Essays on the determinants of student achievement in France and the US : teacher evaluation, teaching practices and social interactions in middle school

Simon Briole

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Ecole des hautes études en sciences sociales

Essais sur les déterminants de la réussite scolaire en France et aux Etats-Unis : pratiques pédagogiques, évaluation des enseignants et interactions sociales au collège

Soutenue par
Simon BRIOLE
Le 21 juin 2019

Ecole doctorale n° 465
Economie Panthéon Sorbonne

Spécialité
Analyse et Politique Économiques

Composition du jury :

Élise HUILLERY Université Paris-Dauphine	*Rapporteur*
Sandra MCNALLY London School of Economics	*Rapporteur*
Manon GARROUSTE Université de Lille	*Examinateur*
Marc GURGAND CNRS-Paris School of Economics	*Examinateur*
Éric MAURIN EHESS- Paris School of Economics	*Directeur de thèse*

L'ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES

École des Hautes Études en Sciences Sociales

THÈSE DE DOCTORAT

Pour l'obtention du grade de docteur en Sciences Économiques
de l'École des Hautes Études en Sciences Sociales

Présentée et soutenue publiquement le 21 juin 2019 par :

Simon BRIOLE

Essais sur les déterminants de la réussite scolaire en France et aux Etats-Unis : évaluation des enseignants, pratiques pédagogiques et interactions sociales au collège

Directeur de thèse : Éric MAURIN

Composition du jury :

Rapporteurs :	Élise HUILLERY, Maître de Conférences à l'Université Paris-Dauphine
	Sandra MCNALLY, Professeure à la London School of Economics
Directeur :	Éric MAURIN, Directeur d'études à l'EHESS-PSE
Examinateurs :	Manon GARROUSTE, Maître de Conférences à l'Université de Lille
	Marc GURGAND, Directeur de recherche CNRS-PSE

École des Hautes Études en Sciences Sociales

PhD Thesis

Submitted to École des Hautes Études en Sciences Sociales
for the Degree of Doctor of Philosophy in Economics

Prepared and defended at Paris School of Economics on June 21, 2019 by:

Simon BRIOLE

Essays on the determinants of student achievement in France and the US: teaching practices, teacher evaluation and social interactions in middle school

Thesis Advisor : Éric MAURIN

Composition du jury :

Rapporteurs :	Élise HUILLERY, Maître de Conférences à l'Université Paris-Dauphine
	Sandra MCNALLY, Professeure à la London School of Economics
Directeur :	Éric MAURIN, Directeur d'études à l'EHESS-PSE
Examinateurs :	Manon GARROUSTE, Maître de Conférences à l'Université de Lille
	Marc GURGAND, Directeur de recherche CNRS-PSE

REMERCIEMENTS

Cette thèse est le fruit d'un long travail de recherche, qui s'est étalé sur cinq années. Cette période de ma vie a été marquée par l'alternance d'un grand enthousiasme et de moments de doute intense. Rien n'aurait été possible sans le soutien sans faille de mes proches et de mes collègues, ni sans la stimulation quotidienne née des interactions avec des personnalités inspirées et inspirantes.

En premier lieu, je voudrais remercier Éric Maurin, qui a été pour moi un excellent directeur de thèse et restera sans aucun doute la personne la plus importante et la plus influente dans mon apprentissage du métier de chercheur, tant au plan intellectuel qu'au plan de l'éthique personnelle. Je lui suis extrêmement reconnaissant pour sa patience, sa bienveillance toujours renouvelée et sa rigueur sans faille. L'un des chapitres de cette thèse est un travail que nous avons écrit en commun. Je suis particulièrement heureux de cette collaboration, qui a marqué un véritable tournant dans ma formation de doctorant.

Je voudrais également remercier Marc Gurgand, qui m'a accompagné tout au long de cette thèse en tant que membre du comité de thèse. Son enthousiasme, sa pédagogie et sa grande disponibilité ont été de précieuses ressources pour moi. Notre collaboration dans le cadre du projet *Active Citizenship* m'a également permis de découvrir une autre facette du métier de chercheur, liée à l'évaluation expérimentale d'une politique publique. Je suis infiniment reconnaissant à Éric et Marc de m'avoir permis de travailler sur ce projet pendant les deux dernières années de ma thèse, qui m'a ouvert de nouveaux horizons et a été la source d'un grand enthousiasme.

Je suis également très reconnaissant envers Élise Huillery et Sandra McNally d'avoir accepté de participer à mon jury de thèse en tant que rapporteuses. Je tiens également à les remercier pour leurs commentaires avisés lors de la pré-soutenance. Je suis également très reconnaissant envers Manon Garrouste d'avoir accepté de participer au jury de thèse en qualité d'examinatrice. Je souhaiterais également remercier Andrew Clark pour les conseils avisés et bienveillants qu'il m'a prodigués tout au long de cette thèse.

Je voudrais remercier mes parents, Alain et Sophie, qui sont depuis toujours essentiels dans ma formation intellectuelle et mon parcours académique. Ils m'ont transmis le goût de la réflexion, le plaisir de la découverte et l'amour du savoir. Ils m'ont aussi transmis le goût de la discussion ouverte et raisonnée ainsi que la volonté de participer au bien commun et d'œuvrer à construire un monde plus juste. Enfin, ils ont été d'un immense soutien moral et affectif tout au long de cette thèse, sans lequel il aurait été difficile de réussir. Je leur suis éternellement reconnaissant pour cela, ainsi que pour tout le reste, et leur dédie cette thèse.

Enfin je voudrais remercier mes proches pour leur soutien. En particulier, je voudrais remercier mon épouse, Justine, qui a toujours été un moteur dans ma vie, et a plus que jamais joué ce rôle pendant ces cinq années de thèse. Je voudrais aussi la remercier de m'avoir toujours soutenu, ainsi que pour l'abnégation exceptionnelle dont elle a fait preuve au cours de cette dernière année. Je souhaite également remercier mes amis de toujours, Raph, Ben et Philippe pour leur amitié et leur soutien sans faille. Enfin, je souhaiterais également remercier mes amis plus récents et mes collègues du JPAL, notamment Clément, Paul, Fanny, Laura, Paul, Anthony, Simon, Ilf, Hélène, et Gaëlle, pour leurs conseils souvent avisés, leurs idées parfois saugrenues, leur grande solidarité et leur flexibilité.

Simon Briole
21 mai 2019
Paris

À mes parents, Alain et Sophie, et à mon épouse, Justine.

RÉSUMÉ

Initiée dans les années 1960 par l'économiste américain Gary Becker, le développement de la théorie du capital humain a considérablement ouvert le champ d'investigation de la science économique. Au cours des deux dernières décennies, de nombreux travaux de recherche en économie de l'éducation ont cherché à identifier les caractéristiques du système ou de l'environnement scolaire qui permettent aux individus d'acquérir un maximum de compétences, de savoirs et d'informations au cours de leur scolarité. Cette thèse s'inscrit dans ce courant de recherche et étudie plus spécifiquement deux aspects de l'environnement scolaire qui ont retenu l'attention des économistes : la productivité des enseignants et l'influence des camarades de classe.

Le premier chapitre de cette thèse étudie dans quelle mesure les pratiques pédagogiques des enseignants aux États-Unis permettent d'expliquer les différences de performances de leurs élèves en mathématiques. Dans un premier temps, nous montrons que chaque heure passée en classe à étudier les mathématiques engendre une progression significative des élèves dans cette discipline. Nous montrons ensuite que la productivité de l'heure d'enseignement est très fortement corrélée avec la mise en place de pratiques pédagogiques interactives, qui requièrent une participation active de la part des élèves. Plus spécifiquement, chaque heure passée avec un enseignant mettant l'accent sur ce type de pratiques est 2 à 3 fois plus productive qu'une heure passée avec un enseignant mettant l'accent sur des pratiques plus traditionnelles telles que le cours magistral.

Le deuxième chapitre de cette thèse étudie l'impact d'une politique publique visant à améliorer les pratiques des enseignants, à savoir le système d'inspection individuelle des enseignants du second degré en France. Dans ce chapitre, nous montrons que les performances des élèves en maths au Diplôme National du Brevet (DNB) s'améliorent significativement à la suite d'une inspection de leur enseignant de mathématiques, non seulement pour les élèves assignés à l'enseignant l'année de l'inspection, mais également pour les élèves assignés à cet enseignant les années suivantes, suggérant une amélioration durable de ses compétences pédagogiques. De surcroît, l'inspection des enseignants de maths de 3ème produit des effets bénéfiques persistants chez les élèves, qui se traduisent par une augmentation de leur probabilité de choisir une filière scientifique en première et d'obtenir un baccalauréat scientifique au cours des années suivantes. Finalement, les effets bénéfiques d'une inspection sur les performances des élèves en maths et sur leur trajectoire scolaire sont particulièrement marqués pour les enseignants de l'éducation prioritaire, lesquels font face à des contextes d'enseignement plus difficiles.

Le troisième et dernier chapitre de cette thèse étudie les effets du genre des camarades de classe de 3ème sur le parcours scolaire des élèves en France. Deux ensembles de résultats se dégagent de l'analyse. D'une part, l'influence des camarades de classes est persistante au cours du temps, puisque la proportion de filles parmi les camarades de classes en 3ème influence non

seulement la réussite au brevet, mais également le taux de décrochage scolaire, le choix des filières après le collège et le taux d'obtention du baccalauréat. D'autre part, la proportion de filles parmi les camarades de classes en 3ème a des effets bénéfiques sur la scolarité des filles alors qu'elle a des effets négatifs sur celle des garçons. Plus spécifiquement, cette proportion réduit le taux de décrochage scolaire des filles après la 3ème et augmente leur taux d'obtention d'un baccalauréat général, particulièrement dans la filière scientifique. A l'inverse, elle augmente la proportion de garçons choisissant une filière technique après le collège et réduit leur taux d'obtention d'un baccalauréat général.

Mots clés: *Économie de l'Éducation*, *Politique publique*, *Pratiques pédagogiques*, *Genre*, *Effets de pairs*.

SUMMARY

The Human Capital Theory developed by Gary Becker in the 60's substantially widened the area of investigation of economics. Over the last two decades, many studies in the economics of education intended to identify the characteristics of an educational system which enable individuals to acquire as much skills, knowledge and information as possible. This thesis contributes to this literature by studying two aspects of the educational environment that has particularly attracted economists' attention over recent years: teacher productivity and peer effects in the classroom.

The first chapter of this thesis investigates the extent to which teaching practices implemented by math teachers in the US relate to their students' math performance. First, it shows that every single hour spent in the classroom studying mathematics generates a significant improvement in students' math performance. Second, it shows that the productivity of instructional time strongly relates to the implementation of interactive teaching practices, which require student active participation in the lesson. More precisely, each hour spent with a teacher putting a high weight on this kind of practices is 2 to 3 times more productive than an hour spent with a teacher putting a higher weight on traditional practices, such a teacher lecture.

The second chapter of this thesis studies the impact of a public policy aimed at improving teachers' practices, namely the individual teacher evaluation system in French secondary education. In this chapter, we show that students' performance at the end-of-middle school national exam significantly improve after the evaluation of their math teacher, not only for students taught by an evaluated teacher the year of the evaluation, but also for students taught by the same teacher on subsequent years, suggesting a long-lasting improvement in teacher pedagogical skills. These positive effects persist over time for students, who not only perform better at the end-of-middle school exam but also choose more often and graduate more often from the science track in high school. In addition, the positive effects of teacher evaluation are particularly salient in education priority schools, in contexts where teaching is often very challenging.

The third chapter of this thesis investigates the effect of school peers' gender on students' performance and educational careers in French middle schools. First, it shows that the proportion of female peers' in middle school has persistent effects on students' educational careers as it not only affects students' test score at the end-of-9th-grade national examination, but also influences their track choices and high school graduation rates several years later. Second, it shows that a larger share of girls in the classroom has positive effects for girls and negative effects for boys. More specifically, it reduces girls' dropout rates and increases their probability to graduate from an academic track in high school, especially in the scientific track, while it increases boys' probability to attend a vocational school after 9th grade and decreases their high school graduation rate.

Keywords: *Economics of Education*, *Public policy*, *Teaching practices*, *Gender*, *Peer effects*.

Table of Contents

INTRODUCTION GÉNÉRALE

L'économie de l'éducation : émergence et tendances récentes

Initiée dans les années 60 par l'économiste américain Gary Becker, la théorie du capital humain a profondément marqué la science économique. Dans son article fondateur intitulé "Investment in Human Capital: A theoretical analysis" (1962), Becker développe un cadre conceptuel permettant de comprendre comment les « individus et les sociétés acquièrent des savoirs, des compétences et des informations en dépensant de l'argent et en dédiant du temps à leur scolarité, aux formations professionnelles ou à d'autres formes d'investissement » (Becker (2011)). L'investissement en capital humain peut alors être envisagé comme le fruit d'un choix rationnel, qui permet non seulement aux individus d'améliorer leur situation économique à travers l'acquisition de compétences valorisées sur le marché du travail, mais également à l'économie dans son ensemble d'être plus productive, à travers l'augmentation du niveau général des connaissances et des compétences. Dès lors, la science économique investit pleinement le champ de l'éducation, et cherche notamment à établir de manière empirique des liens entre le niveau moyen de capital humain dans un pays et le niveau de développement économique de ce pays d'une part, et entre le capital humain d'un individu et sa situation sur le marché du travail d'autre part. En particulier, les travaux fondateurs de Mincer (1970, 1974) mettent en lumière la relation positive qui existe entre le niveau d'éducation d'un individu, mesuré par le nombre d'années d'études qu'il a accompli, et son salaire. Néanmoins, le nombre d'années d'études se révèle rapidement être une mesure très limitée pour appréhender la notion de capital humain.

Au cours des trois dernières décennies, le développement de tests standardisés cherchant à mesurer directement les savoirs et les compétences cognitives des élèves de manière systématique, à l'échelle nationale ou internationale, a ouvert de nouvelles perspectives aux chercheurs en sciences sociales et a notamment impulsé de nombreux travaux en économie de l'éducation. Contrairement à la mesure du capital humain basée sur le nombre d'années d'études, qui suppose qu'une année d'étude a un rendement identique d'un individu à un autre et d'un système scolaire à un autre, ces tests fournissent des mesures plus fines des connaissances et des compétences individuelles et permettent d'introduire des différences de capital humain entre deux individus ayant le même niveau d'éducation. Dès lors, il devient possible d'étudier les déter-

minants du rendement d'un système éducatif, autrement dit, d'étudier les caractéristiques du système ou de l'environnement scolaire qui permettent aux individus d'acquérir un maximum de capital humain au cours de leur scolarité. Au cours des deux dernières décennies, deux aspects de l'environnement scolaire ont particulièrement retenu l'attention des économistes : la productivité des enseignants et l'influence des camarades de classe (*peer effects*). Les deux premiers chapitres de cette thèse portent sur la première thématique et le troisième chapitre porte sur la seconde thématique.

La productivité des enseignants

De récents travaux en économie de l'éducation ont permis de mettre en lumière les grandes différences qui existent entre enseignants dans leur capacité individuelle à faire progresser leurs élèves (Rockoff (2004); Rivkin et al. (2005)). Par contraste avec les effets plus modérés des ressources scolaires traditionnelles telles que la taille des classes ou le temps d'instruction, le fait de bénéficier de « bons enseignants » a une influence déterminante sur le parcours scolaire des élèves (Hanushek & Rivkin (2010)). Plus encore, une étude publiée en 2014 montre que ces effets persistent jusqu'à l'âge adulte, puisque le fait de bénéficier de bons enseignants améliore la situation individuelle sur le marché du travail plusieurs années après la fin des études (Chetty et al. (2014)). Dès lors, il est primordial de comprendre ce qui permet à un enseignant d'être productif, c'est-à-dire de faire progresser ses élèves. Très peu de travaux ont permis de dégager les déterminants de la productivité des enseignants à ce jour. En particulier, les études à ce sujet montrent un lien faible, voire inexistant, entre la productivité des enseignants et leurs caractéristiques personnelles, telles que le genre, le diplôme ou le niveau de certification (Hanushek & Woessmann (2011); Harris & Sass (2011)). Dans l'optique d'apporter un éclairage nouveau sur ces questions, le premier chapitre de cette thèse examine le lien entre les pratiques pédagogiques des enseignants et leur productivité et le second chapitre étudie l'impact du système d'inspection des enseignants sur leur productivité.

Le premier chapitre de cette thèse, intitulé « From teacher quality to teaching quality: instructional productivity and teaching practices in the US » étudie dans quelle mesure les pratiques pédagogiques des enseignants aux États-Unis permettent d'expliquer les différences de performances de leurs élèves en mathématiques. Pour éclairer ces questions, nous nous ap-

puyons sur les performances d'élèves de 4ème en math à l'issue du test TIMSS 2011. Dans le cadre de ce test, les élèves sont évalués dans quatre grands sujets des mathématiques : *Algèbre*, *Dénombrement*, *Géométrie* et *Probabilité*. Par ailleurs, la multiplicité des programmes scolaires aux États-Unis génèrent des différences importantes dans le temps d'instruction qui est dédiée chaque année à ces 4 grands sujets. Sur cette base, il est alors possible de déterminer dans quelle mesure une heure passée en classe à étudier l'algèbre plutôt que la géométrie génère une amélioration des performances relatives des élèves en algèbre, par rapport à leur performance en géométrie, tout en neutralisant leur niveau moyen en mathématiques ainsi que l'ensemble des facteurs de l'environnement scolaire qui influencent la réussite en mathématiques. De surcroît, pour chacun des 4 grands sujets, les élèves sont évalués dans une myriade de sous-sujets, dont certains n'ont pas été traités au cours de l'année. Les performances des élèves dans des sous-sujets non traités permettent de vérifier que les enseignants ou les directeurs d'écoles ne décident pas de l'allocation des heures de cours entre les différents sujets sur la base du niveau initial relatif des élèves dans chacun des sujets.

Le premier résultat de ce chapitre est que chaque heure passée en classe à étudier les mathématiques engendre une progression significative des élèves dans cette discipline. Plus spécifiquement, une heure passée en classe chaque semaine par un élève à étudier un des 4 grands sujets va générer un surcroît de performance relatif de 0.04 écart-type dans ce sujet[1]. Ce résultat étant calculé sur l'ensemble des élèves – et donc des enseignants - évalués aux États-Unis, il nous permet également de conclure que la productivité moyenne des enseignants de mathématiques est de 0.04 écart-type de test scores par heure de cours hebdomadaire.

La deuxième partie de ce premier chapitre consiste à examiner dans quelle mesure cette productivité horaire est reliée aux pratiques pédagogiques des enseignants. Pour ce faire, nous nous appuyons sur une enquête menée auprès des enseignants de mathématiques en charge des élèves évalués à la fin de l'année dans le cadre du test TIMSS. Sur la base des informations issues de cette enquête, nous montrons que la productivité des enseignants est très fortement corrélée avec la mise en place de pratiques pédagogiques interactives, qui requièrent une participation active de la part des élèves pendant le cours. Plus spécifiquement, chaque heure passée avec un enseignant mettant l'accent sur ce type de pratiques est 2 à 3 fois plus productive qu'une heure passée avec un enseignant mettant l'accent sur des pratiques plus traditionnelles

[1] À titre de comparaison, réduire la taille des classes de 5 élèves engendrerait une augmentation des résultats équivalente (Piketty et al. (2006)).

telles que le cours magistral ou la mémorisation de concepts et de méthodes de résolution. À notre connaissance, il s'agit du premier résultat empirique faisant état d'une relation forte entre la productivité horaire des enseignants et un ensemble de pratiques clairement identifiées et facilement reproductibles. Une exploration des mécanismes sous-jacents aux effets bénéfiques des pratiques interactives suggère que ces dernières permettent une amélioration de l'appétence des élèves pour les mathématiques et de leur confiance en leur capacité à réussir dans cette discipline.

Dans l'ensemble, les résultats obtenus dans le premier chapitre de cette thèse suggèrent que les grandes disparités de productivité observées entre les enseignants pourraient s'expliquer en grande partie par les différences d'efficacité de leurs approches pédagogiques, laissant place à des politiques publiques de formation des enseignants et d'incitation au renouvellement de leurs pratiques professionnelles. Dans cette optique, le deuxième chapitre de cette thèse, co-écrit avec Eric Maurin et s'intitulant « Does evaluating teachers make a difference? », étudie l'impact d'une politique publique visant à améliorer les pratiques des enseignants, à savoir le système d'inspection individuelle des enseignants du second degré en France.

En France, les enseignants du secondaire sont inspectés tous les 6 à 7 ans par des inspecteurs de l'Éducation nationale, qui sont des fonctionnaires expérimentés et hautement qualifiés. Ces inspections consistent essentiellement en une demi-journée d'observation de l'enseignant en classe et un entretien individuel approfondi avec l'enseignant évalué. Elles ont pour but de permettre aux enseignants de prendre du recul sur leurs pratiques pédagogiques, de leur proposer des formations professionnelles adaptées, et de s'assurer de la qualité de leur enseignement. De plus, ces inspections ont un impact direct sur la carrière des enseignants, qui peuvent bénéficier d'une promotion plus rapide suite à une inspection très favorable. Par ailleurs, tous les enseignants sont régulièrement inspectés au cours de leur carrière, indépendamment de leurs performances.

Dans ce chapitre, nous tirons parti du timing exact de l'inspection individuelle ainsi que de son caractère généralisé pour estimer l'effet d'une inspection sur la productivité d'un enseignant, mesurée par les performances de ses élèves à l'examen national de fin de collège, le *Diplôme National du Brevet* (DNB). Plus spécifiquement, nous comparons les performances des élèves assignés à un enseignant donné au cours des années précédant son inspection avec les performances des élèves assignés à cet enseignant les années suivant son inspection, au cours de

la période 2008-2012. Nous montrons que les performances des élèves en maths s'améliorent de 0.045 écart-type à la suite d'une inspection de leur enseignant de mathématiques, relativement à des élèves assignés à un enseignant de mathématiques non inspecté au cours de la période considérée. Pour s'assurer de la validité de notre analyse, nous montrons que, par contraste, les performances des élèves d'une classe donnée en français ou en histoire-géographie ne sont pas affectées par l'inspection de l'enseignant de mathématiques de cette classe. Plus généralement, nous montrons que les caractéristiques des élèves assignés à un enseignant avant et après inspection sont identiques en termes d'origine sociale, d'âge, ainsi que de proportion de filles et de latinistes dans la classe. De manière intéressante, nous montrons que les performances en maths au DNB s'améliorent non seulement pour les élèves assignés à l'enseignant l'année de l'inspection, mais également pour les élèves assignés à cet enseignant les années suivantes, suggérant une amélioration durable de ses compétences pédagogiques. De surcroît, l'inspection des enseignants de maths de 3ème produit des effets bénéfiques chez les élèves qui persistent dans le temps et se traduisent par une augmentation de leur probabilité de choisir une filière scientifique en première et d'obtenir un baccalauréat scientifique au cours des années suivantes. Finalement, les effets bénéfiques d'une inspection sur les performances des élèves en maths et sur leur trajectoire scolaire sont particulièrement marqués pour les enseignants de l'éducation prioritaire, lesquels font face à des contextes d'enseignement plus difficiles. Par contraste avec les inspections des enseignants de mathématiques, nous observons des effets bien moindres des inspections des enseignants de français sur les performances de leurs élèves en français au DNB.

De manière générale, les résultats obtenus dans ce chapitre apportent un éclairage nouveau sur la question de l'efficacité des politiques publiques visant à améliorer la productivité des enseignants. À notre connaissance, il s'agit de la première étude démontrant l'impact causal d'un système d'évaluation des enseignants sur leur productivité, à l'exception de Taylor & Tyler (2012), qui montrent que le système d'évaluation des enseignants de la ville de Cincinnati (US) génère également une amélioration des performances de leurs élèves en mathématiques.

L'influence des camarades de classe sur la réussite scolaire des élèves

L'idée que les camarades de classe influencent fortement la réussite scolaire des élèves a connu un essor considérable dans les années 60 à la suite d'un rapport concluant qu'il s'agissait d'un des principaux facteurs de réussite scolaire aux États-Unis (Coleman et al. (1966)). Par la suite, de nombreux travaux de recherche ont cherché à établir un lien causal entre la réussite d'un élève et les caractéristiques de ses « pairs », la difficulté méthodologique principale étant de parvenir à démêler les effets de contexte de l'influence réelle des pairs. Un ensemble de travaux récents a permis d'établir de manière convaincante un effet positif du niveau académique moyen des pairs sur les performances d'un élève (Sacerdote (2011)), mais ces travaux montrent également que, si la présence de bons élèves est très bénéfique pour d'autres bons élèves, elle a souvent des effets négatifs sur les élèves ayant un niveau académique plus faible (Hoxby & Weingarth (2005)). Par contraste avec l'impact du niveau académique des camarades de classe, l'impact du genre des camarades de classe a très peu fait l'objet de travaux en économie jusqu'à ces dernières années. Pourtant, de nombreux travaux en sciences sociales montrent que le genre des élèves façonne grandement leurs comportements et leurs interactions, particulièrement au cours de l'adolescence (Galambos (2004); Steinberg & Monahan (2007)).

Le troisième et dernier chapitre de cette thèse, intitulé « Are girls always good for boys? Short and long term effects of school peers' gender », étudie les effets du genre des camarades de classe de 3ème sur le parcours scolaire des élèves en France, au cours de la période 2008-2012. Pour ce faire, nous exploitons des variations dans la proportion de filles parmi les camarades de classe induites par des chocs démographiques dans le secteur scolaire du collège d'inscription. Le territoire français étant découpé en secteurs scolaires qui correspondent chacun à un collège public unique, chaque enfant d'une cohorte donnée en âge d'aller au collège doit s'inscrire dans le collège correspondant à son lieu de résidence[2]. Par conséquent, la proportion de filles parmi les élèves d'un collège donné est déterminée par la proportion de filles parmi les enfants d'une cohorte d'âge résidant dans le secteur scolaire correspondant à ce collège. Ainsi, des variations naturelles - induites par des chocs démographiques - dans la proportion de filles résidant dans un secteur scolaire donné vont se traduire par des variations naturelles dans la proportion de filles inscrites dans le collège de ce secteur d'une année sur l'autre. À leur tour, ces variations naturelles vont influencer la proportion de filles parmi les camarades de classe. Nous tirons

[2] À l'exception des familles inscrivant leurs enfants dans le système privé ou obtenant une dérogation.

parti du caractère exogène de ces variations pour estimer l'effet de la proportion de filles parmi les camarades de classe sur les performances au brevet (DNB) ainsi que sur le parcours scolaire ultérieur des élèves. De manière générale, deux ensembles de résultats se dégagent de l'analyse. Premièrement, l'influence des camarades de classe est persistante au cours du temps, puisque la proportion de filles parmi les camarades de classe en 3ème influence non seulement la réussite au brevet, mais également le taux de décrochage scolaire, le choix des filières après le collège et le taux d'obtention du baccalauréat. Deuxièmement, la proportion de filles parmi les camarades de classe en 3ème a des effets bénéfiques sur la scolarité des filles alors qu'elle a des effets négatifs sur celle des garçons. Plus spécifiquement, cette proportion réduit le taux de décrochage scolaire des filles après la 3ème et augmente leur taux d'obtention d'un baccalauréat général, particulièrement dans la filière scientifique. A l'inverse, elle augmente la proportion de garçons choisissant une filière technique après le collège et réduit leur taux d'obtention d'un baccalauréat général. Une exploration des mécanismes potentiels suggère que ces effets s'expliquent à la fois par une influence directe des camarades de classe sur le comportement des élèves au collège et par une influence indirecte de ces derniers, par le biais d'une adaptation du comportement des enseignants.

Les résultats obtenus dans ce chapitre contribuent à la récente littérature en économie démontrant l'impact causal du genre des camarades de classe sur la réussite scolaire. À notre connaissance, il s'agit de l'une des premières études démontrant un impact significatif et persistant sur le parcours scolaire des élèves, à l'exception de Black et al. (2013), qui montrent l'existence d'effets de long terme sur la situation individuelle sur le marché du travail à l'âge adulte en Norvège. Par ailleurs, ce chapitre constitue l'un des premiers travaux tentant de dégager les mécanismes permettant d'expliquer l'influence du genre des camarades de classe sur la réussite et le parcours scolaires.

Chapter 1

From Teacher Quality to Teaching Quality: Instructional Productivity and Teaching Practices in the US

Abstract

Though teachers are consistently found to play a major role in determining student achievement, little is known about what teachers can do to increase their instructional productivity. This paper develops a new empirical strategy, based on *within-student within math* variations in student test scores, to assess the instructional hourly productivity of math teachers in the US. Building on these estimates, we show that teachers' hourly productivity strongly relates to the use of teaching practices emphasizing student active participation in the lesson (*modern practices*). One weekly hour of math instructional time increases student test scores by 4.4% of a standard deviation on average, but one hour spent with a teacher above the modern practices index median is more than twice as productive as one hour spent with a teacher under this median (+5.9% vs +2.7% standard deviations). A further investigation suggests that the positive effects associated to modern practices are partially mediated by an improvement in student self-confidence and motivation to learn mathematics.

JEL classification: I20; I21; J24

Keywords: teacher quality ; teaching practices; instruction time; TIMSS; test scores; education.

Introduction

It is well established that teachers differ a lot in their individual capacity to raise student test scores (Rockoff (2004), Rivkin et al. (2005), Hanushek & Rivkin (2006, 2010)). Furthermore, being taught by a good teacher matters beyond schooling as it positively affects adult outcomes such as college attendance, earnings or fertility behaviours (Chetty et al. (2014)). Yet, very little is known about what makes a teacher effective in raising student achievement. Since the estimation of teacher value-added is demanding in terms of data and generally requires the use of administrative datasets, most of the works trying to identify the determinants of teacher effectiveness has focused on teacher demographics and other observed characteristics, such as certification or tenure. Nevertheless, the literature fails to establish consistent and powerful relationships between teacher productivity measures and teacher observed characteristics (Aaronson et al. (2007)), with the notable exception of teacher experience, which is systematically related to higher levels of productivity[1].

This paper investigates the role of a largely unexplored and yet intuitive input of teacher productivity, namely the teaching practices she implements in the classroom. Exploiting US 8th grade students' data from the TIMSS 2011 assessment, we show that practices emphasizing student active participation in the lesson positively and strongly relate to math teachers' instructional productivity.

The TIMSS assessment encompasses 4 basic math topics (*Number*, *Algebra*, *Geometry* and *Data & Chance*) and each topic is divided into 3 to 6 subtopics (19 subtopics in total). For each teacher, the dataset provides information on the amount of instructional time devoted to each topic the year before the assessment as well as information on which subtopics were taught during this pre-assessment period. This wealth of data makes it possible to develop a strategy for identifying teachers' hourly productivity by focusing first on the performance of students on subtopics that were *not* taught the year before assessment and, second, on their performance on subtopics that were taught over this period.

Accordingly, when we first focus on subtopics that were *not* taught during the pre-assessment period, we find *no* relationship between the amount of instructional time devoted by teachers to the corresponding topics and the performance of students. This result is consistent with the

[1] See Harris & Sass (2011) for a summary of recent findings on that topic

assumption that the amount of instructional time devoted by teachers to a given topic is not related to students' initial level of ability in this specific topic.

Building on this assumption, we then provide estimates of teachers' hourly productivity by focusing on the subtopics that have been taught during the pre-assessment period and by looking at the relationship between students' performance in these subtopics and the amount of instructional time devoted by teachers to the corresponding topics. This analysis reveals that students' test scores in these subtopics strongly relate to the amount of instructional time devoted by teachers to the corresponding topic. Specifically, we find that a one hour increase in weekly instructional time in a given topic is associated with an average increase of about 4.4% of a SD in students' test scores on the corresponding subtopics.

In a last step, we investigate the extent to which estimated teachers' productivity levels relate to the teaching practices implemented in the classroom. Specifically, we explore the relationship between our measures of teachers' productivity and their use of practices emphasizing student active participation in the lessons, as opposed to teacher-centered practices and to practices based on student memorization and basic problem solving. We explore these issues with the aid of a "Modern Practices" index (MPI) constructed from the TIMSS survey.

Generally speaking, we find large productivity differentials across US math teachers according to the teaching practice they implement in the classroom. The effect of one additional weekly hour of instructional time on students' scores varies from 2.7% of a SD for teachers under the median of the MPI to 5.9% of a SD for teachers above the median of the MPI. Put differently, using the continuous specification of the MPI, we find that a one SD increase in this index relates to a 8% SD increase in test scores, which is roughly equivalent to half the effect of a SD increase in teacher value-added estimates from previous studies (Hanushek & Rivkin (2010)). An investigation of the potential mechanisms at play suggests that the positive effects associated to modern practices are partially mediated by an improvement in student self-confidence and motivation to learn mathematics.

This paper contributes to the small literature that explores the role of teaching practices in shaping teachers' effectiveness. Some recent papers provide evidence that pedagogical skills and the quality of student-teacher interactions strongly relate to teacher productivity (Kane et al. (2011), Blazar (2015) and Araujo et al. (2016)). In parallel, Machin & McNally (2008) and Lavy (2009) argue that the positive effects on student achievement generated by the *Literacy Hour* in the UK and a teacher payment scheme in Israel, respectively, were primarily mediated

by changes in teaching methods. Altogether, these findings suggest that *teaching* - and not only *teachers* - may be a key determinant of instructional productivity, but they do not provide precise information on the teaching practices that are likely to improve teaching quality[2]. On the other hand, some recent papers have directly related *between subjects* or *between classes* variations in student test scores to variations in teaching practices across teachers, but they provide mixed results and do not give an insight into the magnitude of the relationship between teacher productivity and teaching practices[3]. The aim of this paper is to fill the gap between these two literatures by studying the relationship between the teaching practices implemented in the classroom by US math teachers and their instructional hourly productivity.

This paper also contributes to the literature which aims at evaluating the causal effect of instructional time on student math test scores. This effect is an economically meaningful one, as student math skills have recently proven to be important predictors of both aggregate economic growth (Hanushek & Woessmann (2008) ; Hanushek & Woessmann (2011)) and individual's future earnings (Rose & Betts (2004) ; Joensen & Nielsen (2009) ; Goodman (2017)). Yet, there is only scarce evidence on this topic. Several recent papers find a positive impact on student test scores, but most of them rely on small variations or exploit programs that are targeted at specific students and generally accompanied with other changes in school's input[4]. Two notable exceptions are Lavy (2015a) and Rivkin & Schiman (2015), who both exploit *within student between subjects* variations in instructional time across countries, and rely on the assumption that these variations are independent from student subject specific-skills. Building on their work, this paper intends to improve the identification of the causal effect of instructional time through the exploitation of variations across topics of a single subject. This strategy arguably both requires less restrictive identification assumptions and allows for the exploitation of large variations.

The remainder of the paper is organized as follows. The next section describes the data and the construction of the teaching practices index. The second section presents the empirical

[2]This notwithstanding, it is important to note that all the measures of teaching quality used by Kane et al. (2011), Blazar (2015) and Araujo et al. (2016) emphasize the importance of student-teacher interactions.

[3]These recent works include Aslam & Kingdon (2011), Schwerdt & Wuppermann (2011), Van Klaveren (2011), Bietenbeck (2014), Lavy (2015b) and Hidalgo-Cabrillana & Lopez-Mayan (2018).

[4]Recent papers on this topic exploit variations in the number of school days over the year due to bad weather conditions or legal differences in the school start date (Sims (2008), Marcotte (2007) and Marcotte & Hemelt (2008)), remediation programs (Taylor (2014), Cortes et al. (2015)), or policy changes that increased resources allocated to schools, which result in an increased amount of instruction time (Bellei (2009), Lavy (2012) and Fryer (2014)). Two recent papers exploit a recent reform that took place in Germany and which implied a modest increase (+5%) in instructional time (Andrietti (2015) and Huebener et al. (2017)).

strategy and provides some evidence on the validity of the identification assumptions. The third section presents the estimations of instructional productivity and its relationship with the teaching practices index. The final section concludes with a discussion of the implications of the main results.

1.1 The data

1.1.1 The TIMSS 2011 assessment

This paper exploits US data from the TIMSS 2011 assessment, which evaluates the math and science knowledge of eighth-grade students. The national sample is drawn from a two stage sampling procedure, whose objective is to ensure the national representativeness of US schools and students[5]. Every student in a selected class is assessed in math and science, and scores are assigned by independent external evaluators. This paper focuses on students' math test scores, which are important predictors of future earnings. The TIMSS math assessment encompasses 4 basic topics (*Number*, *Algebra*, *Geometry* and *Data and Chance*), each of which is divided into 3 to 6 subtopics (19 subtopics in total). Finally, it is possible to compute a specific test score for each of these subtopics.

Besides the student assessment, every math teacher who teaches a selected class is asked to answer a questionnaire, which provides information on teacher demographics and teaching practices. We restrict the sample to students whose math teacher answered the teacher questionnaire, which amounts to dropping 30% of observations. The final sample is made up with 7258 students, allocated over 387 classes in 359 schools, and taught by 376 different teachers. The available evidence suggests that students in the final sample performed slightly better over the year than students whose math teacher didn't answer the questionnaire, though there doesn't seem to be large differences in terms of school and student characteristics according to teacher non response to the questionnaire[6].

[5]First, schools are randomly selected among the national sample of schools. In a second step, one class is selected in each selected school.

[6]As we can see in tables A2 and A3 in the appendix, students in the final sample performed slightly better at the TIMSS assessment and are slightly older than those dropped from the initial sample due to teacher non response. No other difference appears to be significant between the two groups, regarding student and school characteristics.

1.1.2 Instructional time

The math teacher questionnaire includes detailed information about the total amount of instructional time that math teachers devote every week to each of the four basic topics in their class[7]. Importantly, students are taught these four topics in the same class, by the same teacher. As we can see in table A4, students are given 4.4 hours of instructional time per week in math on average. Half of this time is spent on *Algebra*, and the rest is distributed in a more balanced fashion over the three remaining topics, though a smaller amount of time is devoted to *Data and Chance* on average ($\simeq$ 0.45 hour/week). In addition, there are substantial variations across teachers, both in the total amount of math instructional time per week and in the allocation of this time over the four topics. As we argue in section 1.2, these observed variations in the share of instructional time devoted to the four topics might be mainly driven by the absence of a unique national curriculum in the US. Indeed, according to the TIMSS 2011 US National Research Coordinator, "the United States does not have a federally mandated national curriculum. State education agencies publish state mathematics standards and local school districts publish curriculum based on the standards"[8]. Such a variety of curricula introduces a lot of exogenous variations across schools and teachers in the allocation of instructional time across topics.

1.1.3 Teaching practices

The measures of teaching practices that we use in this paper are drawn from question 19 in the math teacher questionnaire. For each of the 11 teaching practices listed in the questionnaire (cf. table 1.1), teachers are asked the following question: "In teaching math to the students in this class, how often do you usually ask them to do the following?". There are four possible answers to this question: "Every or almost every lesson", "About half the lessons", "Some lessons" or "Never". Table 1.1 exhibits the distribution of teachers' answers to this question for the different practices.

Building on these questions, it is possible to construct for each practice and each teacher a measure of practice intensity, where intensity is set to 0 when the answer is "Never", to 1 when

[7] It is worth noting that the empirical strategy developed in this paper accounts for potential variations in the length of school year across schools that could introduce some measurement error in this measure of instructional time, as it is based on within student (and thus, within school) variations.

[8] Source: TIMSS Curriculum Questionnaire for Grade 8 (http://timssandpirls.bc.edu/timss2011/international-contextual-q.html)

Table 1.1 Definition and distribution of Teaching Practices

Teaching practice ("I ask students to...")	Never	Some lessons	Half lessons	Every lesson
(a) Listen to me explain how to solve problems	1 (%)	16 (%)	16 (%)	67 (%)
(b) Memorize rules, procedures, facts	4	41	32	23
(c) Work pbs (individually or with peers) with my guidance	0	7	18	75
(d) Work pbs in whole class with direct guidance from me	1	12	20	67
(e) Work pbs (individually or with peers) while I am occupied	26	37	10	27
(f) Apply facts, concepts and procedures to solve routine pbs	0	14	24	62
(g) Explain their answers	0	12	27	61
(h) Relate what they learn to their daily lives	3	34	38	25
(i) Decide on their own procedure for solving complex pbs	3	36	35	26
(j) Work pbs for which there's no obvious method of solution	13	52	25	10
(k) Take a written test or quiz	0	58	25	17

the answer is "Every or almost every lesson", 0.5 when the answer is "About half the lessons" and 0.1 when the answer is "Some lessons"[9]. To account for the fact that all teachers may not have the same definition of the different levels of intensity mentioned in the questionnaire (i.e., the same definition of "Every or almost every", for example) we also center these variable at teachers' means[10]. Overall, we obtain a set of variables describing the relative intensity of each practice for each teacher.

1.2 The evaluation of instructional productivity

1.2.1 Estimation strategy

Assessing the causal impact of instructional time on student achievement raises two identification issues. First, schools with more fundings can both attract better teachers and students and give the latter a higher amount of instructional time, which would introduce an upward bias in the estimation of instructional productivity. Second, students could be assigned a better teacher and more instructional time based on their previous math achievement. This would introduce an upward or a downward bias, depending on the direction of this within school sorting. To overcome these issues, we exploit *within student* variations in math instructional time, which occur across math topics that are taught by the same teacher, at the same school. Formally, we

[9] Alternatively, we assign the score 0.25 to the answer "Some lessons" and check the robustness of our results to this alternative score. Results are presented in the section dedicated to robustness.

[10] Investigating relationships among self-declared practices in our dataset, we find that all pairwise correlation coefficients between teaching practices are positive or null (cf. table A5), which tends to support the existence of an individual bias in the way teachers answered these questions in the TIMSS survey.

estimate the following model:

$$A_{it} = \alpha_i + c_t + \beta_1 IT_{it} + \epsilon_{it} \tag{1.1}$$

where A_{it} is the TIMSS score in math topic $t \in \{1; 4\}$ of student i, and IT_{it} is the quantity of instructional time devoted to topic t by student i's math teacher. The model also includes student fixed effects (α_i), which captures student innate ability and motivation to learn mathematics. Importantly, as all math topics are taught by the same teacher at the same school for a given student, student fixed effects also include teacher and school fixed effects. To complete the model, we add topic-specific constants (c_t). Standard errors are systematically clustered at the teacher level.

The only determinants of student achievement that this specification does not control for are student math topic-specific skills. As a consequence, under the assumption that the *within student between topics* variations in instructional time (IT_{it}) are not related to student topic-specific skills (ϵ_{it}), β_1 identifies the causal effect of a weekly hour of instruction time on student test scores and thus provides a valid estimate of teachers' average hourly productivity.

The US educational system is characterized by the absence of a unique mathematics curriculum for 8th grade students. Based on the "state standard" published by the state education agency, each school district defines its own curriculum. As there are more than 14,000 school districts in the US, this system induces a lot of variations in the allocation of math instructional time across topics that is arguably exogenous to teachers and students.The main threat to this assumption is the possibility that, within the curriculum constraint, teachers adopt strategic behaviours which would consists in marginally allocating a higher (or lower) share of their instruction time to the topic in which their students perform relatively better (or worse).

To test the existence of teacher strategic behaviours that would bias the estimates, we take advantage of a particular feature of the TIMSS assessment. For each math topic under consideration, students are evaluated in both subtopics that are taught over the year preceding the test and subtopics that are *not* taught over this period[11]. Consequently, the test provides us with measures of students' topic-specific skills that are unaffected by the amount of instructional time that is dedicated to study the related topics over the year. Indeed, the instructional time devoted

[11] As previously mentioned, students are evaluated in 3 to 6 subtopics per topic (19 subtopics in total). Subtopics that have not been taught the year preceding the test may have been taught over previous years or have never been taught to the students taking the test.

to a given topic the year of the test should positively affect student test scores in the related subtopics that are taught over the year, but not in the related subtopics that are *not* taught. Any relationship between the amount of instructional time devoted to a given topic and students' test scores in the related subtopics that are *not* taught over the year would instead capture teachers' strategic allocation of instructional time across math topics. Building on this argument, for all the estimations of instructional productivity presented in this paper, we implement regression (1.1) on the subtopics taught over the year only, and we show that there is no effect on subtopics *not* taught over the year.

1.2.2 US math teachers' instructional productivity

As we can see in the first column of table 1.2, when considering subtopics that are taught the year of the assessment, we find that one weekly hour of instructional time increases student math test scores by 4,4% of a standard deviation on average, which roughly amounts to a 3,3 points increase in the TIMSS test score[12]. Contrarily, the coefficient associated to Instruction Time is not significant when considering student test scores in the subtopics that are not taught the year of the assessment (cf. column (2)). As discussed in the previous section, this tends to support the main identification assumption. This effect is quite large, compared with the effect of other school's input. For example, doubling the total amount of math instructional time would increase student test scores by 19.3% of a standard deviation over the year, while a 10 students reduction in class size would raise student test scores by 10 to 30% of a standard deviation, as estimated from previous studies (Hanushek & Rivkin (2010)). In addition, this estimation is consistent with previous studies investigating the effect of instruction time on student test scores in comparable settings[13].

1.3 Instructional productivity and teaching practices

Building on the estimates of math teachers' instructional productivity computed from within student variations in math instructional time, the second step of the empirical strategy consists

[12]The mean test score in math in the final sample is 507.

[13]In particular, studies evaluating the effect of mathematics instructional time in the US provide estimates ranging from 2.5% to 5% of a standard deviation (Dobbie & Fryer Jr (2013), Taylor (2014) and Cortes et al. (2015)). Other studies including Bellei (2009), Lavy (2012), Lavy (2015a), Rivkin & Schiman (2015) and Andrietti (2015) find an effect ranging from 2.1% to 7% of a standard deviation.

Table 1.2 Math teachers' instructional productivity

	(1) Subtopics taught	(2) Subtopics not taught
Instruction Time	0.044***	-0.001
	(0.010)	(0.008)
Observations	18888	22263

Note: this table shows the effect of one weekly hour of instructional time on student math test scores, separately for subtopics taught the year of the test (column (1)) and subtopics *not* taught the year of the test (column (2)). All regressions include student and teacher fixed effects, as well as topic constants. Standards errors (in parentheses) are clustered at the teacher level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

in investigating the relationship between teachers' productivity and the teaching practices they implement in the classroom.

In order to describe the teaching style of a teacher in fewer dimensions than the 11 practices included in the questionnaire, we perform a Principal Component Analysis (PCA) at the teacher level. Based on this PCA, we create the *Modern Practices Index* (henceforth MPI), which equals teacher individual average score on practices (g), (h), (i), and (j). This index measures the relative importance of practices involving strong student-teacher interactions (practices (g) and (h)) and complex thinking (practices (i) and (j)) in the teaching style of the teacher, as opposed to teacher lecture ((a)) and basic problem-solving ((b), (c), (d), (e) and (f)), which are generally considered as traditional practices[14]. We complement this index with the frequency of assessment (practice (k)), which poorly relates to the MPI. Finally, we estimate the following model:

$$A_{it} = \alpha_i + c_t + \beta_1 IT_{it} + \beta_2 . IT_{it} . MPI_i + \beta_3 . IT_{it} . Assess_i + \epsilon_{it} \tag{1.2}$$

where MPI_i is the Modern Practices Index and $Assess_{ij}$ the frequency of assessment of student i's math teacher. The parameter β_2 indicates how teacher instructional productivity varies with the MPI. All other variables included in equation (1.2) are similar to those described in the previous paragraph for equation (1.1).

This strategy accounts for the potential endogeneity in the allocation of students to schools and teachers, as well as for the potential adaptation of teachers' teaching practices to the math general ability of the students in their class. Nevertheless, it is possible that the coefficient associated to the MPI reflects the effect of an unobserved teacher characteristics which is both related to the use of modern practices and to student achievement. To mitigate this concern,

[14] A detailed description of the results obtained from the PCA, as well as the construction of the MPI are available in appendix A.

we first show that the MPI is little influenced by the school and classroom environment, and that it is unrelated to teacher demographics (cf. table A6). By contrast, it is strongly and positively correlated with variables that relate to teachers' motivation and behavioral skills, such as collaborative behaviour and self-confidence[15]. Consequently, we sequentially add teacher characteristics as interacted controls in the regression:

$$A_{it} = \alpha_i + c_t + \beta_1 IT_{it} + \beta_2 . IT_{it} . MPI_i + \beta_3 . IT_{it} . Assess_i + \beta_4 . IT_{it} . X_i + \epsilon_{it} \tag{1.3}$$

where X_i is a vector of all teacher characteristics included in table A6, including teacher demographics and teacher behavioral controls. It also includes class size and the teacher perceived level of disruption in the classroom, which are two important determinants of instructional quality (Lazear (2001)). Results of the estimation of equation (1.3) are presented in table 1.3 and are discussed in the next section.

1.3.1 Math teachers' instructional productivity and Modern Practices

The use of practices emphasizing student active participation in the lesson is systematically associated to higher levels of teachers' instructional productivity. As we can see in table 1.3, the coefficient associated to the interaction term between instructional time and the MPI is positive and significant in all specifications. In addition, this coefficient is remarkably stable across specifications. In particular, the inclusion of teacher behavioural controls, which strongly correlate to the MPI, has a very little impact on the MPI's estimated coefficient. This tends to support the idea that the MPI captures the quality of teaching and not solely the effect of some confounding factors such as teacher motivation[16]. In addition to this, the frequency of assessment is posi-

[15] Due to the absence of *within teacher* variations in teaching practices in the dataset, it is difficult to completely rule out the possibility that the MPI includes the effect of some confounding factors. On the whole, though adding teacher controls in the regression alleviates such a concern, one should be cautious regarding a causal interpretation of the effect of modern practices.

[16] To check the consistency of our results, we further check that the use of modern practices is unrelated to the allocation of instructional time across math topics. To do so, we regress the MPI of a given teacher on the percentages of instructional time she devotes to the different topics, controlling for the math average score of the students she teaches. Results are reported in table A7 in the appendix. As we can see, none of the coefficients associated with the shares of instructional time devoted to the different topics is significant. In addition, we also compute the pairwise correlation coefficients between the MPI and the percentages of instructional time devoted to the topics. The only significant relationship that appears at the 10% level is a positive one between the share of instructional time devoted to *Geometry* and the MPI (cf. table A8). On the whole, there doesn't seem to be a strong relationship between the MPI and the allocation of instructional time across topics. Nevertheless, we check the robustness of our results to the exclusion of *Geometry* test scores (cf. table A17 in the appendix).

tively correlated to teachers' instructional productivity, though the corresponding coefficient is no longer significant when teacher behavioural controls are included in the regression.

To give insights about the magnitude of the variability in teacher productivity associated to the MPI, we provide two distinct interpretations. First, we examine how instructional productivity varies when moving along the MPI distribution. Assuming linearity in the effect of the MPI[17], we compute teacher instructional productivity at different points of the MPI distribution. Moving from the teacher at the 25th to the teacher at the 75th percentile of this distribution is equivalent to a 86% increase in instructional productivity (cf. table A9), which is substantial. Put differently, one hour of math instruction time spent with the latter teacher is about twice as productive as one hour spent with the former one. Second, we compute the effect of a standard deviation increase in the MPI on student test scores. A one standard deviation increase in the MPI increases student test scores by 0.018 of a standard deviation for each weekly hour of instruction time. Computing the effect for the whole year, a one standard deviation increase in the MPI increases student test scores by 0.08 of a standard deviation[18]. This is almost equivalent to doubling the total amount of instructional time, holding instructional productivity at its average level.

In a recent review, Hanushek & Rivkin (2010) show that the *teacher quality* literature provides estimates of the variability in teacher value-added that are highly consistent across studies. For mathematics teachers, a one standard deviation increase in teacher value-added is associated to a 0.15 standard deviation increase in student test scores over the year, on average. Similarly, moving from the teacher at the 25th percentile of the value-added distribution to the teacher at the 75th percentile during one single year is equivalent to a 0.2 standard deviation increase in math test scores. Using both interpretations, the MPI effect roughly equals half of the total teacher fixed effect (i.e. a one SD increase in the MPI equals half the effect of a SD increase in teacher value-added).

Furthermore, the magnitude of this effect is comparable to the results obtained by Kane et al. (2011) and Araujo et al. (2016), who use two distinct measures of pedagogical skills in order to assess the impact of teachers on US 3-8th grade and Ecuador 2-5th grade student achievement, respectively. In the first case, a one standard deviation in the TES score, which measures the

[17]We investigate the extent to which the relationship between instructional productivity and the MPI is linear in the robustness checks section.

[18]The effect of a SD increase in the MPI on hourly productivity is computed as follows: $\hat{\sigma}_{MPI} * \hat{\beta}_2 = 0.189 * 0.096 = 0.018$. Over the year, the effect is 4.4*0.018=0.079.

Table 1.3 Teaching Practices and Teachers' Instructional Productivity

	Score (1)	Score (2)	Score (3)	Score (4)	Score (5)
Panel A: subtopics taught (N=18888)					
Instructional Time (IT)	0.050***	0.061***	0.040	0.004	0.033
	(0.010)	(0.009)	(0.051)	(0.053)	(0.062)
IT*Modern Practices Index	0.096***	0.112***	0.108***	0.107***	0.099***
	(0.035)	(0.032)	(0.032)	(0.031)	(0.036)
IT*Assessment		0.050**	0.049**	0.042*	0.032
		(0.021)	(0.021)	(0.021)	(0.021)
Panel B: subtopics not taught (N=22263)					
Instructional Time (IT)	-0.001	-0.001	0.014	0.004	0.028
	(0.009)	(0.009)	(0.041)	(0.044)	(0.053)
IT*Modern Practices Index	0.005	0.005	0.012	0.017	0.006
	(0.024)	(0.024)	(0.024)	(0.025)	(0.025)
IT*Assessment		-0.001	-0.000	-0.001	-0.013
		(0.018)	(0.019)	(0.019)	(0.020)
IT*Teacher demographics	.	.	√	√	√
IT*Class size	.	.	.	√	√
IT*Teacher behaviour	.	.	.	.	√

Note: This table shows the heterogeneity in the effect of math instructional time on student math performance according to the teaching practices implemented in the classroom by the math teacher, separately on subtopics taught the year of the test (Panel A) and subtopics not taught the year of the test (Panel B). All regressions include student and teacher fixed effects, as well as topic constants and the proportion of subtopics taught the year of the test. Teacher demographic controls included in column (3) - (5) are teacher experience, gender and level of education and a dummy indicating if the teacher' major studied area was "Education-Mathematics". Controls included in column (5) include measures of teachers' collaboration with colleagues, self-confidence in teaching math and perceived level of disruption in the class drawn from the TIMSS teacher questionnaire and provided in the dataset, as well as a dummy indicating that the teacher participated in a professional development over the last two years. All controls are interacted with Instructional Time. Standards errors (in parentheses) are clustered at the teacher level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

quality of student-teacher relations and teacher global instructional skills through the assessment of an external evaluator, increases student test scores by 0.05 standard deviation. In the second case, a one standard deviation increase in the CLASS score, which measures the quality of teacher behaviours in terms of emotional support, classroom organization and instructional support through video observations, increases student test scores by 0.06-0.09 standard deviation over the year, depending on the specification. Importantly, both these measures put a high weight on the quality of student-teacher interactions, which is also the case for the MPI. This tends to confirm the idea that these interactions are crucial in shaping teachers' instructional productivity.

1.3.2 Robustness checks and heterogeneity analysis

The main results outlined in this paper are robust to several alternative specifications regarding the definition of teaching practice variables.

First, we provide evidence that the way we assign scores to the teaching practice variables in the main specification is not driving the results. As we can see in tables A10 and A11, respectively, considering a binary definition of teaching practice variables[19] or assigning the score 0.25 to the answer "Sometimes" leads to the same conclusion.

Second, we show that the main results are not driven by the correction applied to teaching practice variables, which objective is to take into account individual biases in teachers' answers. To explore this issue, we construct one Modern Practices Index and one Traditional Practices Index based on the non centered values of teaching practices, and we include both indexes in the regression[20]. As we can see in table A13, this specification gives similar results. Indeed, the "non centered" MPI is strongly and positively associated to instructional productivity while the coefficient associated to the "non centered" Traditional Practices Index is negative, though it's not significant at conventional confidence levels.

Third, the conclusions drawn from the main specification are robust to considering more dimensions of teaching practices than those captured by the MPI and the frequency of assessment. Indeed, including the teacher total score on all practices to take into account the diversity

[19] In the binary model, the score 1 is assigned to the answer "At every lesson" and 0 to the three other answers. The estimated coefficients from this regression are smaller and less significant than those obtained from the main regressions, as considering a binary definition of teaching practice variables amounts to lose a lot of information.

[20] In the main specification, the centered Modern Practices Index is strongly and negatively related to the Traditional Practices index (cf. table A12). By contrast, when computed over non centered values of teaching practice variables, these two indexes exhibit a small and positive correlation coefficient of 0.12. Consequently, both indexes are included in the regression.

of practices leaves the coefficient associated to the MPI roughly unchanged (cf. table A14). Furthermore, including the two traditional practices indexes described in section 1.1 instead of the single MPI also leads to the same conclusion, as the coefficients associated to both indexes are strongly significant and negative (cf. table A15)[21].

Fourth, the magnitude of the effect associated to the use of modern practices is unchanged when comparing the productivity of teachers who rank in the bottom half of the MPI distribution vs the top half, instead of using a continuous definition of the MPI. Indeed, teachers in the top half of the MPI distribution have an average productivity of 0.059 σ-test score per weekly hour, which is twice as large as the productivity of teachers who belong to the top bottom of this distribution (cf. table A16)[22].

In addition to these robustness checks regarding the specification of teaching practice variables, we also check that the results obtained from the main specification are not driven by the inclusion of *Geometry* test scores. As we can see in table A17, these results are robust to the exclusion of *Geometry* test scores from the regression, as it doesn't affect the coefficients associated to the MPI.

Finally, we investigate whether the MPI effect differs by student gender. Implementing equation (1.3) separately on girls and boys, we find no significant differences in the coefficient associated to the MPI (cf. table A18).

1.3.3 Potential mechanisms: Modern Practices and student non cognitive outcomes

This section investigates the extent to which the positive effect associated to the use of modern practices is mediated by an improvement of student non cognitive outcomes. Three measures of non cognitive outcomes are available in the dataset: student self-confidence in learning mathematics and student intrinsic and extrinsic motivation to learn mathematics[23]. As these outcomes are measured at the end of the year, they are plausibly affected by the teachers observed in the dataset and the teaching practices they have implemented in the classroom over the year.

As there is no within student variations in non cognitive outcomes in the dataset, we estimate the relationship between the MPI and student non cognitive outcomes through the following

[21]This specification better accounts for the second dimension of teaching practices highlighted in the principal component analysis.

[22]Unfortunately, the sample size is too small to precisely estimate teachers' instructional productivity at different points of the MPI distribution when the number of categories is higher than 2.

[23]These measures are drawn from questions 14 and 16 in the student questionnaire, and are directly provided in the database. A detailed description of the construction of these measures is available on the TIMSS website: https://timssandpirls.bc.edu/methods/t-context-q-scales.html

model:

$$NCO_{ij} = \alpha + \beta_1 MPI_{ij} + \beta_2 A_{i0} + \beta_3 X_{ij} + \epsilon_{ij} \tag{1.4}$$

where NCO_{ij} is the non cognitive outcome score of student i, taught by teacher j. MPI_{ij} is the Modern Practices Index of teacher j and A_{i0} is student i's math mean score, computed over subtopics *not* taught over the year, that are presumably unaffected by the teaching practices implemented by the observed teacher. This *proxy* for student math ability intends to control for the fact that initially better students, who also have better non cognitive outcomes, could be assigned teachers who rank higher on the MPI. Finally, X_{ij} is a vector of controls including student gender, age, socio-economic background and language spoken at home, as well as the amount of math instructional time per week, school size, indexes of school immediate area's economic affluence and urban density and all teacher characteristics included in equation (1.3).

As we can see in table A19, the use of modern practices is positively associated to the three non cognitive outcomes under consideration, though the relationship is not significant for student self-confidence at conventional levels. This result is consistent with the notion that the use of modern practices leads students to engage more actively in mathematics lesson. This attitude may, in turn, help them improve their math performance. Furthermore, this result is consistent with Algan et al. (2013), who find that teaching practices which imply strong student-teacher interactions and interactions among students are associated with higher levels of self-confidence and positive attitudes toward learning mathematics.

1.4 Conclusion

The results outlined in this paper shed a new light on the determinants of teacher instructional productivity and the mechanisms lying behind the large heterogeneity observed across US teachers. Building on a new empirical strategy to estimate teachers hourly productivity, we show that the use of practices emphasizing student active participation in the lesson is systematically associated with higher levels of productivity. Specifically, we construct an index measuring the relative weight that math teachers put on these practices and we show that teachers above this index's median are twice as productive as teachers under the median. In terms of magnitude, we find that a one SD increase in this index is related to a 0.08 SD increase in student test scores over the year, which is equivalent to half the effect of a SD increase in teacher value-added estimates from previous studies. A further investigation of the potential mecha-

nisms at play suggests that this effect is mediated by an increase in student self-confidence and motivation to learn mathematics.

These results confirm that teachers are a key determinant of student achievement and suggest a new way to improve teachers' productivity through the promotion of better teaching practices. An important area for future research is to determine the extent to which the positive relationship between teacher instructional productivity and practices based on student active participation truly reflects the causal effect of these practices. In particular, it is possible that only teachers endowed with a high level of pedagogical skills are able to efficiently implement these practices. In this case, forcing teachers (including those poorly endowed with pedagogical skills) to implement them could be counterproductive. In addition, it is important to take into account the adjustment costs incurred by a policy aiming at enhancing new practices, as teachers are not necessarily able to instantaneously absorb and retain new teaching methods. To our knowledge, the only paper dealing with these issues is the one by Haeck et al. (2014) who study the effect of a universal school reform implemented in the early 2000's in Quebec. Their findings are consistent with the existence of adjustment costs and therefore confirm that investigating the long term cost effectiveness of such policies in a dynamic and experimental setting is a key area for future research.

Chapter 2

Does evaluating teachers make a difference?

This thesis chapter is based on a joint work with Éric MAURIN.

Abstract

In France, secondary school teachers are evaluated every six or seven years by senior experts of the Ministry of education. These external evaluations mostly involve the supervision of one class session and a debriefing interview, but have nonetheless a direct impact on teachers' career advancement. In this paper, we show that these evaluations contribute to improving students' performance, especially in math. This effect is seen not only for students taught by teachers the year of their evaluations but also for students taught by the same teachers the subsequent years, suggesting that evaluations improve teachers' core pedagogical skills. These positive effects persist over time and are particularly salient in education priority schools, in contexts where teaching is often very challenging.

JEL classification: I20; I28; J24

Keywords: teacher quality; evaluation; feedback; teaching practices; supervision; education.

Introduction

There is a large body of research suggesting that teachers vary a lot in their ability to improve students' performance (Hanushek & Rivkin (2010)). It is also generally admitted that teacher evaluation can be a way to improve teachers' effectiveness, either by making it possible to provide them with useful feedbacks or by creating incentives to implement better practices (Isoré (2009); Taylor & Tyler (2012)). However, despite the recent evidence that existing evaluation systems produce accurate measures of teacher productivity (Jacob & Lefgren (2008); Kane et al. (2011); Bacher-Hicks et al. (2017)) there is still very little evidence on the actual impact of teacher evaluation on student performance. Teacher evaluations take many different forms across the world and vary a lot in terms of resources involved per teacher, but there is no consensus on what a good evaluation system should be and on how intensive it should be (Isoré (2009); OECD (2013a,b); Jackson et al. (2014)).

To shed light on this issue, this paper builds on administrative data with exhaustive information on the exact timing of secondary school teachers' evaluations in France, in a context where evaluations take place every six or seven years, involve very little resources per teacher and year, but have nonetheless a direct impact on teacher career advancement.

Evaluations are conducted by senior experts of the ministry of education, called *inspecteurs d'académie - inspecteurs pédagogiques régionaux* (hereafter *inspecteurs*), but each one of these *inspecteurs* is responsible for more than 350 teachers and has to perform on average about 40 evaluations per year, on top of many other managerial activities within the education system (IGEN (2011); IGEN/IGAENR (2016)). Evaluations mostly encompass the supervision of one class session and a debriefing interview with the teacher and we can estimate the cost to be about 600 euros per evaluation, namely about 100 euros per year and teacher. The results of these evaluations are used, however, to determine teachers' progression in the wage scale[1]. As a consequence, evaluations may not only help teachers improve their skills through the provision of evaluators' feedbacks, but they also give teachers strong incentives to provide effort to improve their teaching practices in order to be as good as possible on the day of evaluation.

[1]In most developed countries, teachers' evaluations are either conducted by internal evaluators only or not related to career advancement (OECD (2013a)). Only in a few countries (including Portugal, Switzerland and some regions of Germany) are teachers' evaluations conducted by external evaluators and have a direct impact on teachers' wage and promotion, as in the French system (OECD (2013b) ; Eurydice (2018)). Another important feature of the French system is that evaluations are conducted each year, in each subject, by the same group of highly qualified civil servants (*inspecteurs*) who likely develop a specific expertise in this task.

Each year, each teacher is assigned to a given set of classes and, consequently, teaches the same group of students over the whole year. Our empirical strategy exploits data on the exact timing of evaluations to compare the average performance of students assigned to a teacher *before* and *after* his/her evaluations, the basic question being whether evaluations coincide with specific improvement in students' average performance. Identification relies on the assumption that external evaluations do not coincide with teachers being assigned to better classes. Empirically, we checked that there is no specific change in students' characteristics before and after evaluations and, in particular, no changes in the proportion of students who have been held back a grade or in the proportion who take prestigious non-compulsory courses (such as Latin or ancient Greek courses). We also checked that external evaluations are not followed by specific changes in the level of teaching experience of colleagues who teach other subjects in the same class. If teachers were systematically assigned to better classes after external evaluations, we would observe a different pattern, namely a mechanical increase in colleagues' level of experience after external evaluations. Eventually, we provide evidence that the timing of teacher evaluations is unrelated to teacher mobility and that, more specifically, teachers don't move to better performing schools after an external evaluation. By contrast, as regards performance, we provide clear evidence that the visit of a math teacher by an external evaluator is followed by a significant increase (of about 4.5% of a SD) in students' scores in math at the end of middle school (9th grade). The effect of math teachers' evaluations is observed on performance in math, not in other subjects, consistent with the assumption that increased performance in math are driven by improved teaching practices of math teachers, not by an increase in students' overall academic ability or in math workload (which would likely be detrimental to performance in other subjects). Furthermore, math teachers' increased effectiveness is observed not only at the end of the evaluation year, but also at the end of the following years. Such persistent effects on teachers' effectiveness are consistent with the assumption that the visit of an evaluator is associated with an improvement in teachers' pedagogical skills, not just a temporary increase in teachers' effort. In the same spirit, the influence of math teachers' evaluations on their students can still be seen several years later, in high-school, as a larger proportion of their former students keep on studying math and succeed in graduating in fields of study which involve taking math exams. These longer term effects on students' outcomes are further suggestive that external evaluations do not simply help math teachers to "teach to the test", but make them able to improve students' core skills as well as students' perception of the discipline. These improve-

ments can be seen for less experienced teachers as well as for more experienced ones. They are even more significant for math teachers assigned to education priority schools, in context where students' academic level is often very weak and teaching more challenging.

Building on the same identification strategy, we show that external evaluations have smaller effects on French language teachers than on math teachers, even though evaluations of French language teachers are followed by significant improvement in French language test scores in education priority schools. The smaller effects of external evaluations on French language teachers are consistent with the existing literature on teacher effectiveness, which typically finds that teacher effects are much weaker on language exams than on math exams, maybe because students learn language in many settings outside schools, so that the influence of teachers is diluted and distorted by that of many other factors (Lavy (2009); Hanushek & Rivkin (2010); Harris & Sass (2011); Taylor & Tyler (2012); Wiswall (2013); Jackson et al. (2014); Papay & Kraft (2015)).

Eventually, when we consider the joint sample of math and French language teachers, we find an average effect of teacher evaluation of about 3% of a SD on test scores. Such an average effect is about the same order of magnitude as the average effect of a 5-student reduction in class size, as estimated by Piketty et al. (2006) for French middle schools. Our program of teacher evaluations involves, however, much smaller cost per teacher and year.

Our paper contributes to the growing literature on the causal impact of policies aimed at improving teachers' effectiveness. These policies include program of peer mentoring for new teachers (Rockoff (2008); Glazerman et al. (2008, 2010)) as well as programs of formal training and professional development (Angrist & Lavy (2001); Harris & Sass (2011)) and policies designed to evaluate and provide feedbacks to teachers (Weisberg et al. (2009); Allen et al. (2011); Taylor & Tyler (2012); Murphy et al. (2018)). Generally speaking, most existing papers focus on US programs and are suggestive that teacher-related programs can make a difference only insofar as they are high intensity. For example, the evaluation program in Cincinnati public schools involve the observation of four classroom sessions during the year of the evaluation, three by an external expert and one by an internal one (Taylor & Tyler (2012)). Both external and internal evaluators have to complete an intensive evaluation training program, so as to be able to measure several dozens of specific skills and practices. Overall, the Cincinnati program has a significant effect on math teacher effectiveness (about +10% of a SD on student' scores), but involves a total budget of about 7,500 dollars per evaluation, namely a cost per evaluation

that we estimate to be about 10 times more important than the budget involved by the program analyzed in our paper.

The remainder of the paper is organized as follows. Section 1 describes the teacher evaluation system as well as the organization of secondary schooling and national exams in France. Section 2 presents the databases exploited in this paper and the construction of our working samples. Section 3 develops our empirical approach and shows the effects of external evaluations on student outcomes through a graphical analysis. Section 4 implements a regression analysis to show the robustness of our main results and to explore the potential heterogeneity in the effects of evaluations. The final section concludes with a brief discussion on the implications of our results.

2.1 Institutional context

In France, secondary school teachers are recruited through national competitive exams organized each year, in each field of study, by the ministry of education[2]. Once recruited, teachers' progression through the wage scale depends not only on internal evaluations conducted each year by school heads, but also on external evaluations conducted every 6 or 7 years by senior experts of the ministry of education[3]. Internal evaluations focus on teachers' behavior at school (punctuality, absenteeism, participation in cross-class collaboration projects...) whereas external evaluations focus on teaching practices and pedagogical skills.

Teacher external evaluations

Teacher external evaluations are under the responsibility of a group of senior civil servants of the ministry of education, called *inspecteurs d'académie - inspecteurs pédagogiques régionaux* (hereafter *inspecteurs*). The vast majority of evaluations are conducted by *inspecteurs*

[2]The vast majority (93%) are granted the basic degree required to teach secondary school students, namely the *Certificat d'Aptitude au Professorat de l'Enseignement Secondaire* (hereafter CAPES). A small minority (about 7%) are recruited through an even more selective examination and hold an advanced degree, called the *Agrégation*. Most *Agrégation* recipients teach in high school or in higher education. In the remainder, given our focus on students' performance at end-of-middle school exams, we will focus on CAPES recipients.

[3] Teachers' basic promotion rate on the wage scale is based on their number of years of experience. But teachers who get good evaluations can be promoted at a faster rate. Going from the first to the last level of the wage scale takes about 30 years with the basic promotion rate versus only 20 years for the 30% teachers with the best evaluations. Teachers' access to the faster promotion track is determined by the weighted sum of the administrative grade that they get from school heads (/40) and the pedagogical grade that they get from external evaluators (/60).

themselves. A small fraction is conducted by senior teachers temporarily appointed to help *inspecteurs*[4].

Inspecteurs are recruited through national competitive exams restricted to experienced civil servants. There is one such competitive examination per field of study each year. Most candidates are experienced teachers who look for a career change. According to the staff directory of the ministry of education, *inspecteurs* are on average about 52 years old and have about 6 years of experience as *inspecteur* (see Table B4 in the appendix). Once recruited, each *inspecteur* is assigned to a specific education region by a centralized assignment system. There are 31 education regions in France and the average number of *inspecteurs* per region and field of study is typically very small compared to the number of teachers. For instance, according to the staff directory of the ministry, there are on average only about 5 math *inspecteurs* per region and they have to evaluate about 1,700 math teachers (Table B4)[5]. According to the same data source, about 250 math teachers are evaluated each year, in each region. Assuming that 85% of these evaluations are conducted by *inspecteurs*, it means that each *inspecteur* conducts on average about 40 evaluations per year.

Each evaluation involves the supervision of one class session. It also involves a debriefing interview with the evaluated teacher, during which the *inspecteur* provides feedbacks and advices. *Inspecteurs* can also provide teachers with suggestions about the specific training sessions that they could attend to improve their teaching practices or class management practices. On the day of the evaluation, *inspecteurs* also examine students' notebooks as well as the class book, namely the book where teachers have to report class sessions' contents, the exams that they give, etc. Eventually, *inspecteurs* have to produce a written report (so called, *rapport d'inspection*) where they provide an analysis of the class session that they supervised and provide explanations for the overall grade that they give to the evaluated teacher. In general, teachers are notified well in advance of the visit of the *inspecteur*, if only because the date of the visit has to coincide with a day when they teach. However, there is no legal constraint on notification delays.

Symbolically, the evaluation of teachers represents the most important task assigned to *inspecteurs*. But, in practice, *inspecteurs* are in charge of many other aspects of the education

[4]According to IGEN (2011), the proportion of external evaluations who are not conducted by *inspecteurs* vary across regions, but is never above 15%. Senior teachers appointed each year to help *inspecteurs* typically belong to the category who intend to take the exam to become *inspecteurs*.

[5]Overall, there were 142 math *inspecteurs* and 165 French language *inspecteurs* in France in 2008.

policy, so that the evaluation of teachers represents only a small part of their activities. As a matter of fact, *inspecteurs* are also in charge of the conception of the many national exams organized each year in France[6]. In each education region, *inspecteurs* also have to contribute to the conception and organization of teacher training and professional development programs. As regards human resources management, they are also expected to play a consulting role with teachers, namely they are expected to answer queries about both career advancement and teaching practices. More generally, *inspecteurs* are expected to supervise the actual enforcement of education policies in each education region and each school. Overall, according to surveys conducted by the ministry of education on the working condition of *inspecteurs*, the evaluation of teachers represents on average only between 20% and 30% of *inspecteurs*' activities (IGEN (2011); IGEN/IGAENR (2016)). Given that the total wage cost of an *inspecteur* is about 100,000 euros per year and assuming that about 20-30% of this cost compensates for evaluation tasks, we can estimate that 20,000-30,000 euros compensate for about 40 evaluations, meaning about 500-700 euros per evaluation[7]. Given that there is only one evaluation every six or seven year, the cost per teacher and year is about 100 euros.

School context and exams

In France, middle school runs from 6th to 9th grade and high school runs from 10th to 12th grade. Students complete 9th grade the year they turn 15. The curriculum is defined by the central government. It is the same in all middle schools and there is no streaming by ability[8]. The 20% most underprivileged middle-schools benefit from education priority programs which provide them with additional resources[9].

An important feature of the French system is that students stay in the same class, in all subjects, (with the same teacher in each subject), throughout the school year. Classes are groups

[6]Most notably, they are in charge of the different types of end-of high school *Baccalauréat*, as well as the different types of end-of-middle school *Brevet*, the different *Certificat d'Aptitudes Professionnelles*, etc.

[7]More information on the duties and compensations of *inspecteurs* can be found at the following address: http://www.education.gouv.fr/cid1138/inspecteur-de-l-education-nationale.html.

[8]9th grade students get about 25 hours of compulsory courses per week: 4 hours of French language, 3.5 hours of mathematics, 3.5 hours of History and Geography, 3 hours of Science, 1.5 hours of Technology, 5.5 hours of foreign languages, 3 hours of sport, 1 hour of art course. They also have the possibility to take additional (non compulsory) courses, such as Latin or ancient Greek. Principals can decide to assign students taking these additional courses to the same classes. Given that these students are typically good students, we may observe some segregation by ability across classes within schools.

[9]As shown in table B5 in appendix B, the proportion of students from low-income families is twice bigger in education priority schools than in non-priority schools. Education priority schools also exhibit higher proportions of repeaters and students in this type of schools get lower scores at the end-of-middle school national examination on average.

of about 25 students which represent, each year, very distinct entities. School principals assign students and teachers to classes before the beginning of the school year. In the remainder of this paper, we will mostly focus on teachers who teach 9th grade classes and our most basic measure of their effectiveness will be defined by the average performance of their students at the (externally set and marked) national exam taken at the end of 9th grade, which is also the end of middle school. This exam involves three written tests (in math, French language, history-geography) and our first question will be whether external evaluations of 9th grade teachers improve their ability to prepare their students for these tests. Specifically, we will mostly focus on math teachers and ask whether their external evaluations are followed by an improvement in the math scores of their students[10].

After 9th grade, students enter into high school, which runs from grade 10th to 12th grade. At the end of their first year of high school (10th grade), French students can either pursue general education or enter a technical or a vocational education program. Furthermore, those who pursue general education have to specialize in a specific field of study. There are three main fields: science (field "S"), economics and social sciences (field "ES") or languages and literature (field "L"). This is a key choice: each field of study corresponds to a specific curriculum, specific high school examinations, and specific opportunities after high school. Another important research question will be whether the effect of 9th grade teachers' evaluation on their students can still be seen one year later, at the end of 10th grade, on students' probability to choose S as field of specialization. The first year of high school (10th grade) is dedicated to exploring the different subjects and to choosing a field of specialization. The two last years of high school (11th and 12th grade) are dedicated to the preparation of the national high school exit exam, the *Baccalauréat*, which is a prerequisite for entry into post-secondary education. Students have to take one exam per subject, and they obtain their diploma if their weighted average mark across subjects is 10/20 or more, where subjects taken and weights depend strongly on their field of specialization. Given our focus on math teachers, a last research question will be whether the effect of 9th grade teachers' evaluation on their students can still be seen three years later, at the end of 12th grade, on students' ability to graduate in science (S).

[10]In the last section of this paper, we also present an analysis of the effects of external evaluations on French language teachers' effectiveness, as measured by their students' French language score. Generally speaking, we find much weaker effects on French language teachers than on math teachers, except in priority education schools.

2.2 Data and samples

In this paper, we use administrative data with detailed information on secondary school teachers for the period between t_0=2008-2009 and t_1=2011-2012. For each teacher j, this dataset gives information on whether (and when) j underwent an external evaluation between t_0 and t_1. It also gives information on whether (and when) teacher j taught 9th grade students and on the average performance of these students at exams taken at the end of 9th grade as well as at exams taken subsequently at the end of high school. Appendix B provides further information on how we build this database.

To construct our working sample of math teachers, we first extract from our main database the sample of math teachers who have less than 25 years of teaching experience, who taught 9th grade students in t_0, but who were not evaluated in t_0[11]. The size of this sample is about 40,000, which represents about 85% of the total number of 9th grade math teachers. About 57% of teachers in our sample are externally evaluated during the period under consideration and our objective is to evaluate the effect of these external evaluations on their students' math performance[12].

To explore this issue, we have to further focus on the subsample who teach 9th grade students at least one additional time after t_0, so as to be able to look at the evolution of students' performance at the end of 9th grade. The size of the corresponding working sample is about 30,000, which represents about 80% of the main sample. Most of our empirical analysis will be conducted on this working sample. One potential issue with this working sample, however, is that external evaluations may have an impact on teachers' probability to teach 9th grade students after t_0, meaning the selection into the working sample may be endogenous to the "treatment" under consideration. To test for such an endogenous selection, we considered the main sample of 40,000 observations and we tested whether the probability to teach 9th grade students on a year t after t_0 is different for teachers who are evaluated between t_0 and t and for those who are not evaluated in this time interval. As shown in Appendix Table B6, we find no

[11] We drop the small fraction of 9th grade teachers who are evaluated on year t_0=2008-2009 because the vast majority (about 96%) are not (re)evaluated before t_1 and cannot contribute to the identification of the effect of external evaluations. We also drop teachers with more than 25 years of teaching experience (on t_0) so as to minimize attrition rate. As it happens, many teachers with more than 25 years of experience are near the end of their working career and about 31% leave the education system between t_0 and t_1 (against only 4% for teachers with less than 25 years of experience). We checked, however, that results remain similar when we keep teachers with more than 25 years of teaching experience in our working sample (see appendix B14 and B15).

[12] The sample of French language teachers used in the last section of the paper will be constructed in a similar way.

significant difference between the two groups of teachers. The probability to teach 9th grade student on a given year after t_0 is on average about 78% for non-evaluated teachers and about 0.8 percentage point higher for evaluated teachers, the difference between the two groups being non-significant at standard level. The same diagnosis holds true when we replicate this sample selection analysis on subsamples defined by type of schools, teachers' experience or teachers' gender. Generally speaking, these results are consistent with the assumption that attrition is negligible.

Overall, our working sample includes 9,451 math teachers who teach 9th grade students at least two times between t_0 and t_1, which represents 30,414 observations in total. We provide some descriptive statistics in Appendix B (see column (1) of Table B7)).

2.3 The effect of evaluations: conceptual framework and graphical evidence

In the remainder of the paper, we ask whether teachers' external evaluations are followed by an improvement in their effectiveness, as measured by their ability to prepare 9th grade students for national exams or for high school. We first focus on math teachers and the last section provides results for French language teachers. The underlying educational production function is straightforward: (a) students' achievement is assumed to depend not only on their individual characteristics, but also on the effectiveness of their teachers and (b) the effectiveness of teachers is assumed to depend not simply on their level of experience, but also on the number of external evaluations they underwent since the beginning of their career. In this framework, assuming that teachers are assigned to the same type of classes on the years before and after the visit of an *inspecteur*, the comparison of the effectiveness of evaluated and non-evaluated teachers before and after an additional evaluation provides a means to identify the impact of such an additional evaluation on effectiveness. Before moving on to our econometric investigations, we start by providing simple graphical evidence on this issue.

The impact of external evaluations: graphical evidence

For each group of evaluated math teachers defined by the year t_e of their evaluation (with $t_0 < t_e \leq t_1$), let us consider Y_{ed} the average performance in math of their 9th grade students at national exams taken at the end of year $t_e + d$ and Y_{-ed} the average performance of the students

of non-evaluated teachers at the end of the same year $t_e + d$. Denoting Y_d and Y_{-d} the average of Y_{ed} and Y_{-ed} across all possible evaluation year t_e, Figure 1(a) shows the evolution of Y_d and Y_{-d} when d increases from d=-3 to d=+2 (i.e., the range of variation of d in our sample). The Figure reveals a marked increase in the average performance of students of evaluated teachers just after evaluations (i.e, for $d \geq 0$). The average performance of the evaluated and non-evaluated groups follows a similar pattern for exams taken before evaluations, but the gap widens for exams taken after evaluations.

To take one step further, Figure 1(b) plots the difference between evaluated and non-evaluated groups, with the last pre-evaluation year (i.e, t_e-1) being taken as a reference. It confirms that the evaluation year coincides with an improvement in the relative performance of evaluated teachers' students. The difference between the two groups of teachers is not statistically different from zero before the evaluation, but becomes statistically different from zero just after the evaluation.

Overall, Figures 1(a) and 1(b) are suggestive that evaluations have an impact on math teachers' effectiveness, as measured by the math scores of their 9th grade students. The basic identifying assumption is that evaluations do not coincide with teachers being assigned to better classes.

To further explore the credibility of our identifying assumption, Figures 2(a) and 2(b) replicate Figures 1(a) and 1(b) using average standardized scores in humanities as dependent variable, where scores in humanities are defined as the average of French language and history-geography scores[13]. Comfortingly, Figures 2(a) and 2(b) do not reveal any improvement in students' performance in humanities after external evaluations of math teachers. These Figures are in line with the assumption that external evaluations do not coincide with any overall improvement in the ability of students assigned to teachers. They are also consistent with the assumption that increased performance in math are driven by improved teaching practices of math teachers, not by an increase in math workload, since an increase in math workload would likely be detrimental to performance in other subjects.

A symmetrical falsification exercise consists in testing whether students' math performances are affected by the evaluation of non-math teachers. Figures 3(a) and 3(b) shows that this is not

[13] As mentioned above, students take three written tests at the end of 9th grade, namely a test in math, a test in French language and a test in history-geography. For each student, the score in humanities correspond to the average of the French language score and the history-geography score. Results are similar when we use separately the French language score and the history-geography score.

the case, namely the Figures do not show any improvement in student math performance after the evaluation of French language teachers, which further suggest that teachers are not assigned to intrinsically better classes after external evaluations.

In appendix B, Figures B5 (a) to B5 (c) provide additional evidence that external evaluations are not associated with teacher mobility (as captured by variation in teachers' seniority level) and do not coincide with teachers moving to better schools. In particular, these figures show that external evaluations do not coincide with any change in teachers' probability to teach in education priority schools. More generally, we do not see any variation in the academic level of the schools where they teach (as measured by the math average performance of 9th grade students at national exams taken in 2008, pre-treatment).

2.4 The effect of teachers' evaluations: regression analysis

The previous subsection provides us with simple graphical evidence on the effects of external evaluations on math teachers' effectiveness, as measured by the performance of their students at externally set and marked examinations. In this section, we explore the robustness of this finding - as well as the potential heterogeneity of effects across teachers and schools - using more parsimonious regression models. Specifically, we keep on focusing on the same working sample of math teachers as Figure 1(a) and we consider the following basic event-analysis model:

$$Y_{jt} = \beta T_{jt} + \theta X_{jt} + u_j + \gamma_t + \epsilon_{jt} \tag{2.1}$$

where Y_{jt} still represents the average standardized math score of teacher j's students at exams taken at the end of year t, while T_{jt} is a dummy indicating that an evaluation took place between t_0 and t. Variable X_{jt} represents a set of controls describing the average characteristics of the students taught by teacher j on year t (proportion of girls, average age, proportion studying ancient languages, etc.). X_{jt} also includes dummies controlling for teachers' number of year of teaching experience and for teachers' seniority level as well as a dummy indicating whether the teacher works in an education priority school and dummies indicating the education region. Eventually, the u_j and γ_t parameters represent a comprehensive set of teacher and year fixed effects while ϵ_{jt} represent unobserved determinants of students' performance.

In this set-up, parameter β can be interpreted as the effect of one additional external evaluation between t_0 and t on students' performance at the end of t. It should be emphasized

that this basic parameter encompasses the effect of evaluations which took place on t (the very year of the exam) and the effect of evaluations which took place between t_0 and $t-1$. To separate these two effects, we will also consider models with two basic independent variables, namely a dummy (denoted T_{1jt}) indicating that the evaluation took place on t and a dummy (T_{2jt}) indicating that the evaluation took place between t_0 and $t-1$.

To identify the parameters of interest in Equation (2.1), we assume that the timing of evaluations (as captured by changes in T_{jt}) is unrelated to changes in unobserved determinants of students' performance in math (as captured by changes in ϵ_{jt}), namely the same identifying assumption as in the previous graphical analysis. It amounts assuming that the evolution of the effectiveness of evaluated and non-evaluated teachers would have been the same across the period under consideration, had evaluated teachers not been evaluated. Table B8 in the appendix shows the results of regressing students' observed characteristics (gender, age, family background as well as the study of ancient languages or the study of German language) on T_{jt} using model (2.1). Consistent with our identifying assumption, the Table shows that the timing of external evaluation does not coincide with any significant variation in students' characteristics. We also checked that when we regress T_{jt} on all student observed characteristics, a F-test does not reject the joint nullity of the estimated coefficients[14]. These results hold true regardless of whether we use the full sample of math teachers or subsamples defined by level of experience, gender or type of schools. Eventually, Table B9 in the appendix confirms that the timing of evaluation does not coincide with teacher mobility (as captured by changes in teachers' seniority level) or with changes in the academic level of the schools where teachers work (as measured by school pre-treatment average scores or by priority education). The Table also reveals that the timing of evaluation does not coincide with changes in the level of experience or in the level of seniority of colleagues teaching other subjects to the same class. This finding is consistent with our assumption that evaluations are not followed by assignment to specific classes. If that were the case, evaluations would also mechanically coincide with assignment to classes with more senior and experienced colleagues.

2.4.1 Main effects on math scores

The first column of Table B1 shows the basic effect of external evaluations on math teachers' effectiveness, as measured by their students' performance in math at end-of-middle school na-

[14] Specifically, we have F(5, 20857) = 0.49 ; p-value = 0.78

tional exams. Consistent with our graphical analysis, it confirms that external evaluations are followed by a significant improvement in math score of about 4.5% of a SD. The second column shows the impact of external evaluations of math teachers on students' performance in humanities and, comfortingly, it shows no effect[15]. Column 3 shows the results of re-estimating the effect of math teachers' evaluations on math scores when we consider separately the effect on exams taken at the end of the evaluation year (T_{1jt}) and the effect on exams taken at the end of the following years (T_{2jt}). Both effects appear to be significant. The effect on exams taken at the end of the following years tend to be stronger (5.3% of a SD), but the difference between the two effects is non-significant at standard level. Eventually, column 4 confirms that math teachers' evaluations have no effect on performance in humanities, be they measured at the end of the evaluation year or later.

2.4.2 Heterogeneous effects

Table B2 shows the results of replicating our basic analysis separately on subsamples of math teachers defined by their gender, number of years of teaching experience (less than 11 years vs 11 years of more, where 11 is the median number of years of experience in our sample) or type of school (education priority schools vs regular schools). The Table shows that the impact of external evaluations on math scores is similar for men and women as well as for teachers with higher and lower level of work experience. By contrast, the impact appears to be significantly stronger for teachers in education priority schools (9.4% of a SD) than for teachers in non-priority schools (+3.1% of a SD). This finding is suggestive that external evaluations tend to be even more effective in school contexts where the average academic level of students is weaker and where teaching is more challenging[16].

Consistent with our identifying assumption, Table B2 also confirms that external evaluations of math teachers have no significant effect on students' performance in humanities, regardless of the subsample. As mentioned above, Tables B8 and B9 in the appendix provide balancing tests for the different subsamples which further confirm that external evaluations are not followed by any systematic variations in class composition, teacher mobility or colleagues' characteristics.

[15] As mentioned above, the score in humanities correspond to the average of the score in French language and the score in history-geography. We have checked that math teachers' evaluation have no effect on any of the two scores when we consider them separately.

[16] A survey conducted in 2006 provides an analysis of the specific challenges faced by teachers in education priority schools, due to students' social environment (poor working conditions at home, fatigue, diet…) as well as to students' disruptive behaviors and low academic ability. The survey report emphasizes that most teachers lack the pedagogical skills that are necessary to adapt teaching to this specific context (IGEN/IGAENR (2006)).

2.4.3 Longer term effects

Previous sections suggest that external evaluations improve the effectiveness of math teachers, as measured by their ability to prepare their 9th grade students for exams taken at the end of 9th grade. Table B3 shows that the influence of math teachers on their 9th grade students can still be seen one year later at the end of 10th grade (when students have to choose their major field of study) or even three years later, at the end of 12th grade, when they have to take their high school exit exams. Specifically, the Table focuses on the same sample of 9th grade math teachers as Tables B1 or B2 and looks at the probability that their students subsequently choose science as major field of study as well as at the probability that they subsequently succeed in graduating in science. The first column of the Table shows an increase in both probabilities. Specifically, it suggests an increase of about 0.5 percentage points in the probability to choose science at the end of 10th grade and to graduate in science at the end of 12th grade, which represent an increase of about 3% in this probability. Consistent with Table B2, the following columns shows that this increase is particularly significant for teachers in education priority schools (+10%). These longer term effects on students' choices and performance are suggestive that external evaluations do not simply help teachers to "teach to the test", but make them able to improve students' core skills as well as students' perception of the discipline.

2.4.4 Effects of external evaluations on French language teachers

Until now, we have focused on math teachers. In this section, we extend our analysis to French language teachers. The corresponding working sample is constructed along the same line as the working sample of math teachers, meaning we focus on those who teach 9th grade students on t_0, who are not evaluated on t_0 and who have less than 25 years of teaching experience on t_0. Figures 4(a) and 4(b) replicate Figures 1(a) and 1(b) using this working sample of French language teachers. In contrast with what we find for math teachers, these Figures do not show any significant variation in performance at French language exams after French language teachers' evaluations. Tables B10 and B11 in appendix B replicate Tables B1 and B2 using the sample of French Language teachers and confirm that external evaluations have only a small and marginally significant effect on their effectiveness, except when we focus on priority education schools (where the effect is about 7.6% of a SD). To further explore this issue, we looked at the effect of French language teachers' evaluations separately on reading test scores and writing test

scores[17]. This analysis shows that the effects of external evaluations tend to be slightly stronger on writing test scores, but the difference across writing and reading tests is not significant at standard level (see Table B12 in appendix B).

Generally speaking, the smaller effects of evaluations observed on French language teachers are in line with the literature on teachers' effects which recurrently finds that these effects are much weaker on language exams than on math exams (see e.g. Lavy (2009); Hanushek & Rivkin (2010); Harris & Sass (2011); Taylor & Tyler (2012); Wiswall (2013); Jackson et al. (2014); Papay & Kraft (2015)). One possible reason is that students learn language in many other settings outside schools, so that the influence of teachers is diluted and distorted by that of many other factors.

Eventually, Table B13 in the Appendix shows the results of replicating our main regression analysis on the joint sample of math and French language teachers, so as to estimate the average effect of teacher evaluations on end-of-middle-school exams. The Table shows a significant effect of about 3% of SD (8% of a SD in priority education). Not surprisingly, this effect is close to the average of the effect for math teachers and the effect for French language teachers estimated in previous sections. Building on the same type of database as those used in this paper, Piketty and Valdenaire (2006) found that a 5-student reduction in class size improves 9th grade students' average score in math and French language by about 4% of SD. Hence, our estimated effect of teacher evaluation is about the same order of magnitude as the effect of a 5-student reduction in class size. The corresponding cost, however, is much smaller[18].

2.5 Conclusion

Despite the general consensus that teachers represent an important determinant of student achievement, there is still little evidence on successful policies aimed at improving teacher effectiveness. In this paper, we study the impact of teacher evaluation on students' performance, in a context where evaluations are conducted every six or seven years by senior experts of the Ministry of Education and represent a key determinant of teacher career advancement. We show

[17]The French language end-of-middle-school exam consists of a set of reading and a set of writing exercises. During the exam, students are given the same amount of time to complete each one of the two sets of exercises.

[18]Given that class size is about 25 students on average, a 5-student reduction corresponds to a class size reduction of about 20%. Hence, the corresponding cost per teacher and year can be estimated to be about 0.20 x 50,000 euros where 50,000 euros is a proxy for the total labor cost of a secondary school teacher. We end up with a cost per teacher and year of about 10,000 euros whereas the cost per teacher and year of the evaluation system is only about 100 euros (as discussed in section 2).

that math teachers' evaluations increase their students' performance in math at end-of-middle school national exams. This effect is seen not only for students taught by the teacher the year of the evaluation but also for students taught by the same teacher the subsequent years, suggesting that evaluations improve teachers' core pedagogical skills. Math teachers' evaluations also generate persistent benefits for their students, who not only perform better at the end-of-middle school exam, but also graduate more often in science at the end of high school, three years later. The impact of evaluation appears to be much smaller for French language teacher, except in education priority schools. For both math and French language teachers, the positive effects of evaluations are actually particularly salient in education priority schools, in contexts where teaching is often very challenging.

In terms of policy implications, our results suggest that a low-intensity low-cost evaluation program can be highly cost effective provided that it is conducted by external authorities and has a significant impact on teachers' career advancement. Our results also show that evaluations can generate significant benefits even after ten years of work experience. In most countries, evaluations tend to be concentrated on beginning teachers, whereas our findings suggest that it can be efficient to evaluate teachers all along their career, not simply at the start. Finally, our findings show that evaluations are particularly worthwhile in contexts where teaching is very challenging, such as education priority schools. Reinforcing teacher evaluations in this type of schools thus appears as an appealing way to reduce educational inequalities.

Chapter 3

Are girls always good for boys? Short and long term effects of school peers' gender

Abstract

This paper exploits idiosyncratic variations in school cohorts' gender composition to investigate the short and long-term effects of school peers' gender. Using French administrative data over the 2008-2012 period, it shows that the proportion of female peers' in middle school not only affects students' contemporaneous performance but also influences their subsequent educational attainment. More specifically, a larger share of girls among school peers increases girls' test scores, reduces their dropout rates and increases their probability to graduate from high school several years later, especially in the scientific track. By contrast, it increases boys' probability to attend a vocational school and decreases their high school graduation rate. We find suggestive evidence that these effects partially operate through a negative effect of opposite-gender peers on students' classroom behaviour and relationships with their teachers.

JEL classification: I20 ; I24; J16

Keywords: peer effects; gender; student performance; dropout; schooling choices; social interactions.

Introduction

Social scientists have long pointed out that peers' influence is a key determinant of student achievement, and recent empirical studies consistently find significant effects of school peers' ability on student own achievement[1]. Nevertheless, until recent years, the effects of peers' gender has only received little attention by economic researchers. Yet, student gender is a crucial determinant of social interactions at school, especially among 14-15 years old students[2]. In turn, these interactions may not only affect students' behaviour and performance at school but they may also shape students' preferences in a persistent manner. From a theoretical point of view, the effects of being exposed to a higher proportion of female peers at school are unclear. On the one hand, a higher proportion of girls should benefit to all students through a positive effect on the quality of the learning environment, as girls tend to be more conscientious and disciplined than boys on average[3]. On the other hand, a higher proportion of opposite-gender peers may provide greater distraction (Coleman (1961); Hill (2015)) and/or may reduce cooperation among students (Lu & Anderson (2014)), resulting in a negative effect of girls on boys' achievement and a positive effect girls' one. As a consequence, it is key to better understand how school peers' gender affects girls' and boys' achievement and educational choices.

This paper builds on administrative data to investigate the effects of the proportion of girls among school peers in middle school on student achievement and long-term educational choices in France. To address causality issues, we take advantage of two key features of the French educational system, namely compulsory schooling up to age 16 and the existence of a catchment area system implying very little school choice in the public education system. Due to this institutional setting, demographic shocks in the gender ratio of cohorts living in a school catchment area generate idiosyncratic variations in the gender ratio of school cohorts. Building on this setting, our empirical strategy exploits variations in the proportion of girls which occur across cohorts of the same middle school.

[1] See Sacerdote (2011) for a review of the economic literature on peer effects. Another recent paper by Monso et al. (2019) summarizes the literature on peer effects in primary and secondary education.

[2] The high prevalence of gender stereotypes and the cost associated to deviation to these stereotypes during adolescence are well documented by developmental psychologists, who consider puberty as a "gender intensification" period (Galambos (2004), Lobel et al. (2004)). In particular, Steinberg & Monahan (2007) find that resistance to peer influence is at its lowest at age 14.

[3] See for example Duckworth & Seligman (2006), Jacob (2002) Bertrand & Pan (2013) or Cornwell et al. (2013).

We first show that *within school cross cohort* variations in the proportion of girls are unrelated to other changes in peers' socio-economic and educational backgrounds, as measured by the proportion of students who have been held back a grade. We further show that variations in the proportion of girls are unrelated to changes in class size or teacher quality, as measured by teacher experience or tenure. By contrast, we provide clear evidence that the proportion of girls among school peers not only affects student contemporaneous achievement, but also affect their subsequent educational attainment. Generally speaking, being surrounded by more female peers benefits to female students while it has detrimental effects on male students, both in the short and the long run. A 20 pp increase in the proportion of girls in the classroom in 9th grade (i.e, 5 students) increases girls' average performance at the end-of-middle school national exam by 2% of a SD, and decreases boys' one by 1.5% of a SD. It also increases girls' probability to attend high school (+1.6%) and to graduate from a high school academic track (+2%) (especially from the scientific track (+3%)), and decreases their probability to attend a vocational school (-2%) and their dropout rate (-3.8%). By contrast, it decreases boys' probability to attend high school (-1.5%) and to graduate from a high school academic track (-2%), and increases their probability to attend a vocational school (+2.5%)..

Our results are consistent with the notion that a greater proportion of opposite-gender students among school peers decreases cooperation in the classroom (Lu & Anderson (2014)) and provides greater distraction to students, resulting in poorer classroom behaviour and student-teacher relationships (Hill (2015)). We exploit teachers' grading practices to test for the second type of mechanism and we find suggestive evidence that the positive (negative) effects of the proportion of girls among school peers on girls' (boys') achievement are partially mediated by an improvement in girls' classroom behaviour and relationships with their teachers relative to boys.

This paper contributes to the growing literature on the causal effects of school peers' gender on students' achievement and educational choices. Generally speaking, this literature focuses on short-term outcomes and consistently finds that a higher proportion of girls benefits to all students. In particular, Hoxby (2000), Whitmore (2005), Lavy & Schlosser (2011) and Hu (2015) find that a higher proportion of girls among school peers increases both girls' and boys' contemporaneous test scores and provide suggestive evidence that these effects operate through an improved learning environment. Related work further show that the proportion of girls in middle school increases both girls' and boys' probability to chose STEM over language sub-

jects in high school (Schøne et al. (2019)) and reduces girls' and boys' dropout rates (Anil et al. (2016))[4]. One notable exception is Black et al. (2013), who exploit Norwegian registry data and find that the proportion of girls in middle school positively affects girls' employment rate and wages several years later but has slightly negative effects on boys' educational attainment[5]. Eventually, a related literature investigates the effects of school peers' gender in the specific context of single-sex education. Generally speaking, this literature finds positive effects of attending a single-sex schools for both boys a girls (Jackson (2012, 2016); Park et al. (2013, 2018); Dustmann et al. (2018)) but the evidence is more mixed as regards the effects of being assigned to single-sex classes in mixed-gender schools (Strain (2013); Lee et al. (2014); Eisenkopf et al. (2015); Booth et al. (2018)).

We add to the literature by providing clear evidence on both short and long term effects of school cohort gender composition on student achievement and educational attainment, in the broad context of mixed-gender education. Our results are suggestive of persistent effects of school peers' gender on students' skills and preferences, which is a key result to understand the long-term consequences of gender imbalances at school, especially in the context of considerable gender imbalances across fields of study observed in developed countries (OECD (2017)). In addition, this paper is one of the first papers providing clear evidence of negative effects of female peers on boys' achievement and educational attainment in a context of mixed-gender education, with the notable exceptions of Black et al. (2013) and Hill (2015).

The remainder of the paper is organized as follows. Section 1 describes the French institutional context. Section 2 presents the datasets exploited in this paper and the outcomes under consideration. Section 3 presents our empirical strategy. Section 4 provides evidence on the validity of the identifying assumption and presents the estimations of school peers' gender effects on student outcomes. Section 5 discusses potential mechanisms driving these effects. The final section concludes with a discussion of the implications of the main results.

[4]A related work by Schneeweis & Zweimüller (2012) shows that a higher proportion of girls among lower secondary school peers increases girls' probability to chose a male-dominated vocational school for the specific population of girls choosing a vocational school at age 14. In addition to this, Oosterbeek & Van Ewijk (2014) and Hill (2017) find that the impact of female peers is much less pronounced in tertiary education than in secondary education.

[5]Another exception is Anelli & Peri (2017), who study the impact of high school peers' gender on college major choice and labour market outcomes in Italy. They find no significant effect of female peers, but their results are limited to very good students (i.e, those graduating from college-preparatory high schools).

3.1 The French educational system

3.1.1 School system, examinations and track choices

Secondary education in France consists in four years of lower secondary education in middle school (from 6th to 9th grade), and three years of upper secondary education in high school (from 10th to 12th grade). Schooling is compulsory up to age 16 in France. This paper focuses on the population of 9th grade students (14-15 years old) enrolled in public middle schools. An important feature of French middle schools is that students stay in the same class with the same teachers throughout the school year. Classes are groups of about 25-30 students which represent, each year, very distinct entities[6]. As a consequence, classmates represent the group of peers with whom students interact the most over the school day.

At the end of 9th grade, students take a national examination called *Diplôme National du Brevet* (hereafter DNB), which is externally set and marked. Students are assessed in three topics: mathematics, French language and history-geography. As early as the following year, students who continue to high school have to decide to take either the academic track or one of the vocational tracks. The academic track is a 3-year high school education preparing for college. In contrast, vocational tracks are generally much shorter and, in most cases, do not enable students to access tertiary education. Students enrolled in the academic track in 10th grade have to specialize in one of the three academic tracks, namely *Science*, *Economics and Social sciences* and *Literature*, or to switch to a vocational (technological) track when they move on to 11th grade. To complete high school, students take a national examination at the end of 12th grade, namely the *Baccalauréat* (hereafter BAC), which is specific to each track. Students who complete one of the three academic tracks are then eligible for college enrollment, but each of these tracks give them access to very distinct opportunities in post-secondary education.

3.1.2 The catchment area system

The French secondary education system is characterized by a catchment area system. Every child is assigned a single public school, depending on her place of residence. That is, every postal address belongs to a single school area (“secteur scolaire"), which corresponds to a single public school. Middle school areas are determined at the regional level, by the local

[6]School principals assign students and teachers to classes before the beginning of the school year.

government[7]. There are two ways to relax the constraint imposed by the catchment area system. The first one is to request and obtain a dispensation. Dispensations are granted by the regional education authority based on the following criteria: special needs students, medical reasons, need-based and merit-based grants, siblings enrolled in the same school, distance to school and special academic tracks. Dispensations are only granted when all places were not fulfilled by children living in the catchment area. In total, only a small amount of students obtain the dispensation[8]. The second possibility to relax the constraint imposed by the catchment area system is to enroll in the private sector, for which there is no legal constraint regarding school choice. Most private schools in France are publicly-funded and must follow the same curriculum as public schools (except for religious instruction). In France, around 20% of students are enrolled in private middle schools, and there is no strong gender selection between the public and the private sectors[9]. Given the absence of legal constraint regarding school choice for students enrolled in the private sector, this paper focuses on 9th grade students enrolled in public middle schools.

3.2 The data

3.2.1 Datasets and sample restrictions

We exploit a comprehensive panel dataset of students enrolled in secondary education in France over the 2007-2012 period[10]. This dataset provides detailed information on student enrollment status every year, basic student demographics (age, gender, nationality and financial aid status) as well as the unique identifier of the class in which the student is enrolled and student performance at national examinations at the end of 9th and 12th grade. In this study, we focus on 9th grade students enrolled in public middle schools, which are subject to the catchment area system. We then match the student dataset with a teacher dataset[11], which is available for the 2008-2011 period. For each academic year, this second dataset provides information on the background characteristics of all teachers from public secondary schools (gender, level of

[7]Namely, the “Conseil Général".

[8]For example, 96% of 6th grade students in a public middle school in 2006 studied enrolled in their catchment area's middle school (cf. graph A2). The system was slightly softened in 2007, due to a reform which increased the possibilities for dispensation. As a results, 6 to 8% of students entering a public middle school over the 2007-2009 period obtained a dispensation.

[9]Over the 2007-2011 period, 49% of students enrolled in private middle schools are female students, as compared to 50% in public schools.

[10]Namely, the *Fichier Anonymisé d'Elèves pour la Recherche et les Etudes*.

[11]Namely, the *Annuaire du Personnel du Secondaire Public*.

experience, seniority, certification level, employment status), as well as the unique identifier of classes in which they teach. Consequently, we match every 9th grade class with their math, French language and history teachers.

Overall, the final sample includes 4 cohorts of students from 82,184 classes allocated over 5,240 public middle schools. Table C1 in the appendix presents some summary statistics regarding student demographics in the sample, by student gender, averaged at the class level. Table C2 presents sample means and standard deviations of classroom and teacher characteristics. In addition, Figure A1 in the appendix plots the distribution of the proportion of girls in our final sample, at the school cohort level.

3.2.2 Outcomes

This study focuses on both short-term and long-term educational outcomes. The main outcomes under consideration are measured as follows:

1. **Student test scores and school behaviour**: every student in the French educational system must take a national examination at the end of middle school (9th grade), which is externally set and marked. Students are evaluated in three topics: mathematics, French language and history-geography. Student average test score at this exam represents our main measure of student performance[12]. In addition to this, we also exploit a measure of student behaviour at school, namely the "*Note de vie scolaire*". Every student is rated by the school principal three times a year, on the basis of three criteria: (1) assiduity and punctuality (2) compliance to school internal rules and (3) participation to school life. We consider the student average grade over the year as the main measure of student behaviour.

2. **Student track choices and educational attainment**: at the end of 9th grade, students choose whether to go to high school (general track), to go to a vocational school or to dropout. This is a key choice, which leads to very distinct labour market outcomes several years later. In particular, dropping out after 9th grade is associated with very poor outcomes, while vocational degrees are associated with much higher employment rates and better wages but generally don't allow students to enroll in higher education[13]. Even-

[12] Test scores are standardized at the topic x regional x year level.

[13] Among individuals who completed their education in 2013, 60% of those who dropped out after 9th grade are unemployed 3 years later, against 25% for the rest of the population. In addition, 69% of individuals who

tually, the academic track is a 3-year education preparing for higher education[14], which leads to better labour market outcomes than vocational degrees. Hence, we consider 3 outcomes: (1) going to high-school (academic track) (2) going to a vocational school and (3) dropping out after middle school. For students who choose the academic track in high school in 10th grade, we observe whether they graduated from one of the three academic tracks or from the technological track 3, 4 or 5 years later. We define 5 outcomes related to high school graduation within the 5 years following 9th grade: (1) graduation from one of the three academic tracks (2) graduation from the *Science* track (3) graduation from the *Economics/Social Sciences* track, (4) graduation from the *Literature* track and (5) graduation from the *Technological* track.

Sample means and standard deviations of all the outcomes considered in this study are presented in table C3 in the appendix, by student gender, averaged at the class level. On average, girls tend to outperform boys at the end-of-middle school examination and to be better behaved. They choose more often the academic track in high school (61%) than boys (54%), and they graduate more often from one of the academic tracks (38% vs 30%), especially in *Economics/Social Sciences* (13% vs 8%) and in *Literature* (9% vs 2%). On the other hand, boys go more often to a vocational school (36%) than girls (30%) and have slightly higher graduation rates in the *Science* academic track (19% vs 16%).

3.3 Empirical strategy

Estimating the causal effect of school peers' gender on student achievement and educational career raises an identification challenge, due to both between and within school student sorting. For instance, schools with particularly high levels of violence may also be schools in which the proportion of female students is low, if parents decide to move their child away from a violent school differentially for boys and girls. In addition to this, principals may sort students across classes according to both gender and ability, resulting in classes with higher proportions of both girls and high achieving (or low achieving) students.

completed a vocational degree are in employment 3 years later, with a median net monthly wage of 1300 euros, against 83% and 1800 euros for individuals with any higher education degree (cf. Gaubert et al. (2017)).

[14] 99% of students graduating from an academic track in high school in 2014 subsequently entered higher education, against 33% of students graduating from a vocational degree (cf. Kabla-Langlois (2016)).

To tackle this endogeneity issue, this paper takes advantage of two key features of the French institutional context. Due to the coexistence of compulsory schooling up to age 16 and of the school catchment area system, natural fluctuations in the gender ratio across birth cohorts living in a given school catchment area translate into variations across cohorts of students enrolled in this area's public middle school[15]. Building on this, we exploit *within school cross cohort* variations in the proportion of female students among 9th grade students. This strategy consists in comparing students who face the same school environment (school resources, teacher quality, peers background) but are exposed to different proportions of female students for exogenous reasons. As students in French middle schools spend the whole year with a smaller group of peers, namely their classmates, we define the model at the class level and instrument the proportion of girls in the class with the proportion of girls in the 9th grade school cohort. Formally, we estimate the following model, defined at the class level:

$$Y_{jst} = \alpha + \beta_1 \widehat{Girls}_{jst} + \beta_2 X_{jst} + \gamma_s + \delta_t + \epsilon_{jst} \tag{3.1}$$

- Y_{jsc}= mean outcome of class j from school cohort st
- $\widehat{Girls}_{jst}$= instrumented proportion of girls in class j from school cohort st
- γ_s and δ_t = school and cohort fixed effects
- X_{jst} = class j's students' (financial aid status, foreign nationality, educational delay) and teachers' (experience, seniority, certification level, employment status and gender) average characteristics and class size

Model (3.1) assumes that the proportion of girls in the school cohort affects student outcomes only through its effect on the proportion of girls in the classroom. If the proportion of girls in other classes also affects student outcomes, this model may lead to biased estimates of β_1. To account for potential spillovers across classes, we also define a similar model in which we replace the instrumented proportion of girls in the classroom by the proportion of girls in

[15]This argument has been used in previous papers studying the effect of school peers' gender on student achievement, including Lavy & Schlosser (2011), Black et al. (2013) and Schøne et al. (2019). It is based on the combination of three features. First, compulsory schooling up to age 16 implies that, before that age, every child of a given birth cohort must enroll in a middle school. Second, the catchment area system imposes heavy constraints on school choice, provided that the child enrolls in a public middle school. Finally, the gender of a child is random by nature. Consequently, the gender ratio of a birth cohort living in a particular geographical area generates natural fluctuations around the 0.5 mean.

the school cohort:

$$Y_{jst} = \alpha + \beta_1 Girls_{st} + \beta_2 X_{jst} + \gamma_s + \delta_t + \epsilon_{jst} \tag{3.2}$$

Models (3.1) and (3.2) account for both between school and within school student sorting. The main identification assumption of these models is that, conditional on the school and neighbourhood time-invariant quality, there is no systematic relationship between short-term shocks on school quality (ϵ_{jst}) and the proportion of girls in the school cohort ($Girls_{st}$). This assumption would be violated if, for example, parents who observe a decline in teachers' or peers' quality were moving their daughters away from their public middle school more often than they would do with their sons.

To tackle this issue, we first provide evidence in the next section that within school cross cohort variations in the proportion of girls are not related to variations in teachers' quality, as measured by their level of experience or tenure, or in peers' quality, as measured by the proportion of students who have been held back a grade or who come from low-income families.

In addition, we test the robustness of our results to an alternative specification which better accounts for short-term trends in school quality. Again, if such trends exists, they would be problematic only insofar as they are related to the proportion of girls in the school cohort. In that case, the proportion of girls in the two adjacent school cohorts would likely capture short term trends in school quality. Consequently, we follow Gould et al. (2009) and we add the lead and the lag of the proportion of girls in the school cohorts as controls in model (3.1) and (3.2). For the remainder of the paper, we consider models (3.1) and (3.2) as our main specification, and we provide estimations of these models augmented with the lead and the lag of the proportion of girls in the school cohort in the appendix.

3.4 Results

3.4.1 Evidence on the validity of the identification assumption

The key identifying assumption to interpret the results as causal is that within school cross cohorts variations in the gender ratio are not correlated with changes in peers' and teachers' quality. To provide evidence on the validity of this assumption, we thus implement balancing tests on student and teacher characteristics, using our main specification.

Table C4 in the appendix shows the results of this test, implemented on student characteristics, separately for female and male students. Consistently with the identification assumption, the instrumented proportion of female students in the classroom is largely unrelated to the proportions of students from low-income families, foreign students and students with one year or more of educational delay. Table C5 further shows the results of this test, implemented on class size and teacher average characteristics. Comfortingly, this balancing test do not reveal any significant relationship between the instrumented proportion of female students in the class and class size, teacher experience, seniority, certification level or tenure. The only teacher characteristics which appear to be significantly related to the proportion of female student is the proportion of female teachers[16]. In total, these balancing tests strengthen the idea that within school cross cohort variations in the proportion of female students are unrelated to other changes in teachers' or peers' quality.

3.4.2 Short-term effects on test scores and behaviour

The first outcome that we consider is student standardized test scores at the end-of-middle-school national exam. As we can see in table 3.1, the proportion of girls among school peers has positive effects on girls' test scores while it negatively affects boys' one. More specifically, replacing 5 boys by 5 girls in a given class (i.e., a 20 pp increase in the proportion of girls in that class) would result in a 2% (1,5%) of a SD increase (decrease) in girls' (boys') average test scores. To give a sense of the magnitude of these effects, replacing 5 boys by 5 girls in a given class would roughly benefits girls as much as a 2-3 students reduction and harm boys as much as a 2 students increase in class size[17]. These negative effects of girls on boys' achievement are in sharp contrast with Hoxby (2000), Lavy & Schlosser (2011) and Hu (2015), who find strong positive effects on boys' test scores. Contrarily, the effects on girls' performance are consistent with these studies, who find that a 20 pp increase in the proportion of female peers increases girls' test scores by 2-6% of a standard deviation in test scores. In addition to this, the proportion of girls among school peers also has a positive and significant effect on girls' school discipline, as measured by their behaviour grade, while the effect is not significant for boys.

[16]When we implement these regressions separately by topics, it appears that this relationship is entirely driven by history-geography teachers.

[17]We draw this estimation from Piketty et al. (2006), who exploit the maximum class size rule to estimate the effect of reductions in class size on student test scores at the end-of-middle school national exam.

Table 3.1 The effect of female peers on student outcomes

	Class level (IV)		School level	
	(1)	(2)	(3)	(4)
	Girls	Boys	Girls	Boys
I. 9th grade outcomes				
End-of-middle-school test score	0.096**	-0.073**	0.093**	-0.071**
	(0.030)	(0.033)	(0.030)	(0.032)
Behaviour grade	0.489**	0.048	0.474**	0.047
	(0.156)	(0.176)	(0.152)	(0.170)
II. Track choices after 9th grade				
High school (general track)	0.053**	-0.040**	0.051**	-0.039**
	(0.015)	(0.015)	(0.015)	(0.015)
	[0.622]	[0.534]	[0.622]	[0.534]
Vocational school	-0.030**	0.039**	-0.030**	0.038**
	(0.014)	(0.014)	(0.013)	(0.014)
	[0.297]	[0.348]	[0.297]	[0.348]
Dropout	-0.019*	0.007	-0.018*	0.007
	(0.010)	(0.009)	(0.010)	(0.009)
	[0.104]	[0.095]	[0.104]	[0.095]
III. High school graduation				
Academic tracks (S, ES, L)	0.052**	-0.030**	0.050**	-0.029**
	(0.014)	(0.013)	(0.013)	(0.012)
	[0.384]	[0.295]	[0.384]	[0.295]
Science (S)	0.031**	-0.014	0.030**	-0.013
	(0.010)	(0.011)	(0.010)	(0.010)
	[0.162]	[0.190]	[0.162]	[0.190]
Economics/Social sciences (ES)	0.021**	-0.012	0.020**	-0.012
	(0.009)	(0.007)	(0.009)	(0.007)
	[0.132]	[0.081]	[0.132]	[0.081]
Literature (L)	-0.000	-0.004	-0.000	-0.004
	(0.008)	(0.004)	(0.008)	(0.004)
	[0.090]	[0.024]	[0.090]	[0.024]
Technological track	0.004	-0.007	0.004	-0.007
	(0.011)	(0.010)	(0.010)	(0.010)
	[0.152]	[0.144]	[0.152]	[0.144]
F-stats	5253	5253	.	.
Observations	82184	82184	82184	82184

Note: this table shows the effect of the proportion of girls among school peers on various student outcomes, using model (3.1) (columns (1) and (2)), and model (3.2) (columns (3) and (4)), estimated separately for girls and boys. The upper part of the table shows the effects of female peers on students' standardized test scores at the end of middle school national examination and behaviour grade. The middle part of the table shows the effects on the proportion of students who (1) attend a general high-school (2) attend a vocational school and (3) drop out of education, within the next 3 years following 9th grade. The lower part of the table shows the effects on the proportion of students who graduate from high school within the next 5 years following 9th grade, separately by tracks. Outcome sample means are within square brackets. Standard errors (in parentheses) are clustered at the school level. * p<0.10, ** p<0.05

3.4.3 Longer-term effects on track choices and educational attainment

At the end of middle school, students basically have three options: (1) going to high school to follow an academic track (2) going to a vocational school (3) dropping out of the educational system. We estimate the effect of the proportion of female students among school peers on these three outcomes, using a linear probability model with our main specification. This proportion has a strong positive effect on girls' probability to choose an academic track, and it also decreases their probability to choose a vocational track or to drop out. The magnitude of these effects is quite large: increasing by 20 pp the proportion of female students among school peers raises the proportion of girls choosing the academic track by 1.6% (1 pp) and will decrease the proportion of girls choosing a vocational track by 2% (0.6 pp) and girls dropout rate by 3.8% (0.38 pp). These positive effects are in contrast with Black et al. (2013), who find no effect on girls' academic track choices and dropout status. Contrarily, the proportion of girls in a class negatively affects boys' probability to choose the academic track in high school, and positively affects the probability to choose a vocational track. More specifically, a 20 pp increase in the proportion of girls will decrease the proportion of boys who choose the academic track by 1.5% (0.80 pp) and will increase the proportion of boys choosing a vocational track by 2.5% (0.88 pp). There is no significant effect on boys' dropout rates.

Eventually, we estimate the effects of the proportion of girls among school peers on the probability to graduate from high school, separately by tracks, using our main specification. We consider five outcomes: (1) graduating from high school in one of the three academic tracks (2) graduating from the *Science* track (3) the *Economics and Social sciences* track (4) the *Literature* track (5) the *Technological* track.

Generally speaking, the proportion of girls among school peers in middle school (9th grade) has long-lasting effects on student achievement, as it still influences their probability to graduate from high school several years later, both for girls and for boys. Consistently with the previous section, girls benefit from having more female peers in middle school, while it has a detrimental effects on boys. Increasing the proportion of girls by 20 pp generates a 2% (1 pp) increase in girls' high school graduation rate and a 2% (0.60 pp) decrease in boys' one. The positive effect of the proportion of girls in middle school on girls' high school graduation rate is particularly salient for the probability to graduate from the *Science* (+3%) and the *Economics and Social sciences* (+2.6%) tracks. This last result is particularly important as girls tend to be under-

represented in scientific tracks in post-secondary education, which are typically associated with better outcomes on the labour market.

3.4.4 Alternative specifications

We check the robustness of our main results to the inclusion of the lead and the lag of the proportion of girls in the school cohort in equations (3.1) and (3.2). Generally speaking, this alternative specification provides very similar results to those obtained from the main specification, for all outcomes under consideration (cf. table C6 in the appendix). This further suggests that our main results are not driven by short-term trends in school quality.

To test for potential non linearities in the effects of the proportion of girls among school peers, we estimate separately model (3.1) in schools in which there is a minority of girls over the period and in schools in which there is a majority of girls. It appears that the negative effects of girls on boys are concentrated in school in which there is a minority of girls, while the positive effects on girls are concentrated in schools where there is a majority of girls (cf. table C7). Put differently, the higher the average proportion of students of the same gender among school peers, the stronger the benefits of being exposed to a even higher proportion of these students.

3.5 Potential mechanisms

There are three potential channels through which a change in the gender composition of a school cohort may affect student achievement: the general quality of the learning environment, interactions among students and student-teacher interactions. Previous studies consistently find that a higher proportion of girls is beneficial to the general quality of the learning environment (Hoxby (2000); Lavy & Schlosser (2011); Hu (2015)), as girls tend to be more disciplined and conscientious than boys on average (Duckworth & Seligman (2006); Jacob (2002); Bertrand & Pan (2013); Cornwell et al. (2013)). On the other hand, the effects of girls on interactions among students and between students and teachers is much more ambiguous and may sharply differ across genders. An early work by Coleman (1961) suggests that studying in mixed-gender schools may provides greater distraction for students and may, in turn, be detrimental to their achievement. A recent study by Hill (2015) provides empirical evidence that a higher proportion of opposite-gender school friends increases students' probability of entering a romantic

relationship and difficulties getting along with their teachers and paying attention in class in US high school. These two effects translate into worse school outcomes. In a related work, Lu & Anderson (2014) show that Chinese middle-school students tend to be more cooperative with same-gender students than with opposite-gender ones, and that this increased cooperation within the classroom results in improved student outcomes.

We build on the recent literature on teachers' grading biases to investigate whether the effects of girls on student achievement are mediated by student classroom behaviour and relationship with teachers in our data. More specifically, Lavy (2008) and Terrier (2015) exploit the difference between girls and boys in the difference between blind and non-blind test scores (i.e., the teacher grading bias) as a measure of teacher gender-biased behaviour in favor of girls. In this paper, we rather interpret this teacher grading bias measure as a mixed index of the quality of girls' classroom behaviour and relationships with their teachers relative to boys. Consistent with our findings, when we regress this index on the proportion of girls among school peers using our main specification, we find a positive relationship between the proportion of girls and teacher grading bias in their favor (cf. table C8). This result supports the notion that a higher proportion of opposite gender students among school peers is detrimental to student classroom behaviour and student-teacher relationships (Hill (2015)).

3.6 Conclusion

This paper shows that school cohort gender composition in middle school have persistent effects on student achievement and human capital acquisition. Being surrounded by more female peers has a positive influence on girls' performance at the end-of-middle school national examination, whereas it has a detrimental effect on boys' performance. Furthermore, it decreases girls' dropout rate after middle school and increases their probability to graduate from high school, especially from the scientific track. By contrast, it decreases boys' probability to graduate from high school and increases their probability to attend a vocational school. An investigation of the potential mechanisms at play suggests that these effects may partially operate through a negative effect of opposite-gender peers on students' classroom behaviour and relationships with their teachers. Altogether, these results tend to show that school peers' gender in middle school shape students skills and preferences in a persistent way.

Generally speaking, our main results are consistent with a "gender" boutique model (Hoxby & Weingarth (2005)), which states that students benefit from similar peers. Grouping girls together would not only help them learning better but would also encourage them to make educational choices that lead them to higher paying jobs, closing the gender gap on the labour market. Moreover, such a policy would be highly cost-effective, as its costs may be close to zero. However, this conclusion requires a caveat, as students may also build non cognitive skills at school, such as tolerance or social and political attitudes, that are crucial for the well functioning of democratic societies. Accordingly, Merlino et al. (2019) show that a higher share of black students of the same gender among high school peers in the US has a positive effect on racial tolerance and interracial romantic relationships in adulthood. Hence, the effects of gender diversity at school on non cognitive outcomes deserves more attention in future works.

CONCLUSION GÉNÉRALE

Cette thèse de doctorat présente trois essais en économie de l'éducation. Ces essais explorent les déterminants de l'efficacité pédagogique des enseignants ainsi que les moyens d'améliorer cette efficacité à travers la mise en place de politiques publiques. Ils cherchent également à mieux comprendre les effets de long terme des interactions sociales qui ont lieu au collège.

En premier lieu, le travail réalisé dans cette thèse a permis d'établir un lien empirique systématique entre l'efficacité pédagogique des enseignants et les pratiques pédagogiques qu'ils mettent en œuvre en classe. En particulier, nous montrons dans le premier chapitre de cette thèse que les enseignants de mathématiques aux Etats-Unis sont d'autant plus efficaces qu'ils mettent en place des pratiques pédagogiques interactives, requérant une participation active de la part des élèves. Si la littérature existante apporte des preuves abondantes quant aux grandes disparités d'efficacité pédagogique entre enseignants, elle fournit en revanche très peu d'éléments empiriques permettant de comprendre d'où proviennent ces disparités. Étant donné l'importance des enseignants dans le processus d'apprentissage des élèves (Hanushek & Rivkin (2010)), les résultats présentés dans ce premier chapitre apparaissent particulièrement pertinents, notamment dans l'optique d'élaborer des politiques publiques de formation des enseignants et d'amélioration de leurs pratiques pédagogiques. Néanmoins, si ce chapitre permet de conclure que les enseignants qui mettent en place des pratiques pédagogiques interactives sont plus efficaces, il ne permet pas d'en conclure que chaque enseignant *serait* plus efficace s'il mettait en place de telles pratiques. En particulier, l'adoption de nouvelles pratiques pédagogiques peut comporter un coût, lié à l'apprentissage et à l'adaptation du style d'enseignement, et nécessiter d'autres compétences préalables, telles que la capacité à mettre en place un climat de classe propice aux interactions entre élèves et avec l'enseignant. De nouveaux travaux de recherche basé sur un cadre méthodologique expérimental semblent fortement souhaitables pour apporter des éléments de réponse à cette question.

Dans la continuité du premier chapitre, le deuxième chapitre de cette thèse apporte un éclairage nouveau sur l'efficacité des politiques publiques visant à améliorer l'efficacité pédagogique des enseignants. En effet, nous démontrons dans ce chapitre que le système d'inspection

individuelle des enseignants qui caractérise l'enseignement secondaire en France permet de générer des améliorations durables et généralisées de l'efficacité pédagogique des enseignants. Ces améliorations sont constantes au cours de la carrière des enseignants et sont particulièrement visibles dans les contextes d'éducation prioritaire. L'inspection apparaît alors comme un outil pertinent de politique éducative, tant du point de vue de son efficacité que de son effet sur les inégalités scolaires. La plupart des travaux portant sur les effets des différentes politiques éducatives visant à augmenter l'efficacité pédagogique des enseignants montrent qu'il est très difficile d'y parvenir, à moins d'investir des ressources considérables (Jackson et al. (2014)). Les résultats présentés dans le cadre de ce chapitre démontrent qu'il est possible d'y parvenir à un coût bien moindre, à travers un système d'évaluations des enseignants généralisées et répétées tout au long de leur carrière. Néanmoins, ces résultats soulèvent également de nouvelles questions quant aux mécanismes permettant de générer de tels effets. En particulier, il se pourrait que ces effets proviennent des conseils prodigués aux enseignants par les inspecteurs, du travail individuel des enseignants en préparation de leurs inspections ou encore d'une collaboration accrue avec les autres enseignants du collège (Jackson & Bruegmann (2009)). Une étude approfondie des mécanismes permettant aux enseignants d'améliorer leurs compétences pédagogiques serait particulièrement intéressante à cet égard.

Le dernier chapitre de cette thèse étudie un autre élément essentiel de l'environnement scolaire susceptible d'influencer la réussite des élèves : les caractéristiques de leurs camarades de classe. Nous montrons dans ce chapitre que le genre des camarades de classe au collège a un effet important sur l'apprentissage des élèves et sur leur trajectoire scolaire future. Pour une fille, être entourée d'un plus grand nombre de filles parmi ses camarades de classe a des effets bénéfiques sur ses performances au brevet et augmente ses chances d'obtenir le baccalauréat plusieurs années plus tard, notamment dans la filière scientifique. À l'inverse, les garçons semblent être négativement affectés par la présence d'un plus grand nombre de filles parmi leurs camarades de classe au collège. Dans l'ensemble, ces résultats confirment l'importance des camarades de classe comme facteur de réussite scolaire d'un élève (Sacerdote (2011)) et suggèrent que les interactions sociales au collège influencent les compétences et les préférences des élèves de manière durable. Par ailleurs, l'impact négatif d'une présence plus grande d'élèves du sexe opposé est cohérente avec divers travaux montrant que cette présence est généralement associée avec un niveau de coopération plus faible entre les élèves (Lu & Anderson (2014)) et accroît les

possibilités de distraction qui détournent les élèves de l'instruction (Hill (2015)). Néanmoins, il est important de souligner que l'ensemble de ces travaux ainsi que le troisième chapitre de cette thèse ne prennent pas en compte d'autres dimensions primordiales de l'apprentissage des élèves. En effet, le collège est également le lieu où un ensemble d'attitude nécessaire au vivre-ensemblee et au bon fonctionnement d'une société démocratique, telle que la tolérance ou la non-violence, et ces attitudes sont également fortement influencées par les caractéristiques des pairs (Merlino et al. (2019)). Dans ce contexte, de nouvelles études sur les effets des caractéristiques des pairs sur ce type de compétences et d'attitudes semblent particulièrement souhaitables.

Bibliography

Aaronson, D., Barrow, L., & Sander, W. 2007. "Teachers and student achievement in the Chicago public high schools". *Journal of Labor Economics*, 25(1):95–135.

Algan, Y., Cahuc, P., & Shleifer, A. 2013. "Teaching Practices and Social Capital". *American Economic Journal: Applied Economics*, 5(3):189–210.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. 2011. "An interaction-based approach to enhancing secondary school instruction and student achievement". *Science*, 333 (6045):1034–1037.

Andrietti, V. 2015. "The causal effects of increased learning intensity on student achievement: Evidence from a natural experiment". UC3M Working Paper Economic Series 15-06.

Anelli, M. & Peri, G. 2017. "The effects of high school peers' gender on college major, college performance and income". *The Economic Journal*, 129(618):553–602.

Angrist, J. D. & Lavy, V. 2001. "Does teacher training affect pupil learning? Evidence from matched comparisons in Jerusalem public schools". *Journal of Labor Economics*, 19(2): 343–369.

Anil, B., Guner, D., Delibasi, T., & Uysal, G. 2016. "Does Classroom Gender Composition Affect School Dropout?". IZA Discussion Paper No. 10238.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. 2016. "Teacher Quality and Learning Outcomes in Kindergarten". *The Quarterly Journal of Economics*, 131(3):1415–1453.

Aslam, M. & Kingdon, G. 2011. "What can teachers do to raise pupil achievement?". *Economics of Education Review*, 30(3):559–574.

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. 2017. "An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys". National Bureau of Economic Research.

Becker, G. S. 1962. "Investment in human capital: A theoretical analysis". *Journal of political economy*, 70(5, Part 2):9–49.

Becker, G. S. 2011. "Reflections on the Economics of Education". In *Handbook of the Economics of Education*, volume 4. Elsevier.

Bellei, C. 2009. "Does lengthening the school day increase students' academic achievement? Results from a natural experiment in Chile". *Economics of Education Review*, 28(5):629–640.

Bertrand, M. & Pan, J. 2013. "The trouble with boys: Social influences and the gender gap in disruptive behavior". *American Economic Journal: Applied Economics*, 5(1):32–64.

Bietenbeck, J. 2014. "Teaching practices and cognitive skills". *Labour Economics*.

Black, S. E., Devereux, P. J., & Salvanes, K. G. 2013. "Under pressure? The effect of peers on outcomes of young adults". *Journal of Labor Economics*, 31(1):119–153.

Blazar, D. 2015. "Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement". *Economics of Education Review*, 48:16–29.

Booth, A. L., Cardona-Sosa, L., & Nolen, P. 2018. "Do single-sex classes affect academic achievement? An experiment in a coeducational university". *Journal of Public Economics*, 168:109–126.

Bruner, J. S. 1961. "The act of discovery.". *Harvard educational review*.

Chetty, R., Friedman, J. N., & Rockoff, J. E. 2014. "Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood". *The American Economic Review*, 104(9):2633–2679.

Coleman, J. S. 1961. *The adolescent society: The social life of the teenager and its impact on education*. New York: Free Press.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & Robert, L. 1966. "York, 1966: Equality of educational opportunity". *Washington: US Government Printing Office*.

Cornwell, C., Mustard, D. B., & Van Parys, J. 2013. "Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school". *Journal of Human Resources*, 48(1):236–264.

Cortes, K. E., Goodman, J. S., & Nomi, T. 2015. "Intensive math instruction and educational attainment long-run impacts of double-dose algebra". *Journal of Human Resources*, 50(1): 108–158.

Dobbie, W. & Fryer Jr, R. G. 2013. "Getting beneath the veil of effective schools: Evidence from New York City". *American Economic Journal: Applied Economics*, 5(4):28–60.

Duckworth, A. L. & Seligman, M. E. 2006. "Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores.". *Journal of educational psychology*, 98 (1):198.

Dustmann, C., Ku, H., & Kwak, D. W. 2018. "Why are single-sex schools successful?". *Labour Economics*, 54:79–99.

Eisenkopf, G., Hessami, Z., Fischbacher, U., & Ursprung, H. W. 2015. "Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland". *Journal of Economic Behavior & Organization*, 115:123–143.

Eurydice. 2018. *Teaching Careers in Europe: Access, Progression and Support. Eurydice Report*. Eurydice Report. Luxembourg: Publications Office of the European Union.

Fack, G. & Grenet, J. 2012. "Rapport d'évaluation de l'assouplissement de la carte scolaire". École d'économie de Paris.

Fryer, R. G. 2014. "Injecting charter school best practices into traditional public schools: Evidence from field experiments". *The Quarterly Journal of Economics*, 129(3):1355–1407.

Galambos, N. L. 2004. "Gender and gender role development in adolescence". In *Handbook of adolescent psychology*, volume 2, pages 233–262.

Gaubert, É., Henrard, V., Robert, A., & Rouaud, P. 2017. "Enquête 2016 auprès de la Génération 2013. Pas d'amélioration de l'insertion professionnelle pour les non-diplômés". *Céreq bref*, (356).

Glazerman, S., Dolfin, S., Bleeker, M., Johnson, A., Isenberg, E., Lugo-Gil, J., Grider, M., Britton, E., & Ali, M. 2008. "Impacts of Comprehensive Teacher Induction: Results from the First Year of a Randomized Controlled Study. NCEE 2009-4034.". *National Center for Education Evaluation and Regional Assistance*.

Glazerman, S., Isenberg, E., Dolfin, S., Bleeker, M., Johnson, A., Grider, M., & Jacobus, M. 2010. "Impacts of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study. NCEE 2010-4027.". *National Center for Education Evaluation and Regional Assistance*.

Goodman, J. 2017. "The Labor of Division: Returns to Compulsory High School Math Coursework". National Bureau of Economic Research.

Gould, E. D., Lavy, V., & Daniele Paserman, M. 2009. "Does immigration affect the long-term educational outcomes of natives? Quasi-experimental evidence". *The Economic Journal*, 119 (540):1243–1269.

Haeck, C., Lefebvre, P., & Merrigan, P. 2014. "The distributional impacts of a universal school reform on mathematical achievements: A natural experiment from Canada". *Economics of Education Review*, 41:137–160.

Hanushek, E. A. & Rivkin, S. G. 2006. “Teacher quality”. In *Handbook of the Economics of Education*, volume 2, pages 1051–1078. Elsevier, 2006.

Hanushek, E. A. & Rivkin, S. G. 2010. “Generalizations about using value-added measures of teacher quality”. *American Economic Review*, 100(2):267–71.

Hanushek, E. A. & Woessmann, L. 2008. “The role of cognitive skills in economic development”. *Journal of Economic Literature*, 46(3):607–668.

Hanushek, E. A. & Woessmann, L. 2011. “The Economics of International Differences in Educational Achievement”. In *Handbook of the Economics of Education*, volume 3, pages 89–200. Elsevier, 2011.

Harris, D. N. & Sass, T. R. 2011. “Teacher training, teacher quality and student achievement”. *Journal of Public Economics*, 95(7-8):798–812.

Hidalgo-Cabrillana, A. & Lopez-Mayan, C. 2018. “Teaching styles and achievement: Student and teacher perspectives”. *Economics of Education Review*, 67:184–206.

Hill, A. J. 2015. “The girl next door: The effect of opposite gender friends on high school achievement”. *American Economic Journal: Applied Economics*, 7(3):147–77.

Hill, A. J. 2017. “The positive influence of female college students on their male peers”. *Labour Economics*, 44:151–160.

Hoxby, C. 2000. “Peer effects in the classroom: Learning from gender and race variation”. National Bureau of Economic Research.

Hoxby, C. M. & Weingarth, G. 2005. “Taking race out of the equation: School reassignment and the structure of peer effects”. Working paper.

Hu, F. 2015. “Do girl peers improve your academic performance?”. *Economics Letters*, 137: 54–58.

Huebener, M., Kuger, S., & Marcus, J. 2017. “Increased instruction hours and the widening gap in student performance”. *Labour Economics*.

IGEN. 2011. “Mission sur le rôle et l’activité des inspecteurs pédagogiques du second degré, Note à Monsieur le ministre de l’Education nationale, de la jeunesse et de la vie associative”. Note n 2011-02.

IGEN/IGAENR. 2006. “La contribution de l’éducation prioritaire à l’égalité des chances des élèves”. Rapport n 2006-076.

IGEN/IGAENR. 2016. “Rôle et positionnement des inspecteurs du second degré en académie”. Rapport n 2016-070.

Isoré, M. 2009. "Teacher evaluation: Current practices in OECD countries and a literature review". OECD Education Working Papers, No. 23, OECD Publishing, Paris.

Jackson, C. K. 2012. "Single-sex schools, student achievement, and course selection: Evidence from rule-based student assignments in Trinidad and Tobago". *Journal of Public Economics*, 96(1-2):173–187.

Jackson, C. K. 2016. "The effect of single-sex education on test scores, school completion, arrests, and teen motherhood: evidence from school transitions". National Bureau of Economic Research.

Jackson, C. K. & Bruegmann, E. 2009. "Teaching students and teaching each other: The importance of peer learning for teachers". *American Economic Journal: Applied Economics*, 1(4):85–108.

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. 2014. "Teacher effects and teacher-related policies". *Annu. Rev. Econ.*, 6(1):801–825.

Jacob, B. A. 2002. "Where the boys aren't: Non-cognitive skills, returns to school and the gender gap in higher education". *Economics of Education review*, 21(6):589–598.

Jacob, B. A. & Lefgren, L. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education". *Journal of Labor Economics*, 26(1): 101–136.

Joensen, J. S. & Nielsen, H. S. 2009. "Is there a causal effect of high school math on labor market Outcomes?". *Journal of Human Resources*, 44(1):171–198.

Kabla-Langlois, I. 2016. "Les jeunes et l'enseignement supérieur : s'orienter, réussir, s'insérer". In *France, protrait social*, pages 27–42. Insee.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. 2011. "Identifying effective classroom practices using student achievement data". *Journal of Human Resources*, 46(3):587–613.

Lavy, V. 2008. "Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment". *Journal of Public Economics*, 92(10-11):2083–2105.

Lavy, V. 2009. "Performance pay and teachers' effort, productivity, and grading ethics". *American Economic Review*, 99(5):1979–2011.

Lavy, V. 2012. "Expanding school resources and increasing time on task: effects of a policy experiment in Israel on student academic achievement and behavior". National Bureau of Economic Research.

Lavy, V. 2015a. "Do differences in schools' instruction time explain international achievement gaps? Evidence from developed and developing countries". *The Economic Journal*, 125 (588):F397–F424.

Lavy, V. 2015b. "What makes an effective teacher? Quasi-experimental evidence". *CESifo Economic Studies: Oxford Journals*.

Lavy, V. & Schlosser, A. 2011. "Mechanisms and impacts of gender peer effects at school". *American Economic Journal: Applied Economics*, 3(2):1–33.

Lazear, E. P. 2001. "Educational production". *The Quarterly Journal of Economics*, 116(3): 777–803.

Lee, S., Turner, L. J., Woo, S., & Kim, K. 2014. "All or nothing? The impact of school and classroom gender composition on effort and academic achievement". National Bureau of Economic Research.

Lobel, T. E., Nov-Krispin, N., Schiller, D., Lobel, O., & Feldman, A. 2004. "Gender discriminatory behavior during adolescence and young adulthood: A developmental analysis". *Journal of youth and adolescence*, 33(6):535–546.

Lu, F. & Anderson, M. L. 2014. "Peer effects in microenvironments: The benefits of homogeneous classroom groups". *Journal of Labor Economics*, 33(1):91–122.

Machin, S. & McNally, S. 2008. "The literacy hour". *Journal of Public Economics*, 92(5): 1441–1462.

Marcotte, D. E. 2007. "Schooling and test scores: A mother-natural experiment". *Economics of Education Review*, 26(5):629–640.

Marcotte, D. E. & Hemelt, S. W. 2008. "Unscheduled school closings and student performance". *Education*, 3(3):316–338.

Mayer, R. E. 2004. "Should there be a three-strikes rule against pure discovery learning?". *American psychologist*, 59(1):14.

Merlino, L. P., Steinhardt, M. F., & Wren-Lewis, L. 2019. "More than just friends? School peers and adult interracial relationships". *Journal of Labor Economics*, 37(3):000–000.

Mincer, J. 1970. "The distribution of labor incomes: a survey with special reference to the human capital approach". *Journal of economic literature*, 8(1):1–26.

Mincer, J. 1974. *Schooling, Experience, and Earnings. Human Behavior & Social Institutions No. 2*. ERIC.

Monso, O., Fougere, D., Givord, P., & Pirus, C. 2019. "Les camarades influencent-ils la réussite et le parcours des élèves?". Sciences Po LIEPP Working Paper.

Murphy, R., Weinhardt, F., & Wyness, G. 2018. "Who Teaches the Teachers? A RCT of Peer-to-Peer Observation and Feedback in 181 Schools". CEP Discussion Paper No 1565.

National Council of Teachers of Mathematics. 1991. *Professional standards for teaching mathematics*. Commission on Teaching Standards for School Mathematics.

OECD. 2013a. *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*. OECD Reviews of Evaluation and Assessment in Education, Editions OCDE, Paris.

OECD. 2013b. *Teachers for the 21st Century: Using Evaluation to Improve Teaching*. OECD Publishing.

OECD. 2017. "Education at a Glance 2017". OECD Publishing.

Oosterbeek, H. & Van Ewijk, R. 2014. "Gender peer effects in university: Evidence from a randomized experiment". *Economics of Education Review*, 38:51–63.

Papay, J. P. & Kraft, M. A. 2015. "Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement". *Journal of Public Economics*, 130:105–119.

Park, H., Behrman, J. R., & Choi, J. 2013. "Causal effects of single-sex schools on college entrance exams and college attendance: Random assignment in Seoul high schools". *Demography*, 50(2):447–469.

Park, H., Behrman, J. R., & Choi, J. 2018. "Do single-sex schools enhance students' STEM (science, technology, engineering, and mathematics) outcomes?". *Economics of Education Review*, 62:35–47.

Piaget, J. 1970. *Science of education and the psychology of the child.* Trans. D. Coltman. Orion.

Piketty, T., Valdenaire, M., et al. 2006. *L'impact de la taille des classes sur la réussite scolaire dans les écoles, collèges et lycées français: estimations à partir du panel primaire 1997 et du panel secondaire 1995*. Direction de l'évaluation et de la prospective.

Rivkin, S. G. & Schiman, J. C. 2015. "Instruction time, classroom quality, and academic achievement". *The Economic Journal*, 125(588):F425–F448.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. 2005. "Teachers, schools, and academic achievement". *Econometrica*, 73(2):417–458.

Rockoff, J. E. 2004. "The impact of individual teachers on student achievement: Evidence from panel data". *American Economic Review*, pages 247–252.

Rockoff, J. E. 2008. "Does mentoring reduce turnover and improve skills of new employees? Evidence from teachers in New York City". National Bureau of Economic Research.

Rose, H. & Betts, J. R. 2004. "The effect of high school courses on earnings". *Review of Economics and Statistics*, 86(2):497–513.

Sacerdote, B. 2011. "Peer effects in education: How might they work, how big are they and how much do we know thus far?". In *Handbook of the Economics of Education*, volume 3, pages 249–277. Elsevier.

Schneeweis, N. & Zweimüller, M. 2012. "Girls, girls, girls: Gender composition and female school choice". *Economics of Education review*, 31(4):482–500.

Schøne, P., Simson, K. v., & Strøm, M. April 2019. "Peer gender and educational choices". *Empirical Economics*. doi: 10.1007/s00181-019-01697-2.

Schwerdt, G. & Wuppermann, A. C. 2011. "Is traditional teaching really all that bad? A within-student between-subject approach". *Economics of Education Review*, 30(2):365–379.

Sims, D. P. 2008. "Strategic responses to school accountability measures: It's all in the timing". *Economics of Education Review*, 27(1):58–68.

Steinberg, L. & Monahan, K. C. 2007. "Age differences in resistance to peer influence.". *Developmental psychology*, 43(6):1531.

Strain, M. R. 2013. "Single-sex classes & student outcomes: Evidence from North Carolina". *Economics of Education Review*, 36:73–87.

Taylor, E. 2014. "Spending more of the school day in math class: Evidence from a regression discontinuity in middle school". *Journal of Public Economics*, 117:162–181.

Taylor, E. S. & Tyler, J. H. 2012. "The effect of evaluation on teacher performance". *American Economic Review*, 102(7):3628–51.

Terrier, C. 2015. "Giving a little help to girls? Evidence on grade discrimination and its effect on students' achievement". Centre for Economic Performance, London School of Economics and Political Science.

Van Klaveren, C. 2011. "Lecturing style teaching and student performance". *Economics of Education Review*, 30(4):729–739.

Vygotsky, L. S. 2012. *Thought and language*. MIT press.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. 2009. *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. ERIC.

Whitmore, D. 2005. "Resource and peer impacts on girls' academic achievement: Evidence from a randomized experiment". *American Economic Review*, 95(2):199–203.

Wiswall, M. 2013. "The dynamics of teacher quality". *Journal of Public Economics*, 100: 61–78.

Appendices

Appendix A - Chapter 1

Principal Component Analysis and construction of the Modern Practices Index

This section of the appendix describes the construction of the *Modern Practices Index*, which is the main measure of teaching practices used in the first chapter. We first perform a Principal Component Analysis (PCA) at the teacher level, including the 11 teaching practice variables described in section 1.1.3. Figure A1 plots the different practices on the two first axis of the PCA, which summarizes 37% of the between teacher total variation in these 11 variables. The first axis clearly opposes *student-centered* practices, which are based on student active participation, to *teacher-centered* practices and practices based on memorization and routine problems solving. These two sets of practices roughly correspond to what has been called *Modern practices* and *Traditional practices* in the economics of education literature, and this classification is consistent with the main psychological theories of learning. In particular, these theories oppose the *transmissive* approach, where the teacher delivers knowledge to a passive learner, and the *constructivist* or *socio-constructivist* approach[18], which has been promoted in the US by the National Council of Teachers of Mathematics (1991) over the last two decades and for which "learning is an active process in which learners are active sense makers who seek to build coherent and organized knowledge" (Mayer (2004)).

To sum up the opposition between the two sets of practices, we create the *Modern Practices Index* (MPI), which is equal to the individual teacher's average score over practices (g), (h), (i) and (j). The MPI goes from -0.5 and 0.5, with a mean of -0.07 (cf. table A1), and is roughly normally distributed (cf. figure A2). In order to take into account the second axis of the PCA, the frequency of assessment (practice (k)) is included separately in the regressions. We additionally create two *Traditional* indexes corresponding to the two subsets of traditional practices, in order to check the robustness of our results to considering more dimensions of teaching. These two indexes equal the teacher's average score over practices (a), (c) and (d), and practices (b), (e) and (f), respectively.

[18] See Piaget (1970), Bruner (1961) and Vygotsky (2012)

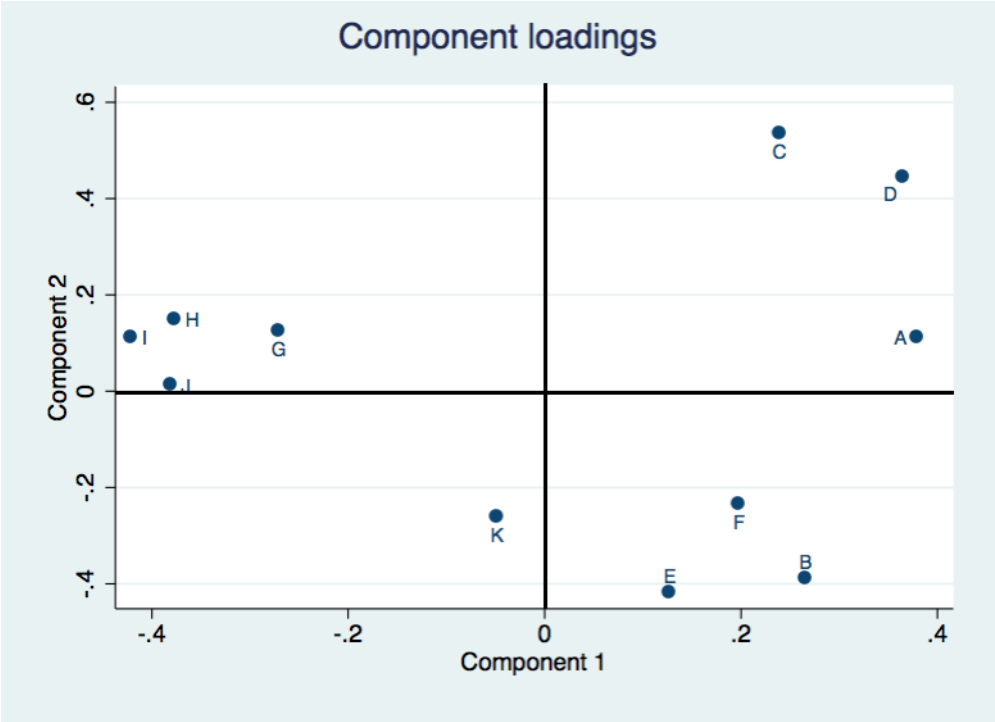


Figure A1 Principal Component Analysis - Teaching Practices

Note: Figure A1 plots the component loadings of the 11 teaching practices listed in table 1.1 on the two first axis of the principal component analysis, which is performed at the teacher level. Teaching practices are denoted with a letter, which refers to table 1.1.

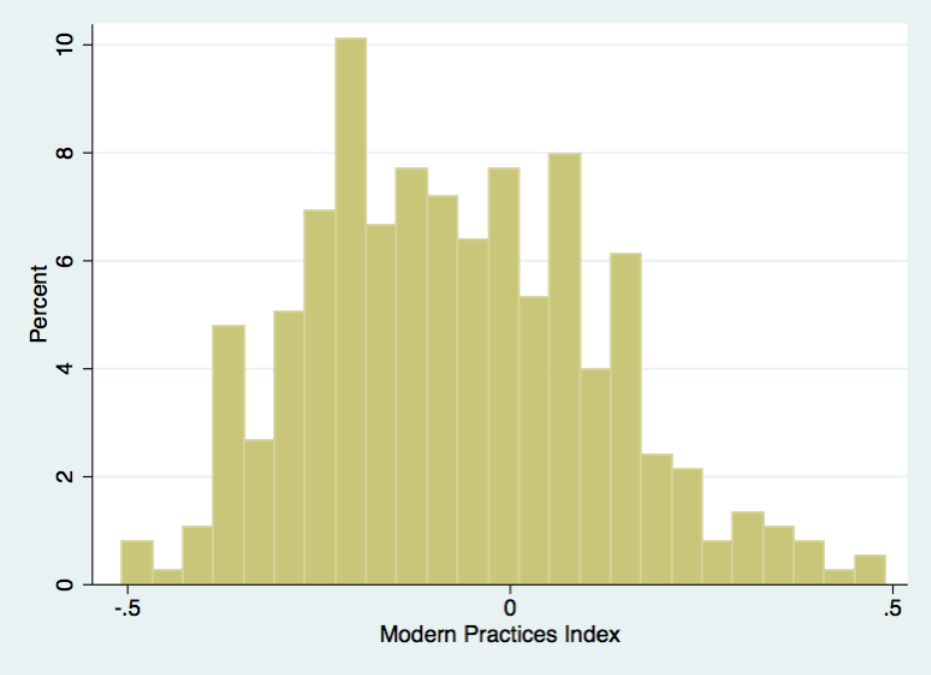


Figure A2 Distribution of the Modern Practices Index

Table A1 Distribution of the Modern Practices Index

Variable	Mean	SD	Min	p25	p50	p75	Max
Modern Practices Index	-0.07	0.19	-0.51	-0.21	-0.08	0.06	0.49

Note: This table shows the mean and standard deviation of the Modern Practices Index (MPI). It also shows the minimum, the maximum, and the quartiles (p25, p50 and p75) of the MPI.

Additional Tables and Graphs

Table A2 Teacher non response and student characteristics

Variable	Final sample (1)	Dropped students (2)	Mean Difference (1) - (2)
Female	0.508	0.493	0.0128 (0.89)
Age	14.26	14.22	0.0361* (1.82)
Foreign language spoken at home	1.374	1.387	-0.0122 (-0.35)
Educational aspirations	5.303	5.263	0.0434 (1.14)
Nb of books at home	2.882	2.884	0.00379 (0.06)
Parents' education	2.033	2.016	0.0171 (0.27)
Math test score	507	496	11* (1.95)
N	372	163	

Note: This table shows the mean characteristics of students whose math teachers answered the teacher questionnaire (column (1)) and students whose math teacher didn't answer the questionnaire (column (2)), in terms of student age, gender, language spoken at home, educational aspirations, parental education and math performance at the TIMSS test, computed at the teacher level. Eventually, column (3) shows the difference between these two groups of students and provides t-test of the significance of the average difference in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3 Teacher non response and school characteristics

Variable	Final sample (1)	Dropped schools (2)	Mean Difference (1) - (2)
School size	727	740	-13 (-0.37)
School remoteness	3.51	3.26	0.254 (1.53)
Average income level of area	2.333	2.272	0.06 (0.89)
Total number of computers	116.28	118.46	-2.175 (-0.22)
Shortage of math teacher	1.559	1.512	0.046 (0.48)
Math resource shortages	11.02	10.96	0.051 (0.20)
N	329	125	456

Note: This table shows the mean characteristics of schools in which the math teachers answered the teacher questionnaire (column (1)) and schools in which the math teacher didn't answer the questionnaire (column (2)), in terms of school size, remoteness, average income level of area, number of computers and math teachers' and resources' shortages, computed at the school level. Eventually, column (3) shows the difference between these two groups of schools and provides t-test of the significance of the average difference in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4 Distribution of math instructional time across math topics

	Mean	SD	p25	p50	p75
Math instructional time in hours/week:					
Number	0.85	0.73	0.35	0.72	1.15
Algebra	2.37	1.44	1.26	2.22	3.33
Geometry	0.75	0.79	0.21	0.58	1.00
Data & Chance	0.45	0.38	0.19	0.39	0.66
Total	4.42	1.63	3.75	4.17	5.00

Note: This table shows the mean and standard deviation of math instructional time per topics, expressed in hours per week and computed at the teacher level. $p25$, $p50$ and $p75$ respectively represent the 25th, the 50th and the 75th percentile of the instructional time variable distribution.

Table A5 Pairwise correlation coefficients among teaching practice variables

Teaching Practice	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
(a)	1.000										
(b)	0.364**	1.000									
(c)	0.334**	0.242**	1.000								
(d)	0.486**	0.324**	0.495**	1.000							
(e)	0.200**	0.274**	0.215**	0.202**	1.000						
(f)	0.346**	0.377**	0.358**	0.316**	0.327**	1.000					
(g)	-0.003	0.046	0.107**	0.072	0.065	0.238**	1.000				
(h)	-0.078	0.001	0.048	-0.013	-0.014	0.090*	0.277**	1.000			
(i)	-0.055	0.032	0.118**	-0.002	0.127**	0.144**	0.284**	0.406**	1.000		
(j)	0.001	0.083	0.098*	-0.003	0.129**	0.127**	0.270**	0.295**	0.509**	1.000	
(k)	0.092*	0.164**	0.034	0.041	0.024	0.142**	0.064	0.106**	0.019	0.099*	1.000

Note: This table shows all the pairwise correlation coefficients among teaching practice variables. Teaching practices are denoted with a letter, which refers to table 1.1. * $p < 0.10$, ** $p < 0.05$.

Table A6 Modern Practices Index and Teacher, School and Student characteristics

	Correlation coefficient
Teacher characteristics (N=372)	
Experience	-0.06
Female	0.03
Education level	-0.01
Major area of study = mathematics	0.05
Major area of study = education - mathematics	0.11**
Professional development in math content	0.18***
Professional development in math pedagogy	0.13**
Professional development in math curriculum	0.11**
Confidence in teaching math	0.38***
Collaboration with colleagues	0.13**
School characteristics (N=355)	
School size	-0.01
School remoteness	-0.05
Average income level of area	0.03
Total number of computers	-0.05
Math resource shortages	-0.03
Student and class characteristics (N=372)	
Female	0.09*
Age	0.05
Foreign language spoken at home	0.06
Educational aspirations	0.04
Nb of books at home	-0.06
Parents' education level	-0.08
Class size	-0.03
Classroom disruption (perceived by the teacher)	-0.07

Note: This table shows pairwise correlation coefficients between the Modern Practices Index (MPI) and teacher, school and student characteristics. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table A7 Modern Practices Index and the allocation of Instructional Time across topics - regression

	(1) Modern Practices Index
Number	-0.0019 (0.0012)
Algebra	-0.0014 (0.0010)
Geometry	-0.0003 (0.0012)
Data & chance	0.0008 (0.0018)
Observations	372

Note: This table shows the estimated coefficients from the regression of the Modern Practices Index (MPI) on the percentages of instructional time devoted to each of the four math topics, controlling for the class mean score in math, computed over subtopics not taught the year of the TIMSS assessment. The regression is implemented at the teacher level. Standard errors are in parenthesis. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A8 Modern Practices Index and the allocation of Instructional Time across topics - pairwise correlation coefficients

	Correlation coefficient
Number	-0.06
Algebra	-0.07
Geometry	0.09*
Data & chance	0.07
Observations	372

Note: This table shows the correlation coefficients between the Modern Practices Index (MPI) and the percentage of math instructional time dedicated to each of the four topics. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A9 Instructional Productivity along the MPI distribution

Position of the teacher in the *MPI* distribution (1)	$MPI_{p^{th}}$ value (2)	Teacher Instructional Productivity (in σ-test score) (3)	Change in productivity relative to the median teacher (4)
10th percentile	-0.30	0.021	- 50%
25th percentile	-0.21	0.030	- 28%
50th percentile	-0.08	0.042	0%
75th percentile	0.06	0.056	+ 33%
90th percentile	0.16	0.065	+ 55%

Note: This table shows the effect of one weekly hour of math instructional time on student performance in math, estimated at different points of the *Modern Practices Index* (MPI) distribution, assuming a linear relationship bewteen the MPI and teachers' instructional productivity. Point estimates shown in column (3) are computed as follows: $Productivity_{p^{th}} = \hat{\beta}_1 + \hat{\beta}_2 MPI_{p^{th}}$, with $MPI_{p^{th}}$ the value of MPI in column (2) and $\hat{\beta}_1$ and $\hat{\beta}_2$ the coefficients associated to Instructional Time and to the interaction term between Instructional Time and MPI, respectively, estimated from our main regression. The first line of the Table might be interpreted as follows: one weekly hour of instructional time given by the teacher at the 10th percentile of the MPI distribution increases student test scores by 2.1% of a standard deviation, which is 50% less productive than one hour taught by the teacher at the median of the MPI distribution.

Table A10 Robustness check - Different score for Teaching Practice variables

	Score (1)	Score (2)	Score (3)	Score (4)	Score (5)
Panel A: subtopics taught (N=18888)					
Instructional Time	0.050***	0.061***	0.040	0.004	0.033
	(0.010)	(0.009)	(0.051)	(0.053)	(0.062)
IT*Modern Practices 2	0.096***	0.112***	0.108***	0.107***	0.099***
	(0.035)	(0.032)	(0.032)	(0.031)	(0.036)
IT*Assessment		0.050**	0.049**	0.042*	0.032
		(0.021)	(0.021)	(0.021)	(0.021)
Panel B: subtopics not taught (N=22263)					
Instructional Time	-0.001	-0.001	0.014	0.004	0.028
	(0.009)	(0.009)	(0.041)	(0.044)	(0.053)
IT*Modern Practices 2	0.005	0.005	0.012	0.017	0.006
	(0.024)	(0.024)	(0.024)	(0.025)	(0.025)
IT*Assessment		-0.001	-0.000	-0.001	-0.013
		(0.018)	(0.019)	(0.019)	(0.020)
IT*Teacher demographics	.	.	√	√	√
IT*Class size	.	.	.	√	√
IT*Teacher behaviour	.	.	.	.	√

Note: This table replicates table 1.3, using an alternative definition of the Modern Practices Index (MPI), which is based on a different way of scoring teachers' answers to the questions related to teaching practices. To compute this alternative MPI, we assign the score 0.25 (instead of 0.1) to the answer "sometimes" for all teaching practices variables. All regressions include student and teacher fixed effects, as well as topic constants and the proportion of subtopics taught the year of the test. Controls included in columns (3) - (5) are similar to those described in table 1.3. Standards errors (in parentheses) are clustered at the teacher level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A11 Robustness check - Binary Teaching Practice variables

	Score (1)	Score (2)	Score (3)	Score (4)	Score (5)
Panel A: subtopics taught (N=18888)					
Instructional Time (IT)	0.052***	0.059***	0.029	-0.008	0.032
	(0.011)	(0.011)	(0.054)	(0.056)	(0.065)
IT*Modern Practices (binary)	0.060*	0.063*	0.057*	0.056*	0.050
	(0.034)	(0.034)	(0.033)	(0.033)	(0.034)
IT*Assessment (binary)		0.027	0.028	0.022	0.015
		(0.018)	(0.017)	(0.017)	(0.018)
Panel B: subtopics not taught (N=22263)					
Instructional Time (IT)	-0.001	-0.001	0.015	0.006	0.029
	(0.009)	(0.010)	(0.041)	(0.044)	(0.054)
IT*Modern Practices (binary)	0.002	0.002	0.007	0.009	0.002
	(0.020)	(0.020)	(0.021)	(0.022)	(0.021)
IT*Assessment (binary)		-0.001	-0.000	-0.000	-0.004
		(0.014)	(0.015)	(0.015)	(0.015)
IT*Teacher demographics	.	.	√	√	√
IT*Class size	.	.	.	√	√
IT*Teacher behaviour	.	.	.	.	√

Note: This table replicates table 1.3, using an alternative definition of the Modern Practices Index (MPI), which is based on a categorical definition of teaching practice variables. To compute this alternative MPI, we assign the score 1 to the answer "Every or almost every lesson" and 0 to all other answers. All regressions include student and teacher fixed effects, as well as topic constants and the proportion of subtopics taught the year of the test. Controls included in columns (3) - (5) are similar to those described in table 1.3. Standards errors (in parentheses) are clustered at the teacher level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A12 Pairwise correlation coefficients among Teaching Practices Indexes

	MPI	TPI_1	TPI_2	TPI_3
Modern Practices Index (MPI)	1			
Traditional Practices Index 1 (TPI_1)	-0.912***	1		
Traditional Practices Index 2 (TPI_2)	-0.650***	0.712***	1	
Traditional Practices Index 3 (TPI_3)	-0.612***	0.672***	-0.042	1

Note: this table exhibits pairwise correlation coefficients between the Modern Practices Index and the Traditional Practices Indexes. TPI_1 includes all the 6 traditional practices, whereas TPI_2 only include practices (a), (c) and (d) and TPI_3 only include practices (b), (e) and (f), respectively. * $p < 0.10$, ** $p < 0.05$ *** $p < 0.01$.

Table A13 Robustness check - Non centered Value of Teaching Practices

	Score (1)	Score (2)	Score (3)	Score (4)	Score (5)
Panel A: subtopics taught (N=18888)					
Instructional Time (IT)	0.039	0.027	-0.001	-0.044	-0.012
	(0.027)	(0.026)	(0.060)	(0.063)	(0.072)
IT*Traditional Practices Index'	-0.053*	-0.035	-0.031	-0.027	-0.028
	(0.030)	(0.030)	(0.031)	(0.031)	(0.031)
IT*Modern Practices Index'	0.077**	0.101***	0.104***	0.107***	0.100***
	(0.035)	(0.033)	(0.034)	(0.033)	(0.035)
IT*Assessment		0.050**	0.050**	0.044**	0.035*
		(0.022)	(0.021)	(0.021)	(0.021)
Panel B: subtopics not taught (N=22263)					
Instructional Time (IT)	0.001	0.001	0.022	0.011	0.046
	(0.023)	(0.023)	(0.042)	(0.047)	(0.059)
IT*Traditional Practices Index'	-0.004	-0.005	-0.017	-0.017	-0.021
	(0.024)	(0.024)	(0.026)	(0.026)	(0.026)
IT*Modern Practices Index'	0.002	0.002	0.001	0.005	-0.010
	(0.019)	(0.019)	(0.018)	(0.020)	(0.022)
IT*Assessment		-0.002	-0.004	-0.005	-0.020
		(0.019)	(0.020)	(0.020)	(0.022)
IT*Teacher demographics	.	.	√	√	√
IT*Class size	.	.	.	√	√
IT*Teacher behaviour	.	.	.	.	√

Note: This table replicates table 1.3, using two distinct teaching practices indexes, one Traditional and one Modern, computed over the non centered values of teaching practice variables. All regressions include student and teacher fixed effects, as well as topic constants and the proportion of subtopics taught the year of the test. Controls included in columns (3) - (5) are similar to those described in table 1.3. Standards errors (in parentheses) are clustered at the teacher level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A14 Robustness check - Including the diversity of Teaching Practices

	Score (1)	Score (2)	Score (3)	Score (4)	Score (5)
Panel A: subtopics taught (N=18888)					
Instructional Time (IT)	0.029	0.027	-0.001	-0.044	-0.012
	(0.027)	(0.026)	(0.060)	(0.063)	(0.072)
IT*Modern Practices Index	0.102***	0.124***	0.125***	0.125***	0.119***
	(0.037)	(0.034)	(0.034)	(0.033)	(0.037)
IT*Teaching Practices Diversity	0.004	0.006	0.007	0.007	0.007
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
IT*Assessment		0.056***	0.055***	0.049**	0.039*
		(0.021)	(0.021)	(0.021)	(0.021)
Panel B: subtopics not taught (N=22263)					
Instructional Time (IT)	0.001	0.001	0.022	0.011	0.046
	(0.022)	(0.023)	(0.042)	(0.047)	(0.059)
IT*Modern Practices Index	0.005	0.005	0.012	0.016	0.004
	(0.024)	(0.024)	(0.024)	(0.025)	(0.025)
IT*Teaching Practices Diversity	-0.000	-0.000	-0.001	-0.001	-0.003
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
IT*Assessment		-0.001	-0.001	-0.002	-0.016
		(0.018)	(0.019)	(0.019)	(0.021)
IT*Teacher demographics	.	.	√	√	√
IT*Class size	.	.	.	√	√
IT*Teacher behaviour	.	.	.	.	√

Note: This table replicates table 1.3, using an index of teaching practice diversity in addition to the main Modern Practices Index. The index of diversity equals the total score of the teacher on the 11 teaching practices. All regressions include student and teacher fixed effects, as well as topic constants and the proportion of subtopics taught the year of the test. Controls included in columns (3) - (5) are similar to those described in table 1.3. Standards errors (in parentheses) are clustered at the teacher level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A15 Robustness check - Two distinct Traditional Practices Indexes

	Score (1)	Score (2)	Score (3)	Score (4)	Score (5)
Panel A: subtopics taught (N=18888)					
Instructional Time (IT)	0.049***	0.052***	0.034	0.001	0.031
	(0.012)	(0.012)	(0.049)	(0.052)	(0.061)
IT*Traditional Practices Index 2 (TPI_2)	-0.067**	-0.062**	-0.060*	-0.063**	-0.052
	(0.032)	(0.031)	(0.031)	(0.030)	(0.033)
IT*Traditional Practices Index 3 (TPI_3)	-0.114***	-0.109***	-0.106***	-0.102***	-0.101**
	(0.038)	(0.039)	(0.040)	(0.039)	(0.041)
IT*Assessment		0.022	0.022	0.016	0.007
		(0.022)	(0.021)	(0.022)	(0.022)
Panel B: subtopics not taught (N=22263)					
Instructional Time (IT)	0.004	0.004	0.020	0.009	0.035
	(0.010)	(0.010)	(0.041)	(0.043)	(0.052)
IT*Traditional Practices Index 2 (TPI_2)	-0.023	-0.023	-0.023	-0.029	-0.017
	(0.029)	(0.029)	(0.029)	(0.032)	(0.029)
IT*Traditional Practices Index 3 (TPI_2)	0.020	0.020	0.008	0.007	0.010
	(0.030)	(0.029)	(0.027)	(0.027)	(0.027)
IT*Assessment		0.000	-0.002	-0.003	-0.013
		(0.019)	(0.019)	(0.019)	(0.020)
IT*Teacher demographics	.	.	√	√	√
IT*Class size	.	.	.	√	√
IT*Teacher behaviour	.	.	.	.	√

Note: This table replicates table 1.3, using two distinct Traditional Practices Indexes instead of one unique Modern Index. TPI_2 includes practices (a), (c) and (d) and TPI_2 includes practices (b), (e) and (f). All regressions include student and teacher fixed effects, as well as topic constants and the proportion of subtopics taught the year of the test. Controls included in columns (3) - (5) are similar to those described in table 1.3. Standards errors (in parentheses) are clustered at the teacher level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A16 Non linearity in the MPI effect

	Score (1)	Score (2)	Score (3)	Score (4)	Score (5)
Panel A: subtopics taught (N=18888)					
Instructional Time (IT)	0.027**	0.034***	0.017	-0.015	0.011
	(0.011)	(0.011)	(0.048)	(0.051)	(0.061)
IT*$MPI_{tophalf}$	0.032**	0.037***	0.035**	0.033**	0.027*
	(0.014)	(0.013)	(0.014)	(0.014)	(0.015)
IT*Assessment		0.044**	0.043**	0.037*	0.026
		(0.021)	(0.021)	(0.022)	(0.021)
Panel B: subtopics not taught (N=22263)					
Instructional Time (IT)	0.001	0.001	0.014	0.007	0.029
	(0.010)	(0.010)	(0.040)	(0.044)	(0.053)
IT*$MPI_{tophalf}$	-0.004	-0.004	-0.003	-0.002	-0.004
	(0.011)	(0.011)	(0.011)	(0.012)	(0.011)
IT*Assessment		-0.003	-0.003	-0.003	-0.015
		(0.018)	(0.019)	(0.019)	(0.020)
IT*Teacher demographics	.	.	√	√	√
IT*Class size	.	.	.	√	√
IT*Teacher behaviour	.	.	.	.	√

Note: This table shows the heterogeneity in the effect of math instructional time on student math performance according to the position of the teacher in the Modern Practices Index (MPI) distribution, separately on subtopics taught the year of the test (Panel A) and subtopics not taught the year of the test (Panel B). $MPI_{tophalf}$ is a dummy indicating whether the teacher ranks above the median of the MPI. All regressions include student and teacher fixed effects, as well as topic constants and the proportion of subtopics taught the year of the test. Controls included in columns (3) - (5) are similar to those described in table 1.3. Standards errors (in parentheses) are clustered at the teacher level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A17 Robustness check - main regressions without *Geometry*

	Score (1)	Score (2)	Score (3)	Score (4)	Score (5)
Panel A: subtopics taught (N=13899)					
Instructional Time (IT)	0.038***	0.047***	-0.016	-0.037	-0.004
	(0.010)	(0.010)	(0.039)	(0.046)	(0.078)
IT*Modern Practices Index	0.099***	0.112***	0.099***	0.102***	0.089**
	(0.036)	(0.036)	(0.035)	(0.035)	(0.039)
IT*Assessment		0.036	0.032	0.029	0.020
		(0.024)	(0.024)	(0.025)	(0.025)
Panel B: subtopics not taught (N=16711)					
Instructional Time (IT)	0.012	0.009	0.008	0.007	0.020
	(0.008)	(0.010)	(0.038)	(0.043)	(0.055)
IT*Modern Practices Index	0.014	0.013	0.017	0.018	0.013
	(0.028)	(0.028)	(0.027)	(0.028)	(0.029)
IT*Assessment		-0.012	-0.012	-0.012	-0.016
		(0.017)	(0.018)	(0.018)	(0.020)
IT*Teacher demographics	.	.	√	√	√
IT*Class size	.	.	.	√	√
IT*Teacher behaviour	.	.	.	.	√

Note: This table shows the heterogeneity in the effect of math instructional time on student math performance according to the teaching practices implemented by the math teacher, separately on subtopics taught the year of the test (Panel A) and subtopics not taught the year of the test (Panel B) and excluding *Geometry* subtopics. All regressions include student and teacher fixed effects, as well as topic constants and the proportion of subtopics taught the year of the test. Controls included in columns (3) - (5) are similar to those described in table 1.3. Standards errors (in parentheses) are clustered at the teacher level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A18 Instructional Productivity and Teaching Practices - Heterogeneity according to student gender

	Girls				Boys			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: topics taught this year (N_{girls}=9489; N_{boys}=9399)								
Instructional Time (IT)	0.065***	0.071***	0.014	0.026	0.032**	0.049***	-0.012	0.025
	(0.012)	(0.013)	(0.052)	(0.073)	(0.014)	(0.015)	(0.083)	(0.096)
IT*Modern Practices Index	0.118***	0.126***	0.112***	0.111**	0.068	0.094*	0.099*	0.087
	(0.040)	(0.040)	(0.042)	(0.048)	(0.057)	(0.055)	(0.057)	(0.062)
IT*Assessment		0.024	0.020	0.015		0.077**	0.067*	0.051
		(0.027)	(0.027)	(0.027)		(0.035)	(0.036)	(0.035)
Panel B: topics not taught this year (N_{girls}= 11459; N_{boys}=10804)								
Instructional Time (IT)	0.007	0.007	0.025	0.002	-0.009	-0.010	-0.024	0.052
	(0.011)	(0.012)	(0.070)	(0.077)	(0.014)	(0.016)	(0.067)	(0.102)
IT*Modern Practices Index	-0.008	-0.008	0.015	0.006	0.020	0.019	0.023	0.007
	(0.030)	(0.031)	(0.033)	(0.033)	(0.046)	(0.046)	(0.046)	(0.043)
IT*Assessment		0.003	0.003	-0.005		-0.004	-0.002	-0.021
		(0.022)	(0.022)	(0.023)		(0.032)	(0.033)	(0.036)
IT*Teacher demographics	.	.	√	√	.	.	√	√
IT*Class size	.	.	√	√	.	.	√	√
IT*Teacher behaviour	.	.	.	√	.	.	.	√

Note: This table shows the heterogeneity in the effect of math instructional time on student math performance according to the teaching practices implemented by the math teacher, separately on subtopics taught the year of the test (Panel A) and subtopics not taught the year of the test (Panel B), and separately for girls (columns (1)-(4)) and boys (columns (5)-(8)). All regressions include student and teacher fixed effects, as well as topic constants and the proportion of subtopics taught the year of the test. Controls included in columns (3) - (5) are similar to those described in table 1.3. Standards errors (in parentheses) are clustered at the teacher level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A19 The Modern Practices Index and Student Non Cognitive outcomes

	(1) Intrisic motivation	(2) Extrinsic motivation	(3) Self-confidence
Modern Practices Index	0.415*	0.423***	0.189
	(0.213)	(0.145)	(0.213)
Observations	7463	7459	7470

Note: This table shows the results of the regression of student non cognitive outcomes on the Modern Practices Index, controlling for student mean score in math subtopics not taught the year of the test, gender, age, socio-economic background, language spoken at home, math instructional time per week, school size, indexes of school immediate area's economic affluence and urban density and all teacher characteristics included in equation (1.3). Standard errors (in parentheses) are clustered at the teacher level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix B - Chapter 2

Main Tables

Table B1 9th grade math teacher evaluation and student performance

	End of middle school test scores			
	Math	Humanities	Math	Humanities
	(1)	(2)	(3)	(4)
Evaluation	0.045**	0.004		
	(0.014)	(0.014)		
Evaluation on t			0.041**	0.006
			(0.014)	(0.014)
Evaluation before t			0.053**	-0.003
			(0.018)	(0.018)
Observations	30414	30414	30414	30414

Note: The table refers to our working sample of math teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. Column (1) (column (2)) shows the result of regressing their students' average standardized score in math (humanities) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. Column (3) (column (4)) shows the result of regressing the same dependent variable on a dummy indicating that they underwent an external evaluation on t and on a dummy indicating that they underwent an evaluation between t_0 and $t-1$. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Table B2 9th grade math teacher evaluation and student performance - by subgroups

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Math score	0.045**	0.038*	0.052**	0.054**	0.039**	0.094**	0.031**
	(0.014)	(0.020)	(0.020)	(0.020)	(0.020)	(0.029)	(0.016)
Humanities score	0.004	-0.000	0.008	0.007	0.004	0.008	0.006
	(0.014)	(0.019)	(0.020)	(0.020)	(0.019)	(0.031)	(0.015)
Observations	30414	15724	14690	15072	15342	6818	23596

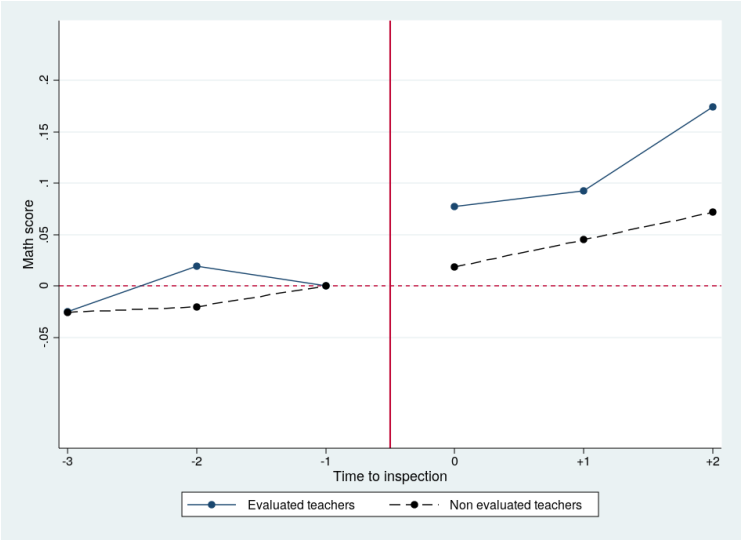
Note: The table refers to our working sample of math teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The first (second) row shows the results of regressing their students' average standardized score in math (humanities) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. The first column refers to the full sample, whereas columns (2) and (3) refer to the subsamples of female and male teachers, columns (4) and (5) to the subsamples of teachers whose number of years of work experience is either above or below the median on t_0 (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in education priority schools on t_0 and the subsample who were in non-priority schools. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Table B3 9th grade math teacher evaluation and student high school outcomes

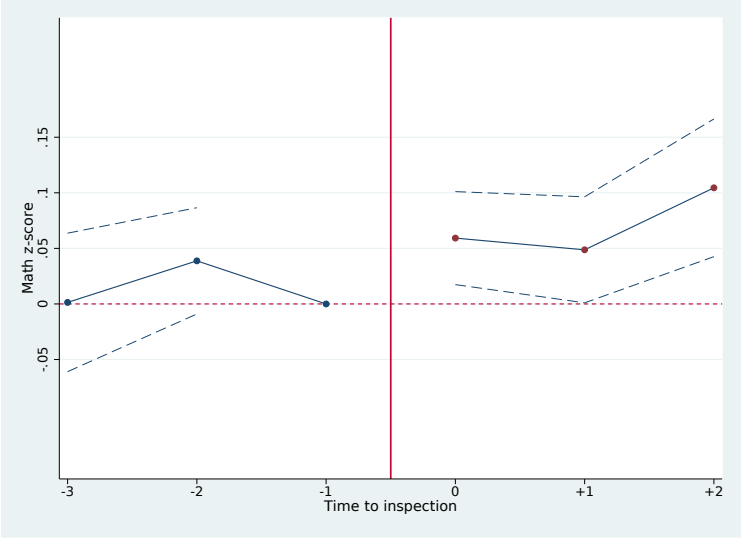
	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Science as major field	0.005**	0.003	0.007**	0.008**	0.002	0.010**	0.003
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.002)
	[0.176]	[0.183]	[0.169]	[0.161]	[0.191]	[0.123]	[0.192]
Graduation in Science	0.004**	0.002	0.007**	0.007**	0.003	0.009**	0.003
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
	[0.149]	[0.155]	[0.142]	[0.135]	[0.163]	[0.099]	[0.163]
Observations	30414	15724	14690	15072	15342	6818	23596

Note: The table refers to the working sample of math teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The first row shows the result of regressing the proportion of their 9th grade students who will choose science as major field of study at the end of 10th grade on a dummy indicating that they underwent an external evaluation between t_0 and t. The second row shows the result of regressing the proportion of their 9th grade students who will graduate in science at the end of 12th grade on the same independent variable. The first column refers to the full sample, whereas columns (2) to (7) refer to subsamples defined by teachers' gender, number of years of teaching experience on t_0 (above/below 11 years) and type of school attended on t_0 (priority/non priority). Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. Sample means of the dependent variables are within square brackets. * p<0.10, ** p<0.05.

Main Graphs



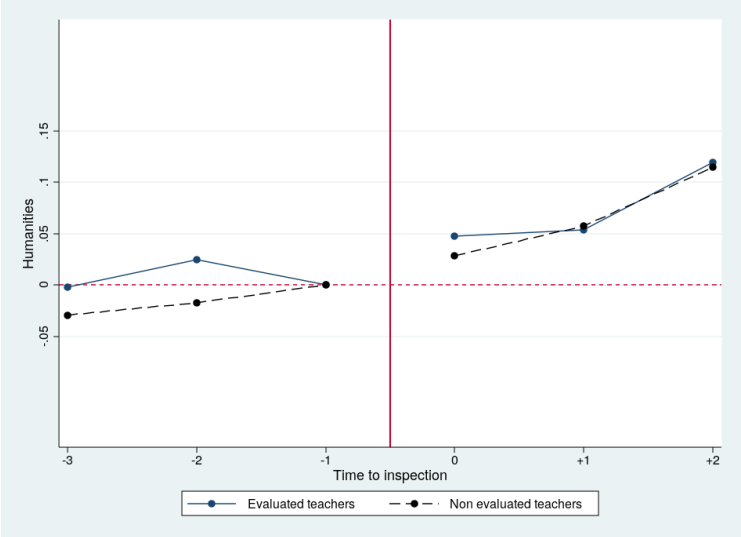
(a)



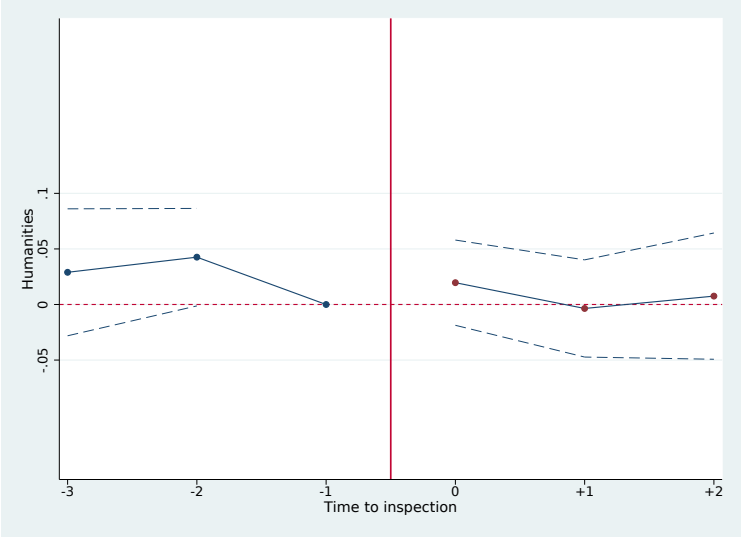
(b)

Figure B1 Math teacher evaluation and student performance in math

Note: The solid line in Figure B1 (a) shows math scores of students of evaluated math teachers before and after teachers' evaluations. The dotted line shows math scores of students of non-evaluated math teachers at exams taken on the same years. The solid line in Figure B1 (b) shows the difference in math scores between students of evaluated and non-evaluated math teachers before and after evaluations. The dotted lines show confidence intervals.



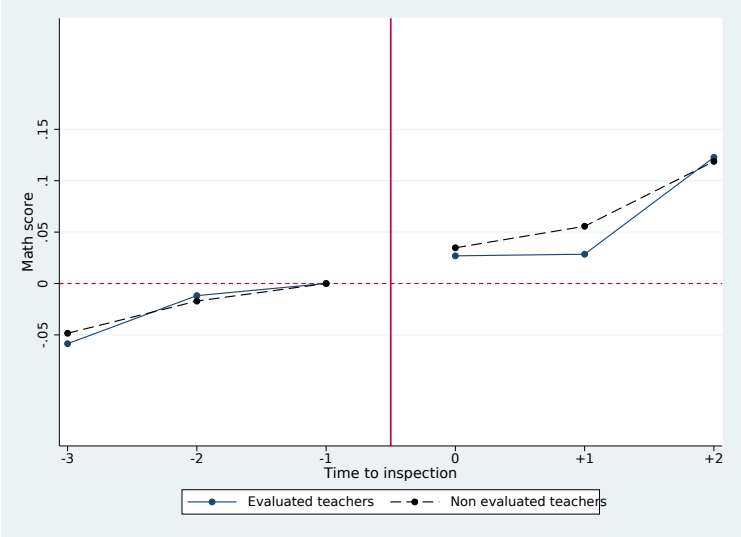
(a)



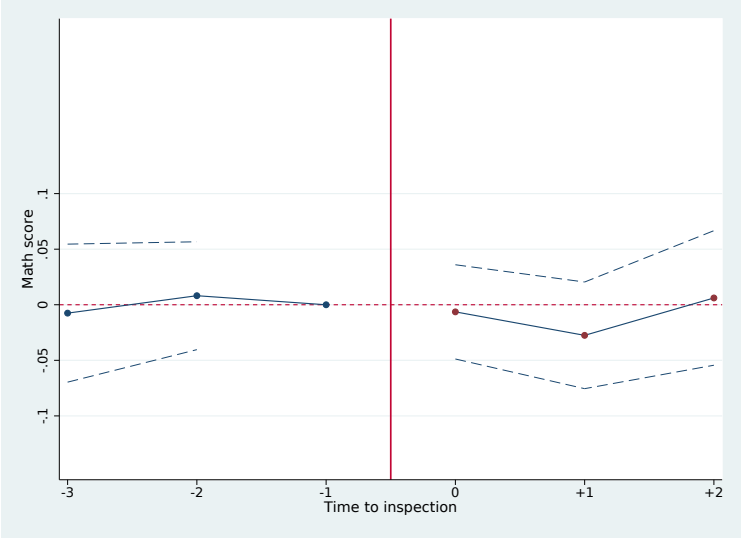
(b)

Figure B2 Math teacher evaluation and student performance in humanities

Note: The solid line in Figure B2 (a) shows humanities scores of students of evaluated math teachers before and after teachers' evaluations. The dotted line shows humanities scores of students of non-evaluated math teachers at exams taken on the same years. The solid line in Figure B2 (b) shows the difference in humanities scores between students of evaluated and non-evaluated math teachers before and after evaluations. The dotted lines show confidence intervals.



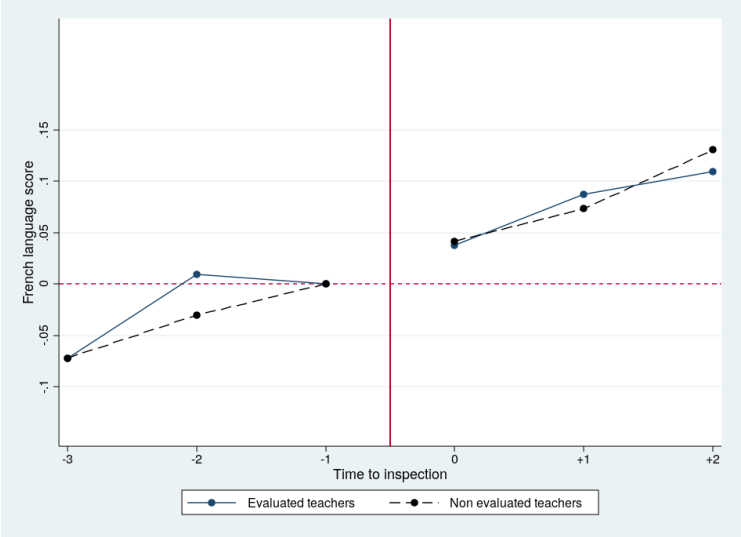
(a)



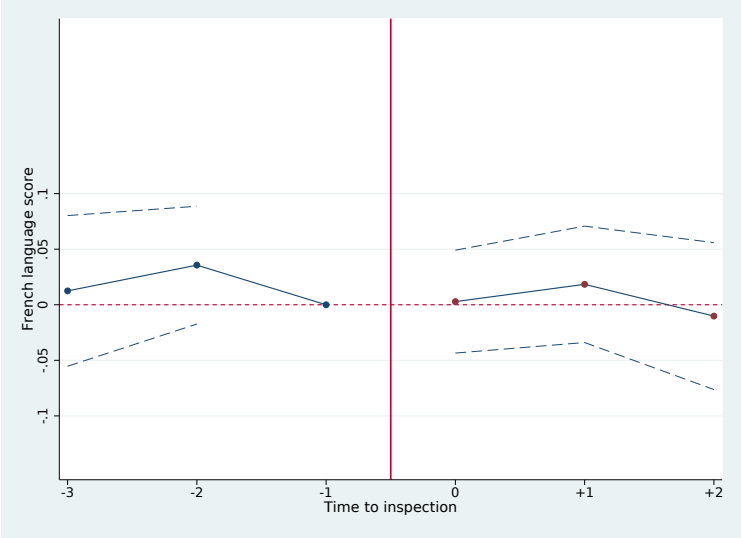
(b)

Figure B3 French language teacher evaluation and student performance in math

Note: The solid line in Figure B3 (a) shows math scores of students of evaluated French language teachers before and after teachers' evaluations. The dotted line shows math scores of students of non-evaluated French language teachers at exams taken on the same years. The solid line in Figure B3 (b) shows the difference in math scores between students of evaluated and non-evaluated French language teachers before and after evaluations. The dotted lines show confidence intervals.



(a)



(b)

Figure B4 French language teacher evaluation and student performance in French language

Note: The solid line in Figure B4 (a) shows French language scores of students of evaluated French language teachers before and after teachers' evaluations. The dotted line shows French language scores of students of non-evaluated French language teachers at exams taken on the same years. The solid line in Figure B4 (b) shows the difference in French language scores between students of evaluated and non-evaluated French language teachers before and after evaluations. The dotted lines show confidence intervals.

Additional Tables and Graphs

Table B4 *Inspecteurs*' characteristics

	(1) Math	(2) French language
Inspecteurs' individual characteristics		
Age	51.50	53.37
	(7.42)	(7.20)
Experience as *inspecteur*	6.31	7.11
	(3.94)	(4.36)
Female	0.33	0.58
	(0.47)	(0.49)
Total nb of *inspecteurs*	142	165
Regional characteristics		
Nb of *inspecteurs* per region	4.7	5.5
	(2.5)	(3)
Nb of teachers per region	1676	2208
	(1013)	(1326)
Nb of evaluations per region	252	310
	(143)	(163)
Total nb of regions	31	31

Note: The table refers to the population of *inspecteurs* working for the Ministry of Education during academic year 2008-2009. The upper part of the table shows their average age, number of years of experience and gender, separately for math *inspecteurs* (column (1)) and French language *inspecteurs* (column (2)). The lower part of the table shows the average number of *inspecteurs*, teachers, evaluations per region (separately for math and French Language). Standard deviations are in parentheses.

Table B5 Student characteristics - difference between priority and non priority schools

	Priority schools (1)	Non priority schools (2)	Difference (1) - (2)
Age	14.64	14.47	0.17**
	(0.24)	(0.18)	(0.01)
Female	0.51	0.51	-0.00
	(0.10)	(0.09)	(0.00)
Low-income	0.45	0.22	0.23**
	(0.20)	(0.14)	(0.01)
Average standardized test scores	-0.62	0.21	-0.83**
	(0.898)	(0.767)	(0.03)
Observations	1091	4144	5235

Note: The table shows the difference in students' average age as well as in the proportion of female students, low-income students and students' average scores at the end-of-middle school national exam, across priority and non-priority schools in 2008-2009. * p<0.10, ** p<0.05.

Table B6 Math teachers' evaluations and 9th grade teaching

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
	0.008	0.013	0.002	0.009	0.006	0.010	0.009
	(0.006)	(0.009)	(0.009)	(0.009)	(0.009)	(0.014)	(0.007)
	[0.78]	[0.78]	[0.79]	[0.76]	[0.80]	[0.76]	[0.79]
Observations	39958	20757	19201	20450	19508	9246	30712

Note: the table refers to the sample of math teachers who teach 9th grade students on year t_0=2008-2009 and who are not evaluated during t_0. The table shows the result of regressing a dummy indicating that teachers teach 9th grade students on year t on a dummy indicating that teachers underwent an external evaluation between t_0 and t. Column (2) refers to the subsample of female teachers, column (3) to male teachers, column (4) to teachers whose number of years of teaching experience is below the median (i.e. above or below 11 years) and column (5) to teachers above this median. Eventually, columns (6) and (7) refer to teachers who were in education priority schools in 2008 and to those who were in non-priority schools in 2008, respectively. Standard errors are in parentheses. Sample means of the dependent variables are within square brackets. * p<0.10, ** p<0.05.

Table B7 Teachers' characteristics

	(1) Math	(2) French language
Experience (in 2008)	12.28	12.74
	(5.11)	(5.01)
Female teacher	0.52	0.83
	(0.50)	(0.37)
Priority schools (in 2008)	0.17	0.18
	(0.37)	(0.38)
Number of evaluations (N_e)		
$N_e = 0$	0.43	0.54
	(0.49)	(0.50)
$N_e = 1$	0.56	0.45
	(0.50)	(0.50)
$N_e > 1$	0.01	0.01
	(0.09)	(0.08)
Observations	30414	30779

Note: The table refers to our working sample of teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The table shows the mean characteristics of teachers in terms of number of years of teaching experience in 2008, gender and type of school in 2008, as well as the number of external evaluations that teachers underwent over the 4-year period under consideration. The first column refers to the subsample of math teachers whereas the second column refers to the subsample of French language teachers.

Table B8 Balancing test - 9th grade math teacher evaluation and student characteristics

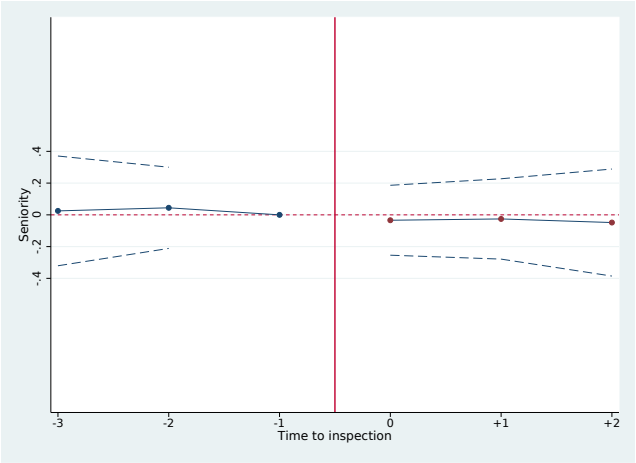
	(1) Age	(2) Female	(3) Low-income	(4) German	(5) Latin/Greek
All teachers (N=30414)					
Evaluation	0.004	-0.001	0.002	-0.000	0.002
	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)
Female teachers (N=15724)					
Evaluation	0.010*	-0.004	0.005	-0.004	0.004
	(0.006)	(0.004)	(0.004)	(0.005)	(0.005)
Male teachers (N=14690)					
Evaluation	-0.001	0.002	-0.001	0.004	0.001
	(0.007)	(0.004)	(0.004)	(0.005)	(0.005)
Low-experience teachers (N=15072)					
Evaluation	0.005	0.002	0.003	-0.003	0.001
	(0.007)	(0.004)	(0.004)	(0.005)	(0.005)
High-experience teachers (N=15342)					
Evaluation	0.002	-0.003	-0.000	0.003	0.003
	(0.006)	(0.004)	(0.004)	(0.005)	(0.005)
Priority schools (N=6818)					
Evaluation	0.010	-0.010*	0.008	-0.000	-0.005
	(0.010)	(0.006)	(0.007)	(0.008)	(0.007)
Non Priority schools (N=23596)					
Evaluation	0.004	0.002	-0.000	0.000	0.005
	(0.005)	(0.003)	(0.003)	(0.004)	(0.004)

Note: the table shows the results of regressing 9th grade classes' average characteristics (average age of students, proportion of girls, proportion from low-income families, proportion studying German and proportion studying Latin or ancient Greek) on a dummy indicating that their math teacher underwent an evaluation between t_0=2008-2009 and t. The first row refers to the full working sample, whereas rows 2 to 7 refer to subsamples defined by teachers' gender, by teachers' number of years of experience (above or below 11 years) or by type of school attended (priority vs non-priority). Standard errors are in parentheses. * p<0.10, ** p<0.05.

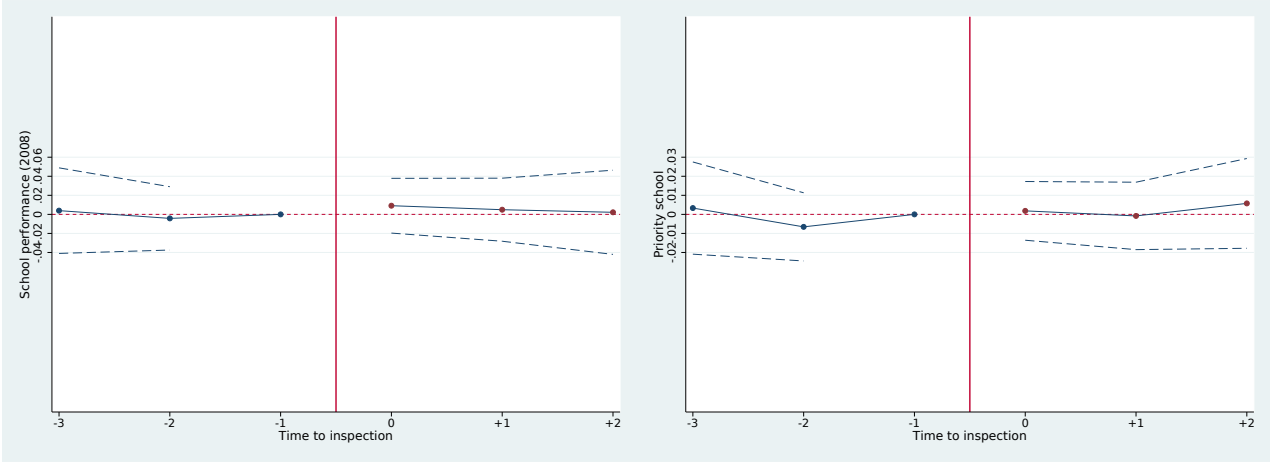
Table B9 Balancing test - 9th grade math teacher evaluation, teacher mobility and colleagues' characteristics

	(1) Teacher seniority	(2) Priority schools	(3) School performance	(4) Colleagues' experience	(5) Colleagues' seniority
All teachers (N=30414)					
Evaluation	0.035	0.004	-0.002	-0.039	-0.004
	(0.030)	(0.003)	(0.004)	(0.093)	(0.084)
Female teachers (N=15724)					
Evaluation	0.058	0.003	-0.002	-0.106	-0.109
	(0.042)	(0.003)	(0.006)	(0.131)	(0.118)
Male teachers (N=14690)					
Evaluation	-0.002	0.004	-0.000	0.042	0.120
	(0.043)	(0.004)	(0.006)	(0.133)	(0.121)
Low-exp (N=15072)					
Evaluation	0.071**	0.006	-0.007	-0.085	0.011
	(0.035)	(0.005)	(0.007)	(0.131)	(0.119)
High-exp (N=15342)					
Evaluation	-0.008	0.002	0.004	-0.007	-0.024
	(0.048)	(0.002)	(0.004)	(0.133)	(0.120)
Priority schools (N=6818)					
Evaluation	0.107	0.007	0.000	-0.048	0.179
	(0.081)	(0.009)	(0.014)	(0.193)	(0.172)
Non priority schools (N=23596)					
Evaluation	0.016	0.003*	-0.002	-0.015	-0.046
	(0.031)	(0.002)	(0.004)	(0.107)	(0.097)

Note: the table shows the results of regressing teacher seniority, school characteristics (priority school, school performance) and colleagues' characteristics (experience, seniority) on a dummy indicating that the math teacher underwent an evaluation between t_0=2008-2009 and t. School performance in column (3) is the average math test score in 2008 of the school in which the math teacher teaches in year t. Eventually, colleagues' experience and seniority in columns (4) and (5) refer to the average characteristics of the 9th grade French language and history teachers who teach the same 9th grade students as the math teacher in year t. Standard errors are in parentheses. * p<0.10, ** p<0.05.



(a) Seniority



(b) School performance

(c) Priority school

Figure B5 Math teacher evaluation and teacher mobility

Note: The solid lines in Figure B5 (a) to B5 (c) show the difference between evaluated and non-evaluated math teachers before and after evaluations in terms of teacher seniority (a), school performance as measured by the school average math test scores in 2008 (b) and teacher probability to teach in a priority school (c). The dotted lines show confidence intervals.

Table B10 9th grade French language teacher evaluation and student performance

	End of middle school test scores			
	French language	Math	French language	Math
	(1)	(2)	(3)	(4)
Evaluation	0.016	0.015		
	(0.016)	(0.015)		
Evaluation on t			0.006	0.015
			(0.016)	(0.016)
Evaluation before t			0.028	0.010
			(0.020)	(0.020)
Observations	30779	30779	30779	30779

Note: The table refers to our working sample of French language teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. Column (1) (column (2)) shows the result of regressing their students' average score in French language (mathematics) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. Column (3) (column (4)) shows the result of regressing the same dependent variable on a dummy indicating that they underwent an external evaluation on t and on a dummy indicating that they underwent an evaluation between t_0 and $t-1$. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Table B11 9th grade French language teacher evaluation and student performance - by subgroups

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	All	Female	Male	Low-exp	High-exp	Priority	Non Priority
French language score	0.016	0.023	-0.003	0.010	0.019	0.076**	-0.003
	(0.016)	(0.017)	(0.039)	(0.023)	(0.021)	(0.035)	(0.017)
Mathematics score	0.015	0.016	0.010	0.011	0.018	0.035	0.008
	(0.015)	(0.017)	(0.038)	(0.022)	(0.021)	(0.032)	(0.017)
Observations	30779	25601	5178	14135	16644	7027	23752

Note: The table refers to our working sample of French language teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The first (second) row shows the results of regressing their students' average score in French language (mathematics) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. The first column refers to the full sample, whereas columns (2) and (3) refer to the subsamples of female and male teachers, columns (4) and (5) to the subsamples of teachers whose number of years of work experience is either above or below the median on t_0 (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in education priority schools on t_0 and the subsample who were in non-priority schools. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Table B12 9th grade French language teacher evaluation and student performance by French language subtopic test scores and by subgroups

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Reading test scores	0.008	0.013	-0.009	-0.005	0.017	0.066*	-0.010
	(0.015)	(0.016)	(0.038)	(0.023)	(0.020)	(0.034)	(0.017)
Writing test scores	0.028	0.036*	0.005	0.038	0.017	0.077*	0.014
	(0.019)	(0.020)	(0.047)	(0.028)	(0.025)	(0.043)	(0.020)
Observations	30778	25600	5178	14135	16643	7027	23751

Note: The table refers to our working sample of French language teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The first (second) row shows the results of regressing their students' average score in reading (writing) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. The first column refers to the full sample, whereas columns (2) and (3) refer to the subsamples of female and male teachers, columns (4) and (5) to the subsamples of teachers whose number of years of work experience is either above or below the median on t_0 (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in education priority schools on t_0 and the subsample who were in non-priority schools. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Table B13 Math and French language teachers' evaluations and student performance - by subgroups

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Score in the subject	0.030**	0.028**	0.038**	0.031**	0.029**	0.083**	0.014
	(0.010)	(0.013)	(0.018)	(0.015)	(0.014)	(0.023)	(0.012)
Score in other subjects	0.008	0.007	0.010	0.009	0.008	0.015	0.007
	(0.010)	(0.013)	(0.017)	(0.015)	(0.014)	(0.022)	(0.011)
Observations	61187	41321	19866	29204	31983	13842	47345

Note: The table refers to joint sample of math and French language teachers who teach 9th grade students between t_0=2008-2009 and t_1=2011-2012. The first (second) row shows the results of regressing their students' average score in the subject they teach (subjects they don't teach) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0 and t. The first column refers to the full sample, whereas columns (2) and (3) refer to the subsamples of female and male teachers, columns (4) and (5) to the subsamples of teachers whose number of years of work experience is either above or below the median on t_0 (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in education priority schools on t_0 and the subsample who were in non-priority schools. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Standard errors are in parentheses. * p<0.10, ** p<0.05.

Table B14 Robustness checks - 9th grade math teacher evaluation and student performance - by subgroups

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Math score	0.043**	0.035*	0.053**	0.054**	0.036*	0.091**	0.030*
	(0.014)	(0.019)	(0.020)	(0.020)	(0.019)	(0.029)	(0.015)
Humanities score	0.005	-0.001	0.010	0.007	0.005	0.009	0.006
	(0.013)	(0.019)	(0.020)	(0.020)	(0.018)	(0.031)	(0.015)
Observations	32379	16906	15473	15072	17307	7029	25350

Note: The table refers to the same working sample of math teachers as Table 1, augmented by teachers with more than 25 years of teaching experience. The first (second) row shows the results of regressing their students' average score in math (humanities) at the end of year t on a dummy indicating that they underwent an external evaluation between t_0=2008-2009 and t. The first column refers to the full sample, whereas columns (2) and (3) refer to the subsamples of female and male teachers, columns (4) and (5) to the subsamples of teachers whose number of years of work experience is either above or below the median (i.e., above or below 11 years), columns (6) and (7) to the subsample of teachers who were in education priority schools on t_0 and the subsample who were in non-priority schools. Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. * p<0.10, ** p<0.05.

Table B15 Robustness check - 9th grade math teacher evaluation and student high school outcomes

	(1) All	(2) Female	(3) Male	(4) Low-exp	(5) High-exp	(6) Priority	(7) Non Priority
Science as major field	0.004**	0.002	0.007**	0.008**	0.001	0.009**	0.003
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.002)
	[0.179]	[0.187]	[0.170]	[0.161]	[0.194]	[0.124]	[0.194]
Graduation in science	0.004**	0.001	0.008**	0.007**	0.002	0.009**	0.002
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.002)
	[0.151]	[0.158]	[0.143]	[0.135]	[0.166]	[0.100]	[0.165]
Observations	32379	16906	15473	15072	17307	7029	25350

Note: The table refers to the same working sample of math teachers as Table 1, augmented by teachers with more than 25 years of teaching experience. The first row shows the result of regressing the proportion of their 9th grade students who will choose science as major field of study at the end of 10th grade on a dummy indicating that they underwent an external evaluation between t_0 and t. The second row shows the result of regressing the proportion of their 9th grade students who will graduate in science at the end of 12th grade on the same independent variable. The first column refers to the full sample, whereas columns (2) to (7) refer to subsamples defined by teachers' gender, number of years of teaching experience (above/below 11 years), type of school attended (priority/non priority). Models include a full set of teachers and year fixed effects as well as controls for students' average age, gender, family social background, German language study and Ancient language study. Sample means of the dependent variables are within square brackets. * p<0.10, ** p<0.05.

Data construction

This chapter uses an administrative database with detailed information on secondary school teachers for the period between t_0=2008-2009 to t_1=2011-2012. For each teacher j, this dataset gives information on whether (and when) j underwent an external evaluation between t_0 and t_1. It also gives information on whether (and when) teacher j taught 9th grade students and on the average performance of these students at exams taken at the end of 9th grade as well as at exams taken subsequently at end of high school. In this appendix, we explain how we build this database.

To construct this working file, we use three exhaustive administrative databases. The first one is the *Fichier Anonymisé d'Élèves pour la Recherche et les Études* (herafter, FAERE). For each academic year, it provides information on all secondary school students, including their socio-demographic characteristics, their ID number, the ID number of their class, their choice of field of study at the end of 10th grade as well as their results at the (externally set and marked) national exams taken at the end of middle school (9th grade) or at the end of high-school (12th grade). The exam taken at the end of middle school involves three written tests (in math, French language and history-geography) and we know students' scores at these different tests. We also know whether students choose science as major field of study at the end of 10th grade and whether they graduated in science at the end of 12th grade.

Using this individual level database, it is possible to build a class level database providing for each 9th grade class observed between 2008-2009 and 2011-2012 (a) the ID of the class and the academic year when the class is observed, (b) the average scores of the students of the class in math and humanities at exams taken at the end of the academic year (i.e. at the end of 9th grade), (c) the proportion of students of the class who will subsequently choose science as major field of study at the end of 10th grade (d) the proportion of students who subsequently succeed in graduating in science at the end of 12th grade.

The second database is an administrative dataset - called base *Relais* - which provides for each class observed between 2008-2009 and 2011-2012 the ID number of the class and the ID number of its teachers. This dataset makes it possible to augment our class-level database with information on the IDs of the math and French language teachers of each 9th grade class.

Eventually, we used the *Annuaire du Personnel du Secondaire Public* (herafter APSP). For each academic year, it provides information on the background characteristics of all teachers from public secondary schools (ID number, age, gender, level of experience, qualifications). For each teacher j and each academic year t, we also know whether j is evaluated during t.

This dataset makes it possible to augment the class level database with information on math and French language teachers, and most notably with information on whether (and when) they underwent an external evaluation between 2008-2009 and 2011-2012[19].

Overall, we get a class level database covering the period from 2008-2009 to 2011-2012 and providing for each 9th grade class observed during this 4-year period (a) the ID number of the class and the academic year when it is observed, (b) the ID number and socio-demographic characteristics of its math and French language teachers, (c) the date of the external evaluations that its math and French language teachers underwent during this 4-year period and (d) the average outcomes of its students at the end of 9th grade as well as their subsequent outcomes at the end of 10th grade or 12th grade.

Eventually, by averaging the variables of this database at the teacher x year level, we build a database which makes it possible to explore the extent to which teachers' external evaluations are followed by an improvement in their effectiveness, as measured by their ability to prepare 9th grade students for the end-of-middle school exams or by their ability to induce 9th grade student to choose science as major field of study in high school and to graduate in science.

[19]For each education region r and each academic year t, the APSP also provide background information on *inspecteurs* assigned to region r during t, namely information on their age, gender, level of experience as well as on their previous position within the French administration. Note, however, that we have no information on the specific teachers that were evaluated by each specific *inspecteurs*. It is not possible to match specific teacher's evaluations with specific inspecteurs.

Appendix C - Chapter 3

Tables

Table C1 Student characteristics

	Girls (1)	Boys (2)
Low-income family	0.28	0.25
	(0.22)	(0.21)
Foreign nationality	0.04	0.03
	(0.08)	(0.08)
Educational delay in 9th grade	0.23	0.27
	(0.19)	(0.20)
Observations	82184	82184

Note: this table shows the mean proportions of financial aid recipients (low-income family), foreign students and students with one year or more of educational delay in 9th grade in our sample, averaged at the class level. The first column refers to female students whereas the second column refers to male students.

Table C2 Classroom and teacher mean characteristics

	(1)
Classroom characteristics	
Class size	23.21
	(4.87)
Proportion of girls	0.50
	(0.11)
Teacher characteristics	
Experience	12.58
	(5.17)
Seniority	7.27
	(4.35)
Female teacher	0.62
	(0.27)
Fixed-term contract	0.02
	(0.08)
Pedagogical grade	44.39
	(3.26)
Observations	82184

Note: this table shows the mean characteristics of classes and teachers in our sample, averaged at the class level. Standard deviations are in parentheses.

Table C3 Outcome sample means and standard deviations, by student gender

	Girls (1)	Boys (2)
I. 9th grade outcomes		
End-of-middle-school test score	-0.05 (0.49)	-0.20 (0.49)
Behaviour grade	17.2 (2)	15.98 (2.3)
II. Track choices after 9th grade		
High school (general tracks)	0.61 (0.24)	0.54 (0.24)
Vocational school	0.30 (0.22)	0.36 (0.21)
Dropout	0.09 (0.12)	0.10 (0.11)
III. High school graduation		
Academic tracks	0.38 (0.23)	0.30 (0.21)
Science	0.16 (0.15)	0.19 (0.17)
Economics/Social sciences	0.13 (0.12)	0.08 (0.10)
Literature	0.09 (0.10)	0.02 (0.05)
Technological track	0.15 (0.13)	0.14 (0.13)
Observations	82184	82184

Note: this table shows the sample means and standard deviations of all student outcomes under consideration in this paper, averaged at the class level. The upper part shows the mean of student standardized test scores at the end of middle school national examination and behaviour grades. The middle part of the table shows the average proportions of students who (1) attend a general track in high school (2) attend a vocational school and (3) drop out of education, within the next 3 years following 9th grade. The lower part of the table shows the average proportions of students who graduate from high school within the next 5 years following 9th grade, by tracks. The first column refers to female students whereas the second column refers to male students. Standard deviations are in parentheses.

Table C4 Balancing test - student characteristics

	Class level (IV)		School level	
	(1)	(2)	(3)	(4)
	Girls	Boys	Girls	Boys
Low-income student	0.005	-0.000	0.005	-0.000
	(0.014)	(0.013)	(0.013)	(0.012)
Non-French student	-0.006	-0.007	-0.006	-0.007
	(0.006)	(0.006)	(0.006)	(0.006)
Educational delay	-0.015	-0.022	-0.014	-0.022
	(0.013)	(0.014)	(0.013)	(0.013)
F-stat	17848	17848	.	.
Observations	82184	82184	82184	82184

Note: the table shows the results of regressing 9th grade classes' average characteristics (average proportions of students from low-income families, foreign students and students with one year or more of educational delay) on the instrumented proportion of girls in the classroom (columns (1) and (2)) and on the proportion of girls in the school cohort (columns (3) and (4)), using our main specification. All regressions include school and cohort fixed effects. Standard errors (in parentheses) are clustered at the school level. * p<0.10, ** p<0.05.

Table C5 Balancing test - teacher and classroom characteristics

	(1)	(2)
	Class level (IV)	School level
Experience	0.292	0.284
	(0.327)	(0.317)
Seniority	0.157	0.153
	(0.297)	(0.288)
Advanced Certification	0.009	0.008
	(0.009)	(0.009)
Non-permanent	0.002	0.002
	(0.007)	(0.007)
Female teacher	-0.041**	-0.040**
	(0.017)	(0.017)
Class size	0.226	0.219
	(0.585)	(0.568)
F-stat	17848	.
Observations	82184	82184

Note: the table shows the results of regressing 9th grade teachers' average characteristics (experience, seniority, a dummy indicating advanced certification level, a dummy indicating non-permanent teachers and gender) and class size on the instrumented proportion of girls in the classroom (column (1)) and on the proportion of girls in the school cohort (column (2)), using our main specification. All regressions include school and cohort fixed effects. Standard errors (in parentheses) are clustered at the school level. * p<0.10, ** p<0.05.

Table C6 The effect of female peers on student outcomes - controlling for lead and lag

	Class level (IV)		School level	
	(1)	(2)	(3)	(4)
	Girls	Boys	Girls	Boys
I. 9th grade outcomes				
End-of-middle-school test score	0.105**	-0.046	0.102**	-0.045
	(0.032)	(0.034)	(0.032)	(0.033)
Behaviour grade	0.615**	0.205	0.598**	0.199
	(0.166)	(0.191)	(0.162)	(0.186)
II. Track choices after 9th grade				
High school (general track)	0.054**	-0.035**	0.053**	-0.034**
	(0.016)	(0.016)	(0.016)	(0.016)
	[0.622]	[0.534]	[0.622]	[0.534]
Vocational school	-0.039**	0.036**	-0.038**	0.035**
	(0.014)	(0.015)	(0.014)	(0.014)
	[0.297]	[0.348]	[0.297]	[0.348]
Dropout	-0.022**	0.004	-0.021**	0.004
	(0.011)	(0.010)	(0.010)	(0.010)
	[0.104]	[0.095]	[0.104]	[0.095]
III. High school graduation				
Academic tracks (S, ES, L)	0.058**	-0.025*	0.057**	-0.025*
	(0.014)	(0.013)	(0.014)	(0.013)
	[0.384]	[0.295]	[0.384]	[0.295]
Science (S)	0.035**	-0.007	0.034**	-0.007
	(0.010)	(0.011)	(0.010)	(0.011)
	[0.162]	[0.190]	[0.162]	[0.190]
Economics/Social sciences (ES)	0.021**	-0.013*	0.021**	-0.013*
	(0.010)	(0.008)	(0.009)	(0.007)
	[0.132]	[0.081]	[0.132]	[0.081]
Literature (L)	0.002	-0.004	0.002	-0.004
	(0.008)	(0.005)	(0.008)	(0.004)
	[0.090]	[0.024]	[0.090]	[0.024]
Technological track	0.003	-0.006	0.002	-0.006
	(0.011)	(0.011)	(0.011)	(0.010)
	[0.152]	[0.144]	[0.152]	[0.144]
F-stats	4710	4710	.	.
Observations	81643	81643	81643	81643

Note: this table shows the effect of the proportion of girls among school peers on various student outcomes, using model (3.1) (columns (1) and (2)), and model (3.2) (columns (3) and (4)) augmented with the lead and the lag of the proportion of girls in the school cohort, estimated separately for girls and boys. The upper part of the table shows the effects of female peers on students' standardized test scores at the end of middle school national examination and behaviour grade. The middle part of the table shows the effects on the proportion of students who (1) attend a general high-school (2) attend a vocational school and (3) drop out of education, within the next 3 years following 9th grade. The lower part of the table shows the effects on the proportion of students who graduate from high school within the next 5 years following 9th grade, separately by tracks. Outcome sample means are within square brackets. Standard errors (in parentheses) are clustered at the school level. * p<0.10, ** p<0.05

Table C7 Non linear effects of female peers on student outcomes

	Girls minority		Girls majority	
	(1) Girls	(2) Boys	(3) Girls	(4) Boys
I. 9th grade outcomes				
End-of-middle-school test score	0.014	-0.132**	0.171**	-0.020
	(0.047)	(0.047)	(0.040)	(0.046)
Behaviour grade	0.078	-0.369	0.859**	0.433*
	(0.237)	(0.257)	(0.206)	(0.240)
II. Track choices after 9th grade				
High school (general track)	0.029	-0.044**	0.074**	-0.037*
	(0.023)	(0.022)	(0.020)	(0.021)
Vocational school	-0.036*	0.027	-0.026	0.050**
	(0.021)	(0.021)	(0.018)	(0.020)
Dropout	0.007	0.007	-0.042**	0.008
	(0.015)	(0.013)	(0.014)	(0.013)
III. High school graduation				
Academic tracks (S, ES, L)	0.021	-0.046**	0.080**	-0.016
	(0.021)	(0.018)	(0.018)	(0.018)
Science (S)	0.010	-0.025*	0.051**	-0.004
	(0.015)	(0.015)	(0.013)	(0.015)
Economics/Social sciences (ES)	0.021	-0.016	0.020	-0.008
	(0.013)	(0.010)	(0.012)	(0.010)
Literature (L)	-0.010	-0.005	0.009	-0.004
	(0.012)	(0.006)	(0.011)	(0.007)
Technological track	-0.001	0.013	0.010	-0.025*
	(0.016)	(0.014)	(0.014)	(0.014)
F-stats	2456	2456	2838	2838
Observations	41102	41102	41082	41082

Note: this table shows the effect of the proportion of girls among school peers on various student outcomes, using model (3.1), estimated separately for schools where there is a minority of girls over the period under study (columns (1) and (2)) and schools where there is a majority of girls (columns (3) and (4)). The upper part of the table shows the effects of female peers on students' standardized test scores at the end of middle school national examination and behaviour grade. The middle part of the table shows the effects on the proportion of students who (1) attend a general high-school (2) attend a vocational school and (3) drop out of education, within the next 3 years following 9th grade. The lower part of the table shows the effects on the proportion of students who graduate from high school within the next 5 years following 9th grade, separately by tracks. Outcome sample means are within square brackets. Standard errors (in parentheses) are clustered at the school level. * $p<0.10$, ** $p<0.05$

Table C8 Female peers and teacher grading biases

	Class level (IV) (1)	School level (2)
Proportion of girls	0.044**	0.042**
	(0.009)	(0.008)
Observations	82184	82184

Note: This table shows the relationship between the proportion of girls among school peers and teacher grading biases in favor of girls, as measured by the class average difference between girls and boys in the difference between (standardized) non-blind and blind test scores, using model (3.1) (column (1)) and model (3.2) (column (2)). Standard errors (in parentheses) are clustered at the school level. * p<0.10, ** p<0.05, *** p<0.01.

Figures

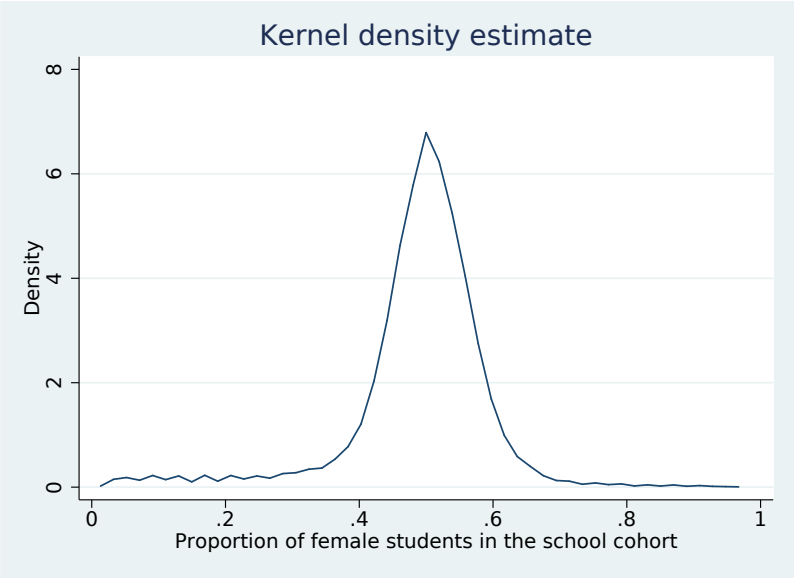


Figure A1 Distribution of the proportion of girls in a school cohort over the 2008-2011 period

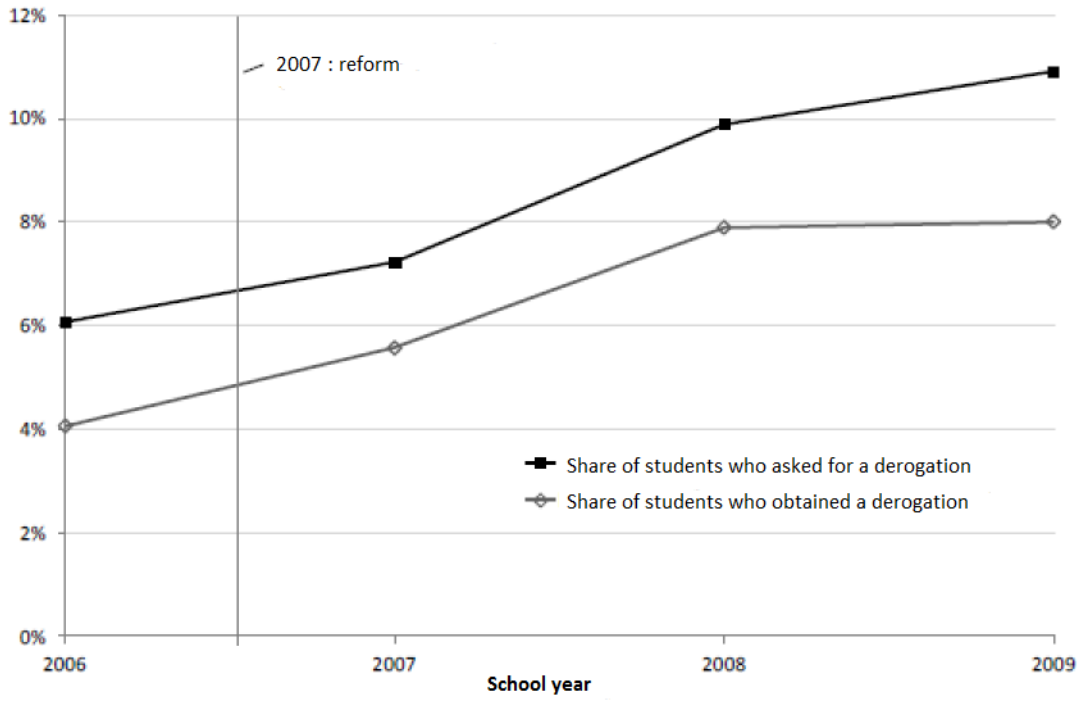


Figure A2 Proportion of students who asked for and obtained a derogation
Source: Fack & Grenet (2012)