, B. (2019). More than number sense: The additional role of executive functions and metacognition in arithmetic. *Journal of Experimental Child Psychology*, 182, 38–60. <https://doi.org/10.1016/j.jecp.2019.01.012>
11. Benasich, A. A., Thomas, J. J., Choudhury, N., & Leppänen, P. H. T. (2002). The Importance of Rapid Auditory Processing Abilities to Early Language Development: Evidence from Converging Methodologies. *Developmental Psychobiology*, 40(3), 278–292. <https://doi.org/10.1002/dev.10032>
12. Benton, D. (2012). Vitamins and neural and cognitive developmental outcomes in children. *Proceedings of the Nutrition Society*, 71(1), 14–26. <https://doi.org/10.1017/S0029665111003247>
13. Bharadwaj, P., De Giorgi, G., Hansen, D., & Neilson, C. A. (2016). The Gender Gap in Mathematics: Evidence from Chile. *Economic Development and Cultural Change*, 65(1), 141–166. <https://doi.org/10.1086/687983>
14. Bianco, M., Bressoux, P., Doyen, A.-L., Lambert, E., Lima, L., Pellenq, C., & Zorman, M. (2010). Early Training in Oral Comprehension and Phonological Skills: Results of a Three-Year Longitudinal Study. *Scientific Studies of Reading*, 14(3), 211–246. <https://doi.org/10.1080/10888430903117518>

15. Bianco, M., Megherbi, H., Sénéchal, M., & Colé, P. (2014). Reading comprehension development: Presentation of the special issue. *L'Année psychologique*, 114(4), 613–621. <https://doi.org/10.3917/anpsy.144.0613>
16. Bianco, M., Pellenq, C., Lambert, E., Bressoux, P., Lima, L., & Doyen, A. (2012). Impact of early code-skill and oral-comprehension training on reading achievement in first grade. *Journal of Research in Reading*, 35(4), 427–455. <https://doi.org/10.1111/j.1467-9817.2010.01479.x>
17. Billard, C., Bricout, L., Ducot, B., Richard, G., Ziegler, J., & Fluss, J. (2010). [Evolution of competence in reading, spelling and comprehension levels in low socioeconomic environments and impact of cognitive and behavioral factors on outcome in two years]. *Revue D'épidemiologie Et De Sante Publique*, 58(2), 101–110. <https://doi.org/10.1016/j.respe.2009.11.002>
18. Billard, C., Fluss, J., Ducot, B., Bricout, L., Richard, G., Ecalle, J., Magnan, A., Warszawski, J., & Ziegler, J. (2009). [Deficits in reading acquisition in primary school: Cognitive, social and behavioral factors studied in a sample of 1062 children]. *Revue D'épidemiologie Et De Sante Publique*, 57(3), 191–203. <https://doi.org/10.1016/j.respe.2009.02.205>
19. Bishop, D. V. M. (2003). *Test for Reception of Grammar: TROG-2 Version 2*. Pearson Assessment.
20. Blok, E., Koopman-Verhoeff, M. E., Dickstein, D. P., Saletin, J., Luik, A. I., Rijlaarsdam, J., Hillegers, M., Kocavska, D., White, T., & Tiemeier, H. (2022). Sleep and mental health in childhood: A multi-method study in the general pediatric population. *Child and Adolescent Psychiatry and Mental Health*, 16(1), 11. <https://doi.org/10.1186/s13034-022-00447-0>
21. Booth, J. L., & Siegler, R. S. (2008). Numerical Magnitude Representations Influence Arithmetic Learning. *Child Development*, 79(4), 1016–1031.

<https://doi.org/10.1111/j.1467-8624.2008.01173.x>

22. Borgonovi, F., Choi, A., & Paccagnella, M. (2021). The evolution of gender gaps in numeracy and literacy between childhood and young adulthood. *Economics of Education Review*, 82, 102119. <https://doi.org/10.1016/j.econedurev.2021.102119>
23. Bouguen, A. (2016). Adjusting content to individual student needs: Further evidence from an in-service teacher training program. *Economics of Education Review*, 50, 90–112. <https://doi.org/10.1016/j.econedurev.2015.12.004>
24. Brankaer, C., Ghesquière, P., & De Smedt, B. (2017). Symbolic magnitude processing in elementary school children: A group administered paper-and-pencil measure (SYMP Test). *Behavior Research Methods*, 49(4), 1361–1373. <https://doi.org/10.3758/s13428-016-0792-3>
25. Breda, T., Jouini, E., Napp, C., & Thebault, G. (2020). Gender stereotypes can explain the gender-equality paradox. *Proceedings of the National Academy of Sciences*, 117(49), 31063–31069. <https://doi.org/10.1073/pnas.2008704117>
26. Breda, T., & Napp, C. (2019). Girls' comparative advantage in reading can largely explain the gender gap in math-related fields. *Proceedings of the National Academy of Sciences*, 116(31), 15435–15440. <https://doi.org/10.1073/pnas.1905779116>
27. Bressoux, P. (2010). *Modélisation statistique appliquée aux sciences sociales* (2nde, De Boeck Supérieur ed.). <https://www.deboecksuperieur.com/ouvrage/9782804163648-modelisation-statistique-appliquee-aux-sciences-sociales>
28. Bressoux, P. (2012). 13. L'influence des pratiques enseignantes sur les acquisitions scolaires des élèves. *Regards croisés sur l'économie*, 12(2), 208–

217. <https://doi.org/10.3917/rce.012.0208>

29. Bressoux, P., Lima, L., & Monseur, C. (2019). Reducing the number of pupils in French first-grade classes: Is there evidence of contemporaneous and carryover effects? *International Journal of Educational Research*, *96*, 136–145. <https://doi.org/10.1016/j.ijer.2018.10.006>
30. Broccolichi, S., & Sinthon, R. (2011). Links between Unequal Learning Outcomes and Course Selection Inequalities: Ignored Relations and Overdue Remedies. *Revue Française de Pédagogie*, *175*(2), 15–38. Cairn.info. <https://doi.org/10.4000/rfp.3017>
31. Buser. (2014). *Gender, Competitiveness, and Career Choices*. The Quarterly Journal of Economics, Oxford Academic. <https://academic.oup.com/qje/article-abstract/129/3/1409/1817134?redirectedFrom=fulltext>
32. Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, Competitiveness, and Career Choices. *The Quarterly Journal of Economics*, *129*(3), 1409–1447.
33. Buser, T., Ranehill, E., & van Veldhuizen, R. (2021). Gender differences in willingness to compete: The role of public observability. *Journal of Economic Psychology*, *83*, 102366. <https://doi.org/10.1016/j.joep.2021.102366>
34. Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*, 1–67. <https://doi.org/10.18637/jss.v045.i03>
35. Byrne, J., Ellsworth, C., Bowering, E., & Vincer, M. (1993). Language development in low birth weight infants: The first two years of life. *Journal of Developmental and Behavioral Pediatrics: JDBP*, *14*(1), 21–27.
36. Cain, K., & Oakhill, J. (2014). Reading comprehension and vocabulary: Is

vocabulary more important for some aspects of comprehension? *L'Année psychologique*, 114(4), 647–662. <https://doi.org/10.3917/anpsy.144.0647>

37. Campbell, J. A., McIntyre, J., & Kucirkova, N. (2021). Gender Equality, Human Development, and PISA Results over Time. *Social Sciences*, 10(12), Article 12. <https://doi.org/10.3390/socsci10120480>
38. Carlana, M. (2019). Implicit Stereotypes: Evidence from Teachers' Gender Bias*. *The Quarterly Journal of Economics*, 134, 1163–1224. <https://doi.org/10.1093/qje/qjz008>
39. Carroll, J. M., Maughan, B., Goodman, R., & Meltzer, H. (2005). Literacy difficulties and psychiatric disorders: Evidence for comorbidity. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 46(5), 524–532. <https://doi.org/10.1111/j.1469-7610.2004.00366.x>
40. Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert: *Psychological Science in the Public Interest*. <https://doi.org/10.1177/1529100618772271>
41. Catts, H. W., Hogan, T. P., & Adlof, S. M. (2005). Developmental changes in reading and reading disabilities. In *The connections between language and reading disabilities* (pp. 25–40). Lawrence Erlbaum Associates Publishers.
42. Catts, H. W., Nielsen, D. C., Bridges, M. S., & Liu, Y.-S. (2016). Early Identification of Reading Comprehension Difficulties. *Journal of Learning Disabilities*, 49(5), 451–465. <https://doi.org/10.1177/0022219414556121>
43. Cavaglia, C., Machin, S., McNally, S., & Ruiz-Valenzuela, J. (2020). Gender, achievement, and subject choice in English education. *Oxford Review of Economic Policy*, 36(4), 816–835. <https://doi.org/10.1093/oxrep/graa050>

44. Chen, Q., Kong, Y., Gao, W., & Mo, L. (2018). Effects of Socioeconomic Status, Parent–Child Relationship, and Learning Motivation on Reading Ability. *Frontiers in Psychology*, 9. <https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01297>
45. Christle, C. A., & Yell, M. L. (2008). Preventing Youth Incarceration Through Reading Remediation: Issues and Solutions. *Reading & Writing Quarterly*, 24(2), 148–176. <https://doi.org/10.1080/10573560701808437>
46. Cimpian, J. R., Lubienski, S. T., Timmer, J. D., Makowski, M. B., & Miller, E. K. (2016). Have Gender Gaps in Math Closed? Achievement, Teacher Perceptions, and Learning Behaviors across Two ECLS-K Cohorts. *AERA Open*, 2(4). <https://eric.ed.gov/?id=EJ1194383>
47. Clayton, F. J., West, G., Sears, C., Hulme, C., & Lervåg, A. (2020). A Longitudinal Study of Early Reading Development: Letter-Sound Knowledge, Phoneme Awareness and RAN, but Not Letter-Sound Integration, Predict Variations in Reading Development. *Scientific Studies of Reading*, 24(2), 91–107. <https://doi.org/10.1080/10888438.2019.1622546>
48. Colmant, M., Dos Santos, S., France, M. de l'Education nationale (MEN), SDPES, & France, M. de l'Education nationale. (2008). *Evolution des performances en lecture des élèves de CM 1. Résultats de l'étude internationale PIRLS*. Ministère de l'Education nationale (MEN). Paris. <https://archives-statistiques-depp.education.gouv.fr/Default/doc/SYRACUSE/9787/evolution-des-performances-en-lecture-des-eleves-de-cm-1-resultats-de-l-etude-internationale-pirls>
49. Contini, D., Tommaso, M. L. D., & Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review*, 58(C), 32–42.

50. Cook, F., Conway, L. J., Giallo, R., Gartland, D., Sciberras, E., & Brown, S. (2020). Infant sleep and child mental health: A longitudinal investigation. *Archives of Disease in Childhood*, *105*(7), 655–660. <https://doi.org/10.1136/archdischild-2019-318014>
51. Cunningham, A., & Carroll, J. (2011). Age and schooling effects on early literacy and phoneme awareness. *Journal of Experimental Child Psychology*, *109*(2), 248–255. <https://doi.org/10.1016/j.jecp.2010.12.005>
52. Currie, N. K., & Cain, K. (2015). Children’s inference generation: The role of vocabulary and working memory. *Journal of Experimental Child Psychology*, *137*, 57–75. <https://doi.org/10.1016/j.jecp.2015.03.005>
53. D’Agostino, A., Schirripa Spagnolo, F., & Salvati, N. (2022). Studying the relationship between anxiety and school achievement: Evidence from PISA data. *Statistical Methods and Applications*, *31*(1), 1–20. <https://doi.org/10.1007/s10260-021-00563-9>
54. Daussin, Rocher, & Keskpaik. (2011). *L’évolution du nombre d’élèves en difficulté face à l’écrit depuis une dizaine d’années – France, portrait social I Insee*. <https://www.insee.fr/fr/statistiques/1373895?sommaire=1373905>
55. De la Haye, F., Gombert, J.-É., Riviere, J.-P., Chabanon, L., France, M. de l’Education nationale (MEN), France, U. D. R. I., France, S. W., & France, E. de B. (2018). *Journée Défense et Citoyenneté 2017: Plus d’un jeune Français sur dix en difficulté de lecture / Léa Chabanon, Fanny De La Haye, Jean-Emile Gombert, Jean-Philippe Rivière*. Ministère de l’éducation nationale. Paris. <https://archives-statistiques-depp.education.gouv.fr/Default/doc/SYRACUSE/43911/journee-defense-et-citoyennete-2017-plus-d-un-jeune-francais-sur-dix-en-difficulte-de-lecture-lea-ch>

56. Dehaene, S. (1997). *The Number Sense: How the Mind Creates Mathematics* (Vol. 47). Taylor & Francis.
57. Dehaene S. (2021). *How we learn: Why Brains Learn Better Than Any Machine... For Now* (Penguin random house).
<https://www.penguinrandomhouse.com/books/579922/how-we-learn-by-stanislas-dehaene/>
58. Dehaene, S., & Cohen, L. (1997). Cerebral pathways for calculation: Double dissociation between rote verbal and quantitative knowledge of arithmetic. *Cortex*, 33, 219–250.
59. Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6), 254–262.
<https://doi.org/10.1016/j.tics.2011.04.003>
60. Dehaene, S., Izard, V., Pica, P., & Spelke, E. (2006). Core knowledge of geometry in an Amazonian indigene group. *Science*, 311, 381–384.
61. Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science (New York, N.Y.)*, 320(5880), 1217–1220.
<https://doi.org/10.1126/science.1156540>
62. Dehaene, S., Nakamura, K., Jobert, A., Kuroki, C., Ogawa, S., & Cohen, L. (2010). Why do children make mirror errors in reading? Neural correlates of mirror invariance in the visual word form area. *NeuroImage*, 49(2), 1837–1848.
<https://doi.org/10.1016/j.neuroimage.2009.09.024>
63. Dehaene-Lambertz, G. (2017). The human infant brain: A neural architecture able to learn language. *Psychonomic Bulletin & Review*, 24(1), 48–55.
<https://doi.org/10.3758/s13423-016-1156-9>

64. Dehaene-Lambertz, G., & Spelke, E. S. (2015). The Infancy of the Human Brain. *Neuron*, *88*(1), 93–109. <https://doi.org/10.1016/j.neuron.2015.09.026>
65. del Río, M. F., Strasser, K., Cvencek, D., Susperreguy, M. I., & Meltzoff, A. N. (2019). Chilean kindergarten children's beliefs about mathematics: Family matters. *Developmental Psychology*, *55*(4), 687–702. <https://doi.org/10.1037/dev0000658>
66. Demir-Lira, Ö. E., Applebaum, L. R., Goldin-Meadow, S., & Levine, S. C. (2019). Parents' early book reading to children: Relation to children's later language and literacy outcomes controlling for other parent language input. *Developmental Science*, *22*(3), e12764. <https://doi.org/10.1111/desc.12764>
67. Department of education. (2020). *Phonic screening checks*. GOV.UK. <https://www.gov.uk/government/collections/statistics-key-stage-1>
68. Desrochers, A. (2018). L'évaluation des difficultés en lecture du français. *Langue française*, *199*(3), 83–97. <https://doi.org/10.3917/lf.199.0083>
69. Dillon, M. R., Kannan, H., Dean, J. T., Spelke, E. S., & Duflo, E. (2017). Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics. *Science*, *357*(6346), 47–55. <https://doi.org/10.1126/science.aal4724>
70. Dong, Y., Tang, Y., Chow, B. W.-Y., Wang, W., & Dong, W.-Y. (2020). Contribution of Vocabulary Knowledge to Reading Comprehension Among Chinese Students: A Meta-Analysis. *Frontiers in Psychology*, *11*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2020.525369>
71. Dotan, D., & Dehaene, S. (2016). On the origins of logarithmic number-to-position mapping. *Psychological Review*, *123*(6), 637–666. <https://doi.org/10.1037/rev0000038>

72. Duff, F. J., Mengoni, S. E., Bailey, A. M., & Snowling, M. J. (2015). Validity and sensitivity of the phonics screening check: Implications for practice. *Journal of Research in Reading, 38*(2), 109–123. <https://doi.org/10.1111/1467-9817.12029>
73. Ecalle, J., Gomes, C., Auphan, P., Cros, L., & Magnan, A. (2019). Effects of policy and educational interventions intended to reduce difficulties in literacy skills in grade 1. *Studies in Educational Evaluation, 61*, 12–20. <https://doi.org/10.1016/j.stueduc.2019.02.001>
74. Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research, 71*(3), 393–447. <https://doi.org/10.3102/00346543071003393>
75. Enge, A., Kapoor, S., Kieslinger, A.-S., & Skeide, M. A. (2023). A meta-analysis of mental rotation in the first years of life. *Developmental Science, 26*(6), e13381. <https://doi.org/10.1111/desc.13381>
76. Etchell, A., Adhikari, A., Weinberg, L. S., Choo, A. L., Garnett, E. O., Chow, H. M., & Chang, S.-E. (2018). A systematic literature review of sex differences in childhood language and brain development. *Neuropsychologia, 114*, 19–31. <https://doi.org/10.1016/j.neuropsychologia.2018.04.011>
77. Ferretti, G., Mazzotti, S., & Brizzolara, D. (2008). Visual scanning and reading ability in normal and dyslexic children. *Behavioural Neurology, 19*(1–2), 87–92. <https://doi.org/10.1155/2008/564561>
78. Fischer, J.-P., & Thierry, X. (2021). Boy's math performance, compared to girls', jumps at age 6 (in the ELFE's data at least). *British Journal of Developmental Psychology, 124*(23). <https://doi.org/10.1111/bjdp.12423>

79. Fluss, J., Ziegler, J. C., Warszawski, J., Ducot, B., Richard, G., & Billard, C. (2009). Poor Reading in French Elementary School: The Interplay of Cognitive, Behavioral, and Socioeconomic Factors. *Journal of Developmental & Behavioral Pediatrics*, *30*(3), 206. <https://doi.org/10.1097/DBP.0b013e3181a7ed6c>
80. Fricke, S., Burgoyne, K., Bowyer-Crane, C., Kyriacou, M., Zosimidou, A., Maxwell, L., Lervåg, A., Snowling, M. J., & Hulme, C. (2017a). The efficacy of early language intervention in mainstream school settings: A randomized controlled trial. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *58*(10), 1141–1151. <https://doi.org/10.1111/jcpp.12737>
81. Fricke, S., Burgoyne, K., Bowyer-Crane, C., Kyriacou, M., Zosimidou, A., Maxwell, L., Lervåg, A., Snowling, M. J., & Hulme, C. (2017b). The efficacy of early language intervention in mainstream school settings: A randomized controlled trial. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *58*(10), 1141–1151. <https://doi.org/10.1111/jcpp.12737>
82. Fryer Jr., R. G., & Levitt, S. D. (2010). An Empirical Analysis of the Gender Gap in Mathematics. *American Economic Journal: Applied Economics*, *2*(2), 210–240. <https://doi.org/10.1257/app.2.2.210>
83. Geary, D. C. (2011). Cognitive Predictors of Achievement Growth in Mathematics: A Five Year Longitudinal Study. *Developmental Psychology*, *47*(6), 1539–1552. <https://doi.org/10.1037/a0025510>
84. Gennari, G., Dehaene, S., Valera, C., & Dehaene-Lambertz, G. (2023). Spontaneous supra-modal encoding of number in the infant brain. *Current Biology*, *33*(10), 1906-1915.e6. <https://doi.org/10.1016/j.cub.2023.03.062>
85. Gentaz, E., Sprenger-Charolles, L., & Theurel, A. (2015a). Differences in the predictors of reading comprehension in first graders from low socio-economic

status families with either good or poor decoding skills. *PloS One*, *10*(3), e0119581. <https://doi.org/10.1371/journal.pone.0119581>

86. Gentaz, E., Sprenger-Charolles, L., & Theurel, A. (2015b). Differences in the Predictors of Reading Comprehension in First Graders from Low Socio-Economic Status Families with Either Good or Poor Decoding Skills. *PLOS ONE*, *10*(3), e0119581. <https://doi.org/10.1371/journal.pone.0119581>
87. Gentaz, E., Sprenger-Charolles, L., Theurel, A., & Colé, P. (2013). Reading Comprehension in a Large Cohort of French First Graders from Low Socio-Economic Status Families: A 7-Month Longitudinal Study. *PLoS ONE*, *8*(11). <https://doi.org/10.1371/journal.pone.0078608>
88. Gershenson, S., Hart, C. M. D., Hyman, J., Lindsay, C. A., & Papageorge, N. W. (2022). The Long-Run Impacts of Same-Race Teachers. *American Economic Journal: Economic Policy*, *14*(4), 300–342. <https://doi.org/10.1257/pol.20190573>
89. Ghasemi, E., & Burley, H. (2019). Gender, affect, and math: A cross-national meta-analysis of Trends in International Mathematics and Science Study 2015 outcomes. *Large-Scale Assessments in Education*, *7*(1), 10. <https://doi.org/10.1186/s40536-019-0078-1>
90. Gilbert, J. K., Compton, D. L., Fuchs, D., Fuchs, L. S., Bouton, B., Barquero, L. A., & Cho, E. (2013). Efficacy of a First-Grade Responsiveness-to-Intervention Prevention Model for Struggling Readers. *Reading Research Quarterly*, *48*(2), 135–154. <https://doi.org/10.1002/rrq.45>
91. Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2007). Symbolic arithmetic knowledge without instruction. *Nature*, *447*(7144), 589–591. <https://doi.org/10.1038/nature05850>

92. Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2010). Non-symbolic arithmetic abilities and mathematics achievement in the first year of formal schooling. *Cognition*, *115*(3), 394–406. <https://doi.org/10.1016/j.cognition.2010.02.002>
93. Gimbert, F., Camos, V., Gentaz, E., & Mazens, K. (2019). What predicts mathematics achievement? Developmental change in 5- and 7-year-old children. *Journal of Experimental Child Psychology*, *178*, 104–120. <https://doi.org/10.1016/j.jecp.2018.09.013>
94. Girelli, L., Lucangeli, D., & Butterworth, B. (2000). The development of automaticity in accessing number magnitude. *Journal of Experimental Child Psychology*, *76*(2), 104–122. <https://doi.org/10.1006/jecp.2000.2564>
95. Gough, P. B., & Tunmer, W. E. (1986). Decoding, Reading, and Reading Disability. *Remedial and Special Education*, *7*(1), 6–10. <https://doi.org/10.1177/074193258600700104>
96. Grainger, J., Dufau, S., & Ziegler, J. C. (2016). A Vision of Reading. *Trends in Cognitive Sciences*, *20*(3), 171–179. <https://doi.org/10.1016/j.tics.2015.12.008>
97. Gunderson, E. A., Ramirez, G., Levine, S. C., & Beilock, S. L. (2012). The role of parents and teachers in the development of gender-related math attitudes. *Sex Roles: A Journal of Research*, *66*(3–4), 153–166. <https://doi.org/10.1007/s11199-011-9996-2>
98. Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the “number sense”: The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457–1465. <https://doi.org/10.1037/a0012682>
99. Hamdan, N., & Gunderson, E. A. (2017). The number line is a critical spatial-numerical representation: Evidence from a fraction intervention. *Developmental*

Psychology, 53(3), 587–596. <https://doi.org/10.1037/dev0000252>

100. Hammerstein, S., König, C., Dreisörner, T., & Frey, A. (2021). Effects of COVID-19-Related School Closures on Student Achievement-A Systematic Review. *Frontiers in Psychology*, 12. <https://www.frontiersin.org/articles/10.3389/fpsyg.2021.746289>
101. Harel, Y., Zuk, L., Guindy, M., Nakar, O., Lotan, D., & Fattal-Valevski, A. (2017). The effect of subclinical infantile thiamine deficiency on motor function in preschool children. *Maternal & Child Nutrition*, 13(4), e12397. <https://doi.org/10.1111/mcn.12397>
102. Heckman, J. J. (2008). Schools, Skills, and Synapses. *Economic Inquiry*, 46(3), 289.
103. Hjetland, H., Brinchmann, E., Scherer, R., & Melby-Lervåg, M. (2017). Preschool predictors of later reading comprehension ability: A systematic review. *Campbell Systematic Reviews*, 14. <https://doi.org/10.4073/csr.2017.14>
104. Hjetland, H. N., Lervåg, A., Lyster, S.-A. H., Hagtvet, B. E., Hulme, C., & Melby-Lervåg, M. (2019). Pathways to reading comprehension: A longitudinal study from 4 to 9 years of age. *Journal of Educational Psychology*, 111(5), 751. <https://doi.org/10.1037/edu0000321>
105. Hogan, T. P., Adlof, S. M., & Alonzo, C. N. (2014). On the importance of listening comprehension. *International Journal of Speech-Language Pathology*, 16(3), 199–207. <https://doi.org/10.3109/17549507.2014.904441>
106. Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, 2(2), 127–160. <https://doi.org/10.1007/BF00401799>

107. Huguet, P., Brunot, S., & Monteil, J. M. (2001). Geometry versus drawing: Changing the meaning of the task as a means to change performance. *Social Psychology of Education: An International Journal*, 4(3–4), 219–234. <https://doi.org/10.1023/A:1011374700020>
108. Huguet, P., & Régner, I. (2009). Counter-stereotypic beliefs in math do not protect school girls from stereotype threat. *Journal of Experimental Social Psychology*, 45, 1024–1027. <https://doi.org/10.1016/j.jesp.2009.04.029>
109. Hutchison, J. E., Lyons, I. M., & Ansari, D. (2019). More Similar Than Different: Gender Differences in Children’s Basic Numerical Skills Are the Exception Not the Rule. *Child Development*, 90(1), e66–e79. <https://doi.org/10.1111/cdev.13044>
110. Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender Similarities Characterize Math Performance. *Science*, 321(5888), 494–495. <https://doi.org/10.1126/science.1160364>
111. Izard, V., Dehaene-Lambertz, G., & Dehaene, S. (2008). Distinct cerebral pathways for object identity and number in human infants. *PLoS Biology*, 6(2), e11. <https://doi.org/10.1371/journal.pbio.0060011>
112. Johns, M., Schmader, T., & Martens, A. (2005). Knowing Is Half the Battle: Teaching Stereotype Threat as a Means of Improving Women’s Math Performance. *Psychological Science*, 16(3), 175–179. <https://doi.org/10.1111/j.0956-7976.2005.00799.x>
113. Josse, J., & Husson, F. (2016). missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software*, 70, 1–31. <https://doi.org/10.18637/jss.v070.i01>
114. Kamps, D., Abbott, M., Greenwood, C., Wills, H., Veerkamp, M., &

- Kaufman, J. (2008). Effects of small-group reading instruction and curriculum differences for students most at risk in kindergartenL: Two-year results for secondary- and tertiary-level interventions. *Journal of Learning Disabilities*, 41(2), 101–114. <https://doi.org/10.1177/0022219407313412>
115. Ken Rowe. (2005). “*Teaching Reading: Report and Recommendations*” by Ken Rowe and *National Inquiry into the Teaching of Literacy (Australia)*. ACER. https://research.acer.edu.au/tll_misc/5/
116. Kendeou, P., van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology*, 101(4), 765–778. <https://doi.org/10.1037/a0015956>
117. Kersey, A. J., Braham, E. J., Csumitta, K. D., Libertus, M. E., & Cantlon, J. F. (2018). No intrinsic gender differences in children’s earliest numerical abilities. *NPJ Science of Learning*, 3, 12. <https://doi.org/10.1038/s41539-018-0028-7>
118. Kim, Y.-S. G. (2016). Direct and mediated effects of language and cognitive skills on comprehension of oral narrative texts (listening comprehension) for children. *Journal of Experimental Child Psychology*, 141, 101–120. <https://doi.org/10.1016/j.jecp.2015.08.003>
119. Kirby, J. R., Desrochers, A., Roth, L., & Lai, S. S. V. (2008). Longitudinal predictors of word reading development. *Canadian Psychology / Psychologie Canadienne*, 49(2), 103–110. <https://doi.org/10.1037/0708-5591.49.2.103>
120. Knowland, V. C. P., Fletcher, F., Henderson, L.-M., Walker, S., Norbury, C. F., & Gaskell, M. G. (2019). Sleep Promotes Phonological Learning in Children Across Language and Autism Spectra. *Journal of Speech, Language, and Hearing Research*, 62(12), 4235–4255.

https://doi.org/10.1044/2019_JSLHR-S-19-0098

121. Kolinsky, R., Morais, J., Cohen, L., & Dehaene, S. (2018). Les bases neurales de l'apprentissage de la lecture. *Langue française*, 199(3), 17–33. <https://doi.org/10.3917/lf.199.0017>
122. Lauer, J. E., Yhang, E., & Lourenco, S. F. (2019). The development of gender differences in spatial reasoning: A meta-analytic review. *Psychological Bulletin*, 145(6), 537–565. <https://doi.org/10.1037/bul0000191>
123. Lauterbach, A. A., Park, Y., & Lombardino, L. J. (2017). The roles of cognitive and language abilities in predicting decoding and reading comprehension: Comparisons of dyslexia and specific language impairment. *Annals of Dyslexia*, 67(3), 201–218.
124. Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1–18.
125. Lecocq. (1996). *L'É.co.s.se une épreuve de compréhension syntaxico-sémantique (manuel et épreuve)*, Deux volumes. <https://www.septentrion.com/fr/livre/?GCOI=27574100290010>
126. Lee, S., Niederle, M., & Kang, N. (2014). Do single-sex schools make girls more competitive? *Economics Letters*, 124(3), 474–477. <https://doi.org/10.1016/j.econlet.2014.07.001>
127. Lees, B., Mewton, L., Jacobus, J., Valadez, E. A., Stapinski, L. A., Teesson, M., Tapert, S. F., & Squeglia, L. M. (2020). Association of Prenatal Alcohol Exposure With Psychological, Behavioral, and Neurodevelopmental Outcomes in Children From the Adolescent Brain Cognitive Development Study. *American Journal of Psychiatry*, 177(11), 1060–1072. <https://doi.org/10.1176/appi.ajp.2020.20010086>

128. Leppänen, U., Aunola, K., Niemi, P., & Nurmi, J.-E. (2008). Letter knowledge predicts Grade 4 reading fluency and reading comprehension. *Learning and Instruction, 18*(6), 548–564. <https://doi.org/10.1016/j.learninstruc.2007.11.004>
129. Lervåg, A., Hulme, C., & Melby-Lervåg, M. (2018a). Unpicking the Developmental Relationship Between Oral Language Skills and Reading Comprehension: It's Simple, But Complex. *Child Development, 89*(5), 1821–1838. <https://doi.org/10.1111/cdev.12861>
130. Lervåg, A., Hulme, C., & Melby-Lervåg, M. (2018b). Unpicking the Developmental Relationship Between Oral Language Skills and Reading Comprehension: It's Simple, But Complex. *Child Development, 89*(5), 1821–1838. <https://doi.org/10.1111/cdev.12861>
131. Levine, S. C., Foley, A., Lourenco, S., Ehrlich, S., & Ratliff, K. (2016). Sex differences in spatial cognition: Advancing the conversation. *Wiley Interdisciplinary Reviews. Cognitive Science, 7*(2), 127–155. <https://doi.org/10.1002/wcs.1380>
132. Linan-Thompson, S., Vaughn, S., Prater, K., & Cirino, P. T. (2006). The response to intervention of English language learners at risk for reading problems. *Journal of Learning Disabilities, 39*(5), 390–398. <https://doi.org/10.1177/00222194060390050201>
133. Lindeman J. (1998). *ALLU–Ala-Asteen Lukutesti [ALLU–Reading Test for Primary School]*.
134. Lonigan, C. J., Burgess, S. R., & Schatschneider, C. (2018). Examining the simple view of reading with elementary school children: Still simple after all these years. *Remedial and Special Education, 39*(5), 260–273. <https://doi.org/10.1177/0741932518764833>

135. Lonigan, C. J., Purpura, D. J., Wilson, S. B., Walker, P. M., & Clancy-Menchetti, J. (2013). Evaluating the components of an emergent literacy intervention for preschool children at risk for reading difficulties. *Journal of Experimental Child Psychology*, *114*(1), 111–130. <https://doi.org/10.1016/j.jecp.2012.08.010>
136. Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1-6. *Developmental Science*, *17*(5), 714–726. <https://doi.org/10.1111/desc.12152>
137. Machin, S., & McNally, S. (2005). Gender and Student Achievement in English Schools. *Oxford Review of Economic Policy*, *21*(3), 357–372.
138. Mamluk, L., Jones, T., Ijaz, S., Edwards, H. B., Savović, J., Leach, V., Moore, T. H. M., von Hinke, S., Lewis, S. J., Donovan, J. L., Lawlor, D. A., Davey Smith, G., Fraser, A., & Zuccolo, L. (2020). Evidence of detrimental effects of prenatal alcohol exposure on offspring birthweight and neurodevelopment from a systematic review of quasi-experimental studies. *International Journal of Epidemiology*, *49*(6), 1972–1995. <https://doi.org/10.1093/ije/dyz272>
139. Martinot, P., Adjibade, M., Taine, M., Davaisse-Paturet, C., Lioret, S., Charles, M.-A., de Lauzon-Guillain, B., & Bernard, J. Y. (2022). LC-PUFA enrichment in infant formula and neurodevelopment up to age 3.5 years in the French nationwide ELFE birth cohort. *European Journal of Nutrition*, *61*(6), 2979–2991. <https://doi.org/10.1007/s00394-022-02863-6>
140. Martinot, P., Bernard, J. Y., Peyre, H., De Agostini, M., Forhan, A., Charles, M.-A., Plancoulaine, S., & Heude, B. (2021). Exposure to screens and children’s language development in the EDEN mother-child cohort. *Scientific Reports*, *11*(1), 11863. <https://doi.org/10.1038/s41598-021-90867-3>
141. Massonnié, J., Bianco, M., Lima, L., & Bressoux, P. (2019). Longitudinal

predictors of reading comprehension in French at first grade: Unpacking the oral comprehension component of the simple view. *Learning and Instruction*, 60, 166–179. <https://doi.org/10.1016/j.learninstruc.2018.01.005>

142. McArthur, G., & Castles, A. (2017). Helping children with reading difficulties: Some things we have learned so far. *Npj Science of Learning*, 2(1), Article 1. <https://doi.org/10.1038/s41539-017-0008-3>
143. McArthur, G., Castles, A., Kohnen, S., & Banales, E. (2016). Low self-concept in poor readers: Prevalence, heterogeneity, and risk. *PeerJ*, 4, e2669. <https://doi.org/10.7717/peerj.2669>
144. McBride-Chang, C., & Kail, R. V. (2002). Cross-Cultural Similarities in the Predictors of Reading Acquisition. *Child Development*, 73(5), 1392–1407. <https://doi.org/10.1111/1467-8624.00479>
145. Melby-Lervåg, M., Lyster, S.-A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin*, 138(2), 322–352. <https://doi.org/10.1037/a0026744>
146. Melhuish, E. C., Sylva, K., Sammons, P., Siraj-Blatchford, I., Taggart, B., Phan, M. B., & Malin, A. (2008). The early years. Preschool influences on mathematics achievement. *Science (New York, N.Y.)*, 321(5893), 1161–1162. <https://doi.org/10.1126/science.1158808>
147. Miller, D. I., & Halpern, D. F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences*, 18(1), 37–45. <https://doi.org/10.1016/j.tics.2013.10.011>
148. Miller, D. I., Nolla, K. M., Eagly, A. H., & Uttal, D. H. (2018). The development of children's gender-science stereotypes: A meta-analysis of 5 decades of U.S. Draw-a-Scientist studies. *Child Development*, 89(6), 1943–1955. <https://doi.org/10.1111/cdev.13039>

149. Miyake, A., Kost-Smith, L. E., Finkelstein, N. D., Pollock, S. J., Cohen, G. L., & Ito, T. A. (2010). Reducing the gender achievement gap in college science: A classroom study of values affirmation. *Science (New York, N. Y.)*, *330*(6008), 1234–1237. <https://doi.org/10.1126/science.1195996>
150. Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., Streiftau, S., Lyytinen, H., Leppänen, P. H. T., Lohvansuu, K., Tóth, D., Honbolygó, F., Csépe, V., Bogliotti, C., Iannuzzi, S., Démonet, J. F., Longeras, E., Valdois, S., George, F., ... Landerl, K. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learning and Instruction*, *29*, 65–77. <https://doi.org/10.1016/j.learninstruc.2013.09.003>
151. Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). TIMSS 2011 International Results in Mathematics. In *International Association for the Evaluation of Educational Achievement*. International Association for the Evaluation of Educational Achievement. <https://eric.ed.gov/?id=ED544554>
152. Mullis, I. V. S., Martin, M. O., Ruddock, G., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 assessment frameworks*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.
153. Nachshon, O., & Horowitz-Kraus, T. (2019). Cognitive and emotional challenges in children with reading difficulties. *Acta Paediatrica (Oslo, Norway: 1992)*, *108*(6), 1110–1114. <https://doi.org/10.1111/apa.14672>
154. Nation, K., Cocksey, J., Taylor, J. S. H., & Bishop, D. V. M. (2010). A longitudinal investigation of early reading and language skills in children with poor reading comprehension. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *51*(9), 1031–1039. <https://doi.org/10.1111/j.1469-7610.2010.02254.x>

155. National Center for Education Statistics. (2011). *ECLS-K Sample Design, Weights, Variance, and Missing Data*. https://nces.ed.gov/training/datauser/ECLS-K_04/assets/ECLK_04_slides.pdf
156. Noble, K. G., Farah, M. J., & McCandliss, B. D. (2006). Socioeconomic background modulates cognition-achievement relationships in reading. *Cognitive Development, 21*(3), 349–368. <https://doi.org/10.1016/j.cogdev.2006.01.007>
157. Nollenberger, N., Rodríguez-Planas, N., & Sevilla, A. (2016). The Math Gender Gap: The Role of Culture. *American Economic Review, 106*(5), 257–261. <https://doi.org/10.1257/aer.p20161121>
158. Nores, M., & Barnett, W. S. (2010). Benefits of early childhood interventions across the world: (Under) Investing in the very young. *Economics of Education Review, 29*(2), 271–282. <https://doi.org/10.1016/j.econedurev.2009.09.001>
159. Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., Bar-Anan, Y., Bergh, R., Cai, H., Gonsalkorale, K., Kesebir, S., Maliszewski, N., Neto, F., Olli, E., Park, J., Schnabel, K., Shiomura, K., Tulbure, B. T., Wiers, R. W., ... Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences, 106*(26), 10593–10597. <https://doi.org/10.1073/pnas.0809921106>
160. Odic, D., Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Developmental Change in the Acuity of Approximate Number and Area Representations. *Developmental Psychology, 49*(6), 1103–1112. <https://doi.org/10.1037/a0029472>
161. OECD. (2013). *Synergies for Better Learning: An international*

Perspective on Evaluation and Assessment. OECD Publishing.
http://www.oecd.org/education/school/Synergies%20for%20Better%20Learning_Summary.pdf

162. OECD. (2015). *The ABC of Gender Equality in Education: Aptitude, Behaviour, Confidence | READ online.* <https://doi.org/10.1787/9789264229945-en>
163. OECD. (2018a). *Is the Last Mile the Longest? Economic Gains from Gender Equality in Nordic Countries.* <https://doi.org/10.1787/6cda329d-en>
164. OECD. (2018b). *PISA - Note per country: France.* OECD. https://www.oecd.org/pisa/publications/PISA2018_CN_FRA_FRE.pdf
165. OECD. (2019a). *PISA 2018 Assessment and Analytical Framework.* OECD. <https://doi.org/10.1787/b25efab8-en>
166. OECD. (2019b). *PISA 2018 Results (Volume I): What Students Know and Can Do.* OECD. <https://doi.org/10.1787/5f07c754-en>
167. OECD. (2019c). *PISA 2018 Results (Volume II): Where All Students Can Succeed.* OECD. <https://doi.org/10.1787/b5fd1b8f-en>
168. OECD. (2019d). *PISA 2018 Results (Volume III): What School Life Means for Students' Lives.* OECD. <https://doi.org/10.1787/acd78851-en>
169. OECD. (2011). *Singapore: Rapid improvement followed by strong performance.* <https://www.oecd.org/pisa/48758240.pdf>
170. Ozernov-Palchik, O., Norton, E. S., Sideridis, G., Beach, S. D., Wolf, M., Gabrieli, J. D. E., & Gaab, N. (2017). Longitudinal stability of pre-reading skill profiles of kindergarten children: Implications for early screening and theories

of reading. *Developmental Science*, 20(5), e12471.
<https://doi.org/10.1111/desc.12471>

171. Penner, A. M., & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research*, 37(1), 239–253. <https://doi.org/10.1016/j.ssresearch.2007.06.012>
172. Perfetti, C., & Stafura, J. (2014). Word Knowledge in a Theory of Reading Comprehension. *Scientific Studies of Reading*, 18(1), 22–37. <https://doi.org/10.1080/10888438.2013.827687>
173. Peterson, R. L., & Pennington, B. F. (2012). Developmental dyslexia. *The Lancet*, 379(9830), 1997–2007. [https://doi.org/10.1016/S0140-6736\(12\)60198-6](https://doi.org/10.1016/S0140-6736(12)60198-6)
174. Pinto, G., Bigozzi, L., Tarchi, C., Vezzani, C., & Accorti Gamannossi, B. (2016). Predicting Reading, Spelling, and Mathematical Skills: A Longitudinal Study From Kindergarten Through First Grade. *Psychological Reports*, 118(2), 413–440. <https://doi.org/10.1177/0033294116633357>
175. Pinto, S. & Sato, M (Eds). (2016). Les premières étapes de l’acquisition du langage, Kabdebon, C. & Dehaene-Lambertz, G. In *Traité de Neurolinguistique, du cerveau au langage* (De Boeck-Solal, p. 416).
176. PIRLS and TIMSS. (2015a). *Student Achievement in Mathematics – TIMSS 2015 and TIMSS Advanced 2015 International Results*. <http://timssandpirls.bc.edu/timss2015/international-results/timss-2015/mathematics/student-achievement/>
177. PIRLS and TIMSS. (2016a). *What Makes a Good Reader: International Findings from PIRLS 2016 – PIRLS 2016*. <http://timssandpirls.bc.edu/pirls2016/international-results/pirls/summary/>

178. PIRLS and TIMSS. (2011). *Relationships Among Reading, Mathematics, and Science Achievement at the Fourth Grade—Implications for Early Learning*. <https://timssandpirls.bc.edu/timsspirls2011/international-database.html>
179. PIRLS and TIMSS. (2015b). *Student Achievement in Sciences—TIMSS 2015 and TIMSS Advanced 2015 International Results*. <http://timssandpirls.bc.edu/timss2015/international-results/timss-2015/science/student-achievement/>
180. PIRLS and TIMSS. (2016b). *Student Achievement in reading – PIRLS 2016*. <http://timssandpirls.bc.edu/pirls2016/international-results/pirls/student-achievement/>
181. *PISA 2012 results*. (2012). OECD Publishing. <https://www.oecd.org/pisa/keyfindings/PISA-2012-results-france.pdf>
182. *Programmes et horaires à l'école maternelle*. (n.d.). Ministère de l'Éducation Nationale et de la Jeunesse. Retrieved October 19, 2023, from <https://www.education.gouv.fr/programmes-et-horaires-l-ecole-maternelle-4193>
183. Psyridou, M., Tolvanen, A., Lerkkanen, M.-K., Poikkeus, A.-M., & Torppa, M. (2020). Longitudinal Stability of Reading Difficulties: Examining the Effects of Measurement Error, Cut-Offs, and Buffer Zones in Identification. *Frontiers in Psychology, 10*, 2841. <https://doi.org/10.3389/fpsyg.2019.02841>
184. Quinn, J. M., Wagner, R. K., Petscher, Y., & Lopez, D. (2015). Developmental Relations Between Vocabulary Knowledge and Reading Comprehension: A Latent Change Score Modeling Study. *Child Development, 86*(1), 159–175. <https://doi.org/10.1111/cdev.12292>
185. Radlowski, E. C., & Johnson, R. W. (2013). Perinatal iron deficiency and

- neurocognitive development. *Frontiers in Human Neuroscience*, 7, 585.
<https://doi.org/10.3389/fnhum.2013.00585>
186. Raudenbush, S. W., Hernandez, M., Goldin-Meadow, S., Carrazza, C., Foley, A., Leslie, D., Sorkin, J. E., & Levine, S. C. (2020). Longitudinally adaptive assessment and instruction increase numerical skills of preschool children. *Proceedings of the National Academy of Sciences*, 117(45), 27945–27953. <https://doi.org/10.1073/pnas.2002883117>
187. Rayner K. (2016). *So Much to Read, So Little Time: How Do We Read, and Can Speed Reading Help?* <https://psycnet.apa.org/record/2005-11115-001>
188. Régner, I., Thinus-Blanc, C., Netter, A., Schmader, T., & Huguet, P. (2019). Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour*, 3(11), 1171–1179. <https://doi.org/10.1038/s41562-019-0686-3>
189. Reynvoet, B., De Smedt, B., & Van den Bussche, E. (2009). Children's representation of symbolic magnitude: The development of the priming distance effect. *Journal of Experimental Child Psychology*, 103(4), 480–489. <https://doi.org/10.1016/j.jecp.2009.01.007>
190. Ribeiro, L. A., Zachrisson, H. D., Schjolberg, S., Aase, H., Rohrer-Baumgartner, N., & Magnus, P. (2011). Attention problems and language development in preterm low-birth-weight children: Cross-lagged relations from 18 to 36 months. *BMC Pediatrics*, 11(1), 59. <https://doi.org/10.1186/1471-2431-11-59>
191. Robinson, J. P., & Lubienski, S. T. (2011). The Development of Gender Achievement Gaps in Mathematics and Reading During Elementary and Middle School: Examining Direct Cognitive Assessments and Teacher Ratings. *American Educational Research Journal*, 48(2), 268–302.

<https://doi.org/10.3102/0002831210372249>

192. Robinson-Cimpian, J. P., Lubienski, S. T., Ganley, C. M., & Copur-Gencturk, Y. (2014). Teachers' perceptions of students' mathematics proficiency may exacerbate early gender gaps in achievement. *Developmental Psychology*, *50*(4), 1262–1281. <https://doi.org/10.1037/a0035073>
193. Rocher, T. (2016). Construction d'un indice de position sociale des élèves. *Education et Formations*, *90*, 5–27.
194. Ross, E. J., Graham, D. L., Money, K. M., & Stanwood, G. D. (2015). Developmental Consequences of Fetal Exposure to Drugs: What We Know and What We Still Must Learn. *Neuropsychopharmacology*, *40*(1), 61–87. <https://doi.org/10.1038/npp.2014.147>
195. Roth, F. P., Speece, D. L., & Cooper, D. H. (2002). A Longitudinal Analysis of the Connection Between Oral Language and Early Reading. *The Journal of Educational Research*, *95*(5), 259–272. <https://doi.org/10.1080/00220670209596600>
196. Schneider, M., Beeres, K., Coban, L., Merz, S., Schmidt, S. S., Stricker, J., & Smedt, B. D. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science*, *20*(3), e12372. <https://doi.org/10.1111/desc.12372>
197. Schneider, M., Merz, S., Stricker, J., De Smedt, B., Torbeyns, J., Verschaffel, L., & Luwel, K. (2018). Associations of Number Line Estimation With Mathematical Competence: A Meta-analysis. *Child Development*, *89*(5), 1467–1484. <https://doi.org/10.1111/cdev.13068>
198. Shanahan, T., & Lonigan, C. J. (2010). The National Early Literacy Panel: A Summary of the Process and the Report. *Educational Researcher*, *39*(4),

279–285.

199. Shinwell, J., & Defeyter, M. A. (2017). Investigation of Summer Learning Loss in the UK—Implications for Holiday Club Provision. *Frontiers in Public Health*, *5*. <https://www.frontiersin.org/articles/10.3389/fpubh.2017.00270>
200. Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science*, *14*(3), 237–243. <https://doi.org/10.1111/1467-9280.02438>
201. Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children’s numerical development. *Developmental Science*, *11*(5), 655–661. <https://doi.org/10.1111/j.1467-7687.2008.00714.x>
202. Smith, E., & Reimer, D. (2023). Understanding gender inequality in children’s reading behavior: New insights from digital behavioral data. *Child Development*. <https://doi.org/10.1111/cdev.14001>
203. Smith, H. (2010). *Individuell Utvecklingsplan (IUP) med skriftliga omdömen ur ett elevperspektiv – i den senare delen av grundskolan*. <https://urn.kb.se/resolve?urn=urn:nbn:se:hh:diva-5840>
204. Smith, R., Snow, P., Serry, T., & Hammond, L. (2021). The Role of Background Knowledge in Reading Comprehension: A Critical Review. *Reading Psychology*, *42*(3), 214–240. <https://doi.org/10.1080/02702711.2021.1888348>
205. Snowling, M. J., Hulme, C., & Nation, K. (2020). Defining and understanding dyslexia: Past, present and future. *Oxford Review of Education*, *46*(4), 501–513. <https://doi.org/10.1080/03054985.2020.1765756>

206. Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science?: A critical review. *The American Psychologist*, *60*(9), 950–958. <https://doi.org/10.1037/0003-066X.60.9.950>
207. Spencer, M., Quinn, J. M., & Wagner, R. K. (2014). Specific Reading Comprehension Disability: Major Problem, Myth, or Misnomer? *Learning Disabilities Research & Practice: A Publication of the Division for Learning Disabilities, Council for Exceptional Children*, *29*(1), 3–9. <https://doi.org/10.1111/ldrp.12024>
208. Sprenger-Charolles L. & Colé. (2013). *Lecture et dyslexie; approche cognitive (2e édition)*—Liliane Sprenger-Charolles, Pascale Colé (Dunod). Librairie Gallimard PARIS. <https://www.librairie-gallimard.com/livre/9782100582921-lecture-et-dyslexie-approche-cognitive-2e-edition-liliane-sprenger-charolles-pascale-cole/>
209. Sprenger-Charolles, L., Siegel, L. S., Béchenec, D., & Serniclaes, W. (2003). Development of phonological and orthographic processing in reading aloud, in silent reading, and in spelling: A four-year longitudinal study. *Journal of Experimental Child Psychology*, *84*(3), 194–217. [https://doi.org/10.1016/s0022-0965\(03\)00024-9](https://doi.org/10.1016/s0022-0965(03)00024-9)
210. Sprugevica, I., & Høien, T. (2003). Early phonological skills as a predictor of reading acquisition: A follow-up study from kindergarten to the middle of grade 2. *Scandinavian Journal of Psychology*, *44*(2), 119–124. <https://doi.org/10.1111/1467-9450.00329>
211. Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. <https://doi.org/10.1598/RRQ.21.4.1>
212. Starkey, G. S., & McCandliss, B. D. (2014). The emergence of

“groupitizing” in children’s numerical cognition. *Journal of Experimental Child Psychology*, 126, 120–137. <https://doi.org/10.1016/j.jecp.2014.03.006>

213. Stoevenbelt, A. H., Wicherts, J. M., Flore, P. C., Phillips, L. A., Pietschnig, J., Verschuere, B., Voracek, M., & Schwabe, I. (2023). Are Speeded Tests Unfair? Modeling the Impact of Time Limits on the Gender Gap in Mathematics. *Educational and Psychological Measurement*, 83(4), 684–709. <https://doi.org/10.1177/00131644221111076>
214. Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology*, 38(6), 934–947. <https://doi.org/10.1037/0012-1649.38.6.934>
215. Stout, J. G., Dasgupta, N., Hunsinger, M., & McManus, M. A. (2011). STEMing the tide: Using ingroup experts to inoculate women’s self-concept in science, technology, engineering, and mathematics (STEM). *Journal of Personality and Social Psychology*, 100(2), 255–270. <https://doi.org/10.1037/a0021385>
216. Thomas, D. P., Hopwood, B., Hatisaru, V., & Hicks, D. (2022). Gender differences in reading and numeracy achievement across the school years. *Australian Educational Researcher*, 1–26. <https://doi.org/10.1007/s13384-022-00583-8>
217. Thompson, B. L., Levitt, P., & Stanwood, G. D. (2009). Prenatal exposure to drugs: Effects on brain development and implications for policy and education. *Nature Reviews. Neuroscience*, 10(4), 303–312. <https://doi.org/10.1038/nrn2598>
218. Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across

grade levels. *Journal of Research in Reading*, 32(4), 383–401.
<https://doi.org/10.1111/j.1467-9817.2009.01401.x>

219. Torgesen, J. K. (2009). The Response to Intervention Instructional Model: Some Outcomes From a Large-Scale Implementation in Reading First Schools. *Child Development Perspectives*, 3(1), 38–40.
<https://doi.org/10.1111/j.1750-8606.2009.00073.x>
220. Torgesen, J. K., & Davis, C. (1996). Individual difference variables that predict response to training in phonological awareness. *Journal of Experimental Child Psychology*, 63(1), 1–21. <https://doi.org/10.1006/jecp.1996.0040>
221. Tsui, J. M., & Mazzocco, M. M. M. (2007). Effects of Math Anxiety and Perfectionism on Timed versus Untimed Math Testing in Mathematically Gifted Sixth Graders. *Roeper Review*, 29(2), 132–139.
<https://doi.org/10.1080/02783190709554397>
222. Vaessen, A., Bertrand, D., Tóth, D., Csépe, V., Faísca, L., Reis, A., & Blomert, L. (2010). Cognitive development of fluent word reading does not qualitatively differ between transparent and opaque orthographies. *Journal of Educational Psychology*, 102(4), 827–842. <https://doi.org/10.1037/a0019465>
223. Van Mier, H. I., Schleepen, T. M. J., & Van den Berg, F. C. G. (2019). Gender Differences Regarding the Impact of Math Anxiety on Arithmetic Performance in Second and Fourth Graders. *Frontiers in Psychology*, 9, 2690.
<https://doi.org/10.3389/fpsyg.2018.02690>
224. Vandormael, C., Schoenhals, L., Hüppi, P. S., Filippa, M., & Borradori Tolsa, C. (2019). Language in Preterm Born Children: Atypical Development and Effects of Early Interventions on Neuroplasticity. *Neural Plasticity*, 2019, 6873270. <https://doi.org/10.1155/2019/6873270>

225. Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: A meta-analysis. *Psychonomic Bulletin & Review*, *18*(2), 267–277. <https://doi.org/10.3758/s13423-010-0042-0>
226. Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. <https://doi.org/10.1037/a0036620>
227. Wang, M.-T., & Degol, J. L. (2017). Gender Gap in Science, Technology, Engineering, and Mathematics (STEM): Current Knowledge, Implications for Practice, Policy, and Future Directions. *Educational Psychology Review*, *29*(1), 119–140. <https://doi.org/10.1007/s10648-015-9355-x>
228. Watson, A., Timperio, A., Brown, H., Best, K., & Hesketh, K. D. (2017). Effect of classroom-based physical activity interventions on academic and physical activity outcomes: A systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity*, *14*(1), 114. <https://doi.org/10.1186/s12966-017-0569-9>
229. Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., ... Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364–369. <https://doi.org/10.1038/s41586-019-1466-y>
230. Zablotzky, B., Black, L. I., Maenner, M. J., Schieve, L. A., Danielson, M. L., Bitsko, R. H., Blumberg, S. J., Kogan, M. D., & Boyle, C. A. (2019). Prevalence and Trends of Developmental Disabilities among Children in the US: 2009-2017. *Pediatrics*, *144*(4), e20190811. <https://doi.org/10.1542/peds.2019-0811>

231. Zhou, S., Rosenthal, D. G., Sherman, S., Zelikoff, J., Gordon, T., & Weitzman, M. (2014). Physical, Behavioral, and Cognitive Effects of Prenatal Tobacco and Postnatal Secondhand Smoke Exposure. *Current Problems in Pediatric and Adolescent Health Care*, 44(8), 219–241. <https://doi.org/10.1016/j.cppeds.2014.03.007>
232. Ziegler, J. C. (2018). Différences inter-linguistiques dans l'apprentissage de la lecture. *Langue française*, 199(3), 35–49. <https://doi.org/10.3917/lf.199.0035>
233. Ziegler, J. C., Bertrand, D., Tóth, D., Csépe, V., Reis, A., Faisca, L., Saine, N., Lyytinen, H., Vaessen, A., & Blomert, L. (2010). Orthographic depth and its impact on universal predictors of reading: A cross-language investigation. *Psychological Science*, 21(4), 551–559. <https://doi.org/10.1177/0956797610363406>
234. Zorman †, M., Bressoux, P., Bianco, M., Lequette, C., Pouget, G., & Pourchet, M. (2015). «PARLER»: Un dispositif pour prévenir les difficultés scolaires. *Revue française de pédagogie. Recherches en éducation*, 193, Article 193. <https://doi.org/10.4000/rfp.4890>

Data and materials availability and transfer agreements (MTAs)

For confidentiality reasons, the raw data were not shared in public but accessible through a data secured convention established with the DEPP in France.

For reproducibility, code and models that were used to generate results, text, figures, and tables both in the main text and in the supplementary information are available on the following github repository: [PauMdlm/Education](https://github.com/PauMdlm/Education).

This work was supported by the Fondation pour la Recherche Médicale, grant number « FDM201906008580 », to « Pauline MARTINOT », the Collège de France, INSERM, CEA and the doctoral school FIRE (Frontières de l'Innovation en Recherche et Education (FIRE) – ED 474), Bettencourt Programme, University of Paris-Cité, France.

Figures list

Introduction

- **Figure 1.** Reading acquisition brain mechanisms (from Dehaene-Lambertz G. et al., PLoS Bio, 2018; Monzalvo et al., Neuroimage, 2012). Learning to read is a new path (vs. oral language pathway in blue) to access meaning through vision (in red).
- **Figure 2.** PISA (left) and PIRLS (right) results since 2000-2001 comparing France and OECD/European countries.
- **Figure 3.** Math PISA results since 2000

Chapter 2

- **Figure 4.** EVALAIDE, design of a country-wide longitudinal assessment of cognitive skills.
- **Figure 5.** Effect of age in Math and Language, reproduced in 2018 (main figure), 2019, 2020 and 2021.
- **Figure 6.** Effect of school categories and SES score in Math and language, reproduced in 2018 (main figure), 2019, 2020 and 2021.
- **Figure 7.** Correlation matrix between all class and school level variables at T1 in 2018, on the total population.
- **Figure 8.** Correlation matrix for cognitive tests both in math and language at T1
- **Figure S1.** Indirect estimation of internal reliability between tests using correlation matrices in math and language.
- **Figure S2.** Math and language at T1 considering all ages available in advance-in-age and late-in-age children.
- **Figure S3.** Panel of Math and Language in function of Age at T1, T2 and T3 in 2018, 2019, 2020 and 2021.
- **Figure S4.** Panel of Math and Language in function of SES and types of school at T1, T2, and T3 in 2018, 2019, 2020 and 2021.

Chapter 3

- **Figure 9.** Distribution of normalized language assessments at T1, T2 and T3 (range 0-100) in 2018.
- **Figure 10.** Distribution of raw scores in reading words and texts at T2 and T3 in 2018 for all typical-age children.
- **Figure 11.** Correlations between language exercises at T1 in 2018.
- **Figure 12.** Correlations between language exercises at T2 in 2018.
- **Figure 13.** Correlations between language exercises at T3 in 2018.
- **Figure 14.** Correlations between language exercises at T1, T2 and T3, results for the 2018 cohort.
- **Figure 15.** Principal component analysis of language items at T1 in 2018.
- **Figure 16.** Principal component analysis of language items at T2 in 2018.
- **Figure 17.** Principal component of language items at T3 in 2018.
- **Figure 18.** Summarize of three PCA in language at T1, T2 and at T3 in 2018.
- **Figure 19.** Predictors for reading words and texts in one minute in second grade.
- **Figure 20.** Predictors for reading comprehension in second grade.
- **Figure 21.** Levels in language at T1, T2, T3 in 2018, 2019, 2020 and 2021.
- **Figure 22.** Comparing Reading abilities for words and texts at T2 and at T3 (A) in 2018, 2019, 2020 and 2021 and (B) between school categories in 2018, 2019, 2020 and 2021.
- **Figure 23.** Comparing Reading comprehension of sentences and of texts at T3 (A) in 2018, 2019, 2020 and 2021 and (B) between school categories in 2018, 2019, 2020 and 2021.
- **Figure 24.** Level trajectories in language in 2018. Language was the mean of all assessments of a period of time, all assessments were normalized.
- **Figure 25.** Reading comprehension at T3 in function of age in month.
- **Figure 26.** Reading comprehension average scores at T3 (in percentage of success) per school categories and SES score.
- **Figure 27.** Distribution of raw scores in reading comprehension of sentences and texts at T3 in 2018 for all typical-age children, comparing children of the lowest quintile in reading comprehension at T3 vs. the others.

- **Figure 28.** Principal component analysis of language assessments at T1 among children facing difficulties in reading comprehension at T3.
- **Figure 29.** Comparing all three PCA between (1) children with oral comprehension difficulties at T1, (2) children with meta phonology difficulties at T1 and (3) children with difficulties in reading comprehension at T3.
- **Figure S5.** Language assessments' distributions in 2018, 2019, 2020 and 2021 at T1
- **Figure S6.** Language assessments' distributions in 2018, 2019, 2020 and 2021 at T2
- **Figure S7.** Language assessments' distributions in 2018, 2019, 2020 and 2021 at T3
- **Figure S8.** Correlation panels for language at T1 in 2018, 2019, 2020 and 2021
- **Figure S9.** Correlation panels for language at T2 in 2018, 2019, 2020 and 202
- **Figure S10.** Correlation panels for language at T3 in 2018, 2019, 2020 and 2021
- **Figure S11.** PCA results for language at T1, T2 and T3 in 2019
- **Figure S12.** PCA results for language at T1, T2 and T3 in 2020
- **Figure S13.** PCA results for language at T1, T2 ans T3 in 2021

Chapter 4

- **Figure 30.** [Fig 1] Rapid emergence of the math gender gap found in the national program Evalaide. **Figure 30C.** Distribution of ranks in math among boys and girls, showing an initially higher density of boys in both high- and low-performers, quickly shifting to a large advantage in favor of boys and, reproducible in 2019, 2020 and 2021.
- **Figure 31.** [Fig 2] A new element: Age is not a modulator of the gender gap in math.
- **Figure 32.** [Fig 3] A massive gender gap emerges and remains at T2 and T3 after the elimination of any gender difference at T1 when implementing matching procedures.
- **Figure 33.** Quantile regression of math at T3 using math deciles (left) and math percentiles (right).

- **Figure S14.** [Fig S4] Gender gap explanatory class-I evel covariates on children with typical age in first grade and among classes with at least 30% of boys and 30% of girls.
- **Figure S15** [Fig S2] Math panels in 2018, 2019, 2020 and 2021 for figures 1 and 3.
- **Figure S16** [Fig S3] Panels for number line and problem solving in 2018, 2019, 2020 and 2021.
- **Figure S17** [Fig S5] Language panels in 2018, 2019, 2020 and 2021.

Tables list

Introduction

- **Table 1.** TIMSS results in Math from 2015 to 2019

Chapter 2

- **Table 2.** Total number of first graders tested in 2018, 2019, 2020 and 2021.
- **Table 3.** Characteristics of the pilot studies population in 2018 (n = 9797).
- **Table 4.** Description of the collected variables at the individual level, at the class level and at the school level.
- **Table 5.** Summary of the tests performed at each time period.
- **Table 6.** Overview of missing values among the four different whole-population cohorts and their data management.
- **Table 7.** Sensitivity analysis presented for year 2018, comparing (1) the selected population (i.e., imputed, using *mice*) to the non-imputed population (i.e., model 1) and comparing (2) the imputed population to the non-imputed population withdrawn from all missing values (i.e., model 2).
- **Table 8.** Description of characteristics of the population continuous variables in 2018 (n = 586,936)
- **Table 9.** Description of characteristics of the population categorial variables in 2018 (n = 586,936)
- **Table 10.** Descriptive analysis of cognitive tests in 2018 (n = 586,936).
- **Table 11.** Description of gender and school categories for advanced and late children when beginning first grade (T1) in 2018
- **Table S1.** Description of children' characteristics and tests in 2018, 2019, 2020 and 2021.
- **Table S2.** Gender and age in function of the four school categories in 2019, 2020 and 2021.

- **Table S3.** Among the total population entering in first grade in 2018, description of their results considering their age category (Advance; typical; late) (n = 586,936).
- **Table S4.** Analyzing the effect of age in function of time on math and language results. Analyses were made among typical-in-age children. Time was defined as T1 = 0; T2 = 4; T3 = 12 months of school.
- **Table S5.** Comparison of linear effect slopes between age categories and math and language at T1, T2 and T3.
- **Table S6.** Description of cognitive tests in function of school social category (Private - Public - PE - HPE) in 2018 (n = 586,936)
- **Table S7.** Analyzing the effect of school categories in function of time on math and language results. Analyses were made among typical-in-age children. Time was defined as T1 = 0; T2 = 4; T3 = 12 months of school.
- **Table S8.** Analysis zooming on T1-T2 and on T2-T3 in math and language, and the school categories effects on the results.

Chapter 3

- **Table 12.** Multilevel models for different language exercises at T2 in 2018.
- **Table 13.** Multilevel regression model testing “reading at T3” as the main outcome among children of typical age.
- **Table 14.** Multilevel regression model testing “reading comprehension at T3” as the main outcome and all the T1 assessments.
- **Table 15.** Multilevel regression model assessing Reading comprehension at T3 with composite variables at T1.
- **Table 16.** Age characteristics of children belonging to the latest quintile vs. the rest, and vs. the best quintile in reading comprehension at T3 in 2018.
- **Table 17.** Description of children belonging to the worst reading comprehension quintile at T3 regarding their age categories and their school types
- **Table 18.** Describing children belonging to the latest quintile in reading comprehension at T3 with other 4 quintiles and with the best quintile, among typical-age children.

- **Table 19.** Description of difficulties presented by children of PE and HPE public schools that belong to the worst quintile in reading comprehension at T3 in 2018.
- **Table 20.** Multilevel models analyzing – among children with the highest difficulties in oral language comprehension at T1 (left model) and among children with the highest difficulties in meta phonology at T1 (right model) – the predictors at T1 associated with a better level of reading comprehension at T3 in 2018.
- **Table S9.** Multilevel regression progressive models for reading at T3 in 2018.
- **Table S10.** Multilevel progressive regression models for Reading comprehension at T3 in 2018.
- **Table S11.** ANOVA differencing Language at T1, T2 and T3, reading at T2 and T3 and reading comprehension at T3 in 2018, 2019, 2020 and 2021.
- **Table S12.** ANOVA testing for reading abilities (i.e., decoding) between T2 and T3, and reading comprehension at T3 for all cohorts in 2018, 2019, 2020 and 2021.
- **Table S13.** Results for reading comprehension of sentences at T3 in 2018, among children belonging to the lowest quintile of level in reading comprehension at T3.
- **Table S14.** Results for reading comprehension of sentences at T3 in 2018, among children belonging to the other four quintiles of level in reading comprehension at T3.
- **Table S15.** Results for reading comprehension of texts at T3 in 2018, among children belonging to the lowest quintile of level in reading comprehension at T3.
- **Table S16.** Results for reading comprehension of texts at T3 in 2018, among children belonging to the other four quintiles of level in reading comprehension at T3.
- **Table S17.** Describing and comparing results in language subtests at T1, T2 and T3 of children belonging to the latest quintile of level vs. the other four quintiles and vs. children belonging to the best quintile in reading comprehension at T3.

Chapter 4

- **Table 21.** Cohen's D effect size for gender gaps in 2018, 2019, 2020 and 2021 among children of normal age at T1.
- **Table 22.** Analysis of fixed and random factors associated with children's math scores at T3, using multilevel mixed regression models on normal-aged children ($n_{\text{Total}} = 1.783.666$ children).
- **Table S18.** Percent success in each test for all four cohorts (2018, 2019, 2020 and 2021), separately for each gender, among children of normal age at T1.
- **Table S19.** Progressive Multilevel model for Math at T3 among children of typical age in first grade ($n = 569,771$).
- **Table S20.** Multilevel regression model for Language at T3 among children of typical age at T1. Language proficiency at T3 differed significantly from those in math proficiency at T3 for all cohorts.
- **Table S21.** Outliers management, see Table 6 in Chapter 2.
- **Table S22.** Multilevel regression model for Math at T1 among children of normal age at T1
- **Table S23.** Multilevel regression model for Math at T2 among children of normal age at T1.
- **Table S24.** Matching experiments and scenarii
- **Table S25.** Results of causal inference methods applied to the gender gap in Math between T1 and T3. CI = 95% confidence interval.
- **Table S26.** T-tests measuring the differences of gender gaps magnitude between 2018, 2019, 2020 and 2021
- **Table S27.** Fixed and random effects for model 1 in quantile regression of Math at T3 in 2018.
- **Table S28.** Fixed and random effects for model 2 in quantile regression of Math at T3 in 2018.

Résumé long en Français

Récemment, la France a mis en place des évaluations nationales visant à mesurer précisément les acquisitions et les difficultés d'apprentissage auxquelles les enfants sont confrontés tout au long de leur première année d'école jusqu'à leur deuxième année, en se basant sur l'évaluation cognitive des compétences en mathématiques et en langage. Chaque année, environ 750 000 enfants ont effectué 46 exercices pour évaluer leurs performances, couvrant un total de 2,9 millions d'enfants entre 2018 et 2022. En analysant cet ensemble riche de données sur l'ensemble de la population, cette thèse visait à fournir une meilleure compréhension des conditions qui favorisent ou entravent l'acquisition de l'apprentissage académique chez les enfants. Pour ce faire, nous avons mené une série d'études en utilisant des données longitudinales provenant de quatre cohortes françaises représentatives de la population, évaluant l'influence relative d'une large gamme de facteurs individuels, de classe et d'établissement sur différents aspects de la réussite scolaire au primaire.

Tout d'abord, nous avons décrit les données obtenues dans le programme national et identifié les prédicteurs des compétences en lecture et en compréhension de la lecture. Nous avons constaté que la différence d'âge avait un impact très fort sur les résultats des enfants, tant en ce qui concerne les tests de langage que les tests de mathématiques, indépendamment de la catégorie scolaire, et qu'une relation linéaire positive liait l'âge en mois et les résultats en mathématique et en langage chez les enfants d'âge typique à l'entrée au CP (i.e., 69 à 80 mois d'âge ou une entrée en CP l'année de leurs 6 ans). Par ailleurs, étant donné que les enfants en avance (i.e., < 69 mois d'âge en CP) ou en retard (i.e., > 80 mois d'âge en CP) peuvent correspondre à des antécédents biologiques, environnementaux et culturels spécifiques, nous avons décidé de ne pas les explorer dans cette thèse et nous nous sommes concentrés sur les enfants ayant l'âge typique en CP en France. Indépendamment de la catégorie scolaire et de l'âge en mois, les enfants d'âge typique en CP ont mieux performé pour chaque tâche, tant en langage qu'en mathématiques, à T1, T2 et T3 lorsqu'ils fréquentaient des écoles privées, par rapport aux écoles publiques, avec un gradient de performance comme suit : écoles privées > écoles publiques classiques > écoles publiques en éducation prioritaire (REP) > écoles publiques en éducation prioritaire

renforcée (REP+). Les exercices à durée limitée et les exercices difficiles favorisaient les enfants issus de milieux socio-économiques plus privilégiés. Les différences entre les catégories sociales des élèves se sont réduites après 4 mois de scolarité, tandis que les inégalités sociales se sont creusées entre T2 et T3, où il y a eu une pause de 2,5 mois de vacances d'été. C'est au sein des écoles REP et REP+ que les enfants ont le plus progressé en langage et en mathématiques entre T1 et T2. Cependant, c'est également parmi ces sous-groupes qu'ils ont fait moins de progrès en langage et en mathématiques entre T2 et T3. En ce qui concerne le genre, davantage de filles étaient en avance d'une année à l'école en première année (55-60 %) par rapport aux garçons. La proportion de garçons était plus élevée dans la catégorie avec un an de retard à l'école en CP par rapport aux filles (60 % des enfants qui avaient un an de retard étaient des garçons). Dans les écoles privées, le double d'enfants avaient un an d'avance (1 %) par rapport aux autres catégories d'écoles (0,5 %). Il y avait des disparités de niveau plus importantes (c'est-à-dire, une hétérogénéité de niveau de classe) au sein de la même classe, à la fois en langage et en mathématiques, parmi les classes ayant des scores socio-économiques plus faibles. Les classes avec un niveau socio-économique plus élevé avaient des performances plus homogènes par classe et des moyennes de classe plus élevées en mathématiques et en langage à T1.

En plus de l'âge, du niveau socioéconomique, du type d'école et du genre, et grâce à ces évaluations linguistiques précises effectuées dès le début de la première année de l'école primaire (c'est-à-dire des évaluations de compréhension orale, des évaluations métaphonologiques et des compétences de décodage), nous avons exploré le pouvoir prédictif des exercices de linguistique en compréhension de la lecture en deuxième année. Plus particulièrement, nous avons pu identifier les besoins d'apprentissage parmi les différentes catégories socio-économiques des écoles. Notamment, nous avons pu confirmer que les prédicteurs des compétences en lecture et de la compréhension en lecture différaient : Les prédicteurs des compétences en lecture à T3 étaient en premier lieu l'« association lettres-sons » et la « manipulation phonémique », tous deux présentant le poids prédictif le plus élevé, et en second lieu, la « manipulation des syllabes » et la « connaissance des lettres », leur poids prédictif étant similaire entre 2018, 2019, 2020 et 2021. Ce résultat contrastait avec la plupart

des recherches sur les compétences en lecture qui identifiaient la conscience phonémique comme le prédicteur le plus élevé des compétences en lecture ultérieures (c'est-à-dire le décodage des mots et des textes). D'une part, on pourrait prétendre que nous avons trouvé un poids prédictif plus élevé pour l'association lettres-sons par rapport à la manipulation phonémique en raison d'une différence dans le contenu et dans le contexte des tests, car notamment dans les études de psychologie du développement réalisées sur des populations plus petites, les tests sont très spécifiques et prennent plus de temps pour évaluer un enfant par rapport aux tests nationaux généralisés « évaluant la conscience phonémique » en contexte de classe et sont effectués en moins de 5 minutes. D'autre part, les données massives, répliquées sur quatre populations exhaustives d'enfants consécutivement, ont apporté une « nouvelle information » et ont indiqué la place importante de l'association lettres-sons comme étant plus importante que la conscience phonémique parmi les prédicteurs des compétences en lecture. De plus, comme vu précédemment, la fenêtre temporelle variait entre les différentes études, avec plus d'importance accordée aux compétences de décodage en maternelle et davantage d'importance accordée à la compréhension orale des mots en première et deuxième année. La fenêtre temporelle de notre étude (c'est-à-dire, du début de la première année et suivie pendant 12 mois) pourrait être associée à un poids prédictif dominant pour l'association lettres-sons par rapport à la manipulation phonémique.

Aussi, comme plusieurs données au niveau de la classe et de l'école ont été analysées, nous avons pu identifier le retard alarmant et les besoins des enfants ayant un score de SES plus faible, notamment en compréhension orale du langage, par rapport aux autres enfants. De plus, nous avons constaté que les disparités de niveau entre les quatre catégories d'écoles (c'est-à-dire une distance plus importante entre les écoles privées et les écoles publiques en REP et REP+) se situaient sur l'axe de la compréhension orale (pour les mots et les textes), qui dépend de l'immersion linguistique depuis la naissance, tandis que les domaines plus formels de l'enseignement (c'est-à-dire la manipulation phonémique, la manipulation syllabique et la connaissance des lettres) sont moins discriminants entre les types d'écoles.

Également, nos modèles nous ont permis d'évaluer divers paramètres liés à l'environnement d'apprentissage (c'est-à-dire, les caractéristiques de la classe) qui

influençaient, ou non, la compréhension en lecture. Les enfants, ayant des difficultés en compréhension en lecture à T3, avaient tendance à être plus jeunes (parmi les enfants typiques en âge) ou à appartenir à la catégorie des enfants en retard en âge, avaient tendance à être associés à un type d'école à faible revenu (c'est-à-dire, REP et REP+), et avaient tendance à être des garçons (57% contre 43%) par rapport à la population générale. Comme décrit précédemment, en comparant les ACP entre les trois groupes de difficultés, nous avons constaté que des dimensions très spécifiques définissaient chaque population, et n'identifiaient pas de description commune des variables associées aux trois types de difficultés. De manière remarquable, et contrairement aux mathématiques (voir chapitre 4), notre enquête a révélé qu'aucune des caractéristiques de la classe ne modifiait de manière significative le niveau de compréhension en lecture, sauf les scores de SES. Un score de SES plus élevé était associé à une meilleure compréhension en lecture à T3. Par exemple, les enfants fréquentant des écoles privées avaient quatre fois plus de chances d'appartenir au meilleur quintile de niveau en manipulation phonémique, par rapport aux enfants fréquentant des écoles REP+. Aussi, parmi les enfants ayant des difficultés, que ce soit en compréhension orale du langage ou en métaphonologie, le fait d'appartenir à un score de SES plus élevé était un prédicteur important d'un meilleur niveau de compréhension en lecture à T3. Enfin, l'hétérogénéité de la compétence linguistique à T1 au sein d'une classe était liée à un niveau plus bas de compréhension en lecture à T3. En d'autres termes, de plus grandes disparités dans la compétence linguistique au sein de la même classe à T1 entravaient la progression des enfants en compréhension en lecture. Cependant, en ce qui concerne les compétences de décodage à T3, la taille de la classe avait de l'importance : des classes plus petites et des scores de SES plus élevés étaient liés à une vitesse de lecture plus rapide à T3 (c'est-à-dire, davantage de mots lus par minute).

Enfin, le contexte unique de la pandémie de la COVID-19 a présenté une expérience naturelle, nous permettant de comparer l'année difficile de 2019, caractérisée par une absence substantielle de l'école (52 jours de congé plus une pause estivale de deux mois) et 2020, avec trois semaines de congé scolaire et la pause estivale habituelle, par rapport aux autres années. Nos résultats ont montré une baisse significative du niveau en lecture, en compréhension en lecture et en langage

en 2019, qui pourrait être attribuée à une moindre exposition à l'enseignement formel de la lecture à l'école. En d'autres termes, en comparant les niveaux linguistiques entre 2018, 2019, 2020 et 2021, plus les enfants allaient à l'école longtemps, meilleurs étaient leurs résultats. L'événement de la COVID-19 a été le seul changement majeur de l'exposition à l'éducation entre les quatre années, et comme les enfants présentaient des résultats similaires avant que la COVID-19 ne se produise (c'est-à-dire à T1 et T2) et ne différaient pas en termes de caractéristiques individuelles ou environnementales. Notamment, les vacances estivales ont été décrites dans d'autres études comme étant associées à un élargissement des écarts de SES en langage et en mathématiques. Pour résumer nos conclusions, notre étude a permis une compréhension plus nuancée des composantes essentielles nécessaires à une compréhension en lecture et à une lecture, efficaces, des paramètres importants pour la recherche future et la mise en œuvre d'interventions de renforcement en langage.

Pour terminer, nous avons centré notre dernière analyse sur les différences de genre en langage et en mathématiques, en estimant l'influence de différents facteurs sur les résultats des enfants. Notamment, nous avons pu identifier que l'écart entre les genres en mathématiques est déclenché par l'école et non par l'âge.

En effet, tandis que des études internationales antérieures transversales ont conclu que l'écart entre les sexes en mathématiques apparaissait vers l'âge de 8-9 ans ou en 4e année d'école primaire, les résultats longitudinaux actuels indiquent une émergence bien plus précoce, en accord avec des résultats antérieurs sur des échantillons plus petits. Grâce aux données massives, nous avons découvert un écart entre les sexes en mathématiques, favorable aux garçons, rapidement induit et ancré après seulement 4 mois de scolarité en première année, sans lien avec l'âge des enfants et émergeant plus tôt parmi les milieux socioéconomiques élevés. Les mathématiques diffèrent donc considérablement du langage, où de grandes différences favorisant les filles existent avant la scolarisation et se développent de manière linéaire avec l'âge de l'enfant et homogène dans tous les milieux socioéconomiques. Fondamentalement, la présente étude élucide les conditions d'émergence de l'écart entre les sexes en mathématiques, qui ne reflète pas de différences entre les sexes préexistantes, ni ne nécessite une longue période d'internalisation. Les stéréotypes de genre ne sont pas internalisés lentement ni en

fonction de l'âge (c'est-à-dire que les enfants plus jeunes ont tendance à internaliser les stéréotypes de genre plus rapidement, en particulier en mathématiques, voir les pentes négatives des écarts entre les sexes en mathématiques). Au contraire, l'écart entre les sexes en mathématiques émerge de manière significative et s'approfondit après avoir été exposé à l'enseignement formel des mathématiques et à une plus longue durée d'exposition à l'école. Les attitudes des enseignants et l'enseignement formel des mathématiques peuvent jouer un rôle important, s'ils interagissent différemment avec les garçons et les filles, transmettent leur anxiété mathématique aux filles ou encouragent les efforts des filles en lecture plus qu'en mathématiques. Cependant, le début de la scolarisation peut également entraîner un changement dans les attitudes des enfants eux-mêmes, des parents, des membres de la famille et d'autres professionnels. La simple croyance que les garçons et les filles ont des intérêts et des capacités différents peut renforcer les disparités entre les sexes. Enfin, les filles peuvent manifester une anxiété plus importante en mathématiques et donc éviter la compétition, un comportement qui pourrait expliquer pourquoi, parmi tous les exercices de mathématiques et de langage, l'avantage masculin est plus prononcé pour les tests plus difficiles, nouveaux ou complexes qui sollicitent les fonctions exécutives. Dans la dernière étude PISA, 21 pays ont réussi à réduire l'écart entre les sexes en mathématiques entre 2009 et 2018, et dans 5 pays, cela a été réalisé grâce à des améliorations du niveau des filles en mathématiques. Les résultats actuels suggèrent que les interventions devraient avoir lieu tôt dans le programme scolaire. L'égalité socio-culturelle, politique et éducative transnationale chez les adultes ne prédit pas nécessairement une réduction de l'écart entre les sexes en mathématiques. D'un point de vue politique, lutter contre l'écart entre les sexes en mathématiques dès les premières étapes (maternelle ou CP) peut être plus rentable et efficace, car cela se produit avant que les filles ne perdent confiance en leurs compétences en mathématiques et ne deviennent réticentes à l'information contre-stéréotypée. Quels facteurs devraient être ciblés ? Nos résultats suggèrent que les variables de niveau de classe telles que la taille de la classe, le ratio des sexes, l'hétérogénéité du niveau en mathématiques ou le sexe de l'élève en tête de classe n'exercent qu'une faible influence. Les écoles ou classes uniquement réservées à un sexe sont également inefficaces. L'intervention la plus importante peut consister à convaincre tous les

enfants que les mathématiques en valent la peine pour les deux sexes. Des recherches antérieures suggèrent que les actions suivantes peuvent être efficaces : soutenir les parents, les informer et encourager le développement d'un environnement d'apprentissage stimulant à la maison ; encourager les deux sexes à jouer à des jeux similaires pour la pensée spatiale; encourager les évaluations et les pratiques justes en matière de genre des enseignants, telles que poser des questions aux filles et aux garçons également souvent pendant les cours de mathématiques et de sciences; exposer les enfants à des modèles masculins et féminins auxquels ils peuvent s'identifier; fournir aux filles des moyens de faire face au stress de la compétition et à l'anxiété en mathématiques; et les informer sur l'impact

Dans l'ensemble, nous avons discuté des preuves récentes en matière de compréhension de la lecture et des compétences en mathématiques, à un niveau populationnel. Ces approches scientifiques peuvent conduire à la conception de programmes d'apprentissage ciblés, tant pour les apprenants normaux que pour les apprenants à risque de développer des difficultés, ainsi que pour les apprenants rencontrant des difficultés d'apprentissage en langage et/ou en mathématiques. Tout au long de cette thèse, nous présentons des exemples de la manière dont les données massives et les analyses basées sur les sciences cognitives peuvent aider les apprenants et informer le système éducatif national. En parallèle de chaque approche, nous discutons des limites de l'approche et proposons des solutions pour les surmonter.