Dispositif d'observation et d'évaluation « CP Dédoublés » : premiers résultats

Linda Ben Ali, Laurent Blouet, Pascal Bressoux, Axelle Charpentier, Isabelle Cioldi, Marianne Fabre, Laurent Lima, Fabrice Murat, Danaé Odin-Steiner, Christelle Raffaëlli, Thierry Rocher, Ronan Vourc'h

Commentaires : Denis Fougère (CNRS, Sciences Po)

Séminaire « Politiques éducatives », LIEPP (16/04/2019)

Le constat international

- ***Education Endowment Foundation, 2018 (une des principales institutions privées finançant des recherches sur l'éducation au Royaume-Uni):***
  - *"Intuitively, reducing the number of pupils in a class should improve the quality of teaching and learning, for example by increasing the amount of high quality feedback or one to one attention learners receive"*
  - *"Small reductions in class size are unlikely to be cost-effective relative to other strategies"*
  - *"Reducing class sizes for younger children may provide longer term benefits"*
  - *"Smaller classes only impact upon learning if the reduced numbers allow teachers to teach differently"*
  - *"Reducing class sizes to a level where a significant benefit is likely is expensive"*
  - *Conclusion : "**Moderate impact for high cost, based on moderate evidence**"*

Méta-analyses (1)

"Across the meta-analyses, summaries of major initiatives, and newer studies, the average effect size is ***d = 0.13***. This typical effect size of about ***d = 0.10–0.20*** could be considered small especially in relation to many other possible interventions—and certainly not worth the billions of dollars that is required to reduce the number of children per classroom. The more important question, therefore, is "Why are the effect sizes from reducing class size so small?"

"It appears that the effects of reducing class size may be higher on teacher and student work-related conditions, which then may or may not translate into effects on student learning."

John Hattie (directeur du *Melbourne Education Research Institute* de l'Université de Melbourne). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement.* NY: Routledge, 2008.

Méta-analyses (2)

G. Whitehurst & M. Chingos (2011): "Class Size: What Research Says and What it Means for State Policy", Brookings Institution's Brown Center on Education Policy

"Assuming even the largest class-size effects, class-size mandates must still be considered in the context of ***alternative uses of tax dollars for education***. Will a dollar spent on class-size reduction generate as much return as a dollar spent on: raising teacher salaries, implementing better curriculum, strengthening early childhood programs, providing more frequent assessment results to teachers to help guide instruction, investments in educational technology, etc.?"

"Class-size reduction has been shown to work for some students in some grades in some states and countries, but its impact has been found to be mixed or not discernable in other settings and circumstances that seem similar. ***It is very expensive.*** The costs and benefits of class-size mandates need to be carefully weighed against all of the alternatives when difficult budget and program decisions must be made."

Des résultats contrastés

- ***Effets positifs :***
  - *Krueger (1999): expérimentation STAR, Tennessee, 1985, GS-CE2, 11000 élèves*
  - *Résultats contestés par Hanushek (EEPA, 1999): "the results show effects that are limited to very large (and expensive) reductions in kindergarten or possibly first grade class sizes"*
  - *Angrist & Lavy (QJE, 1999) sur les écoles israéliennes, résultats remis en question par les mêmes auteurs (AER, 2019, à paraître): "newer estimates show no evidence of class size effects"*
- ***Effets négligeables (ou ambigus)***
  - *Hoxby (QJE, 2000): Connecticut, "no relationship between class size and achievement in fourth and sixth grades"*
  - *Jepsen and Rivkin (JHR, 2009) : Californie, "increases in the numbers of new and not-fully-certified teachers offset much of the gains"*
  - *Chingos (EER, 2012): Floride, "mandated CSR in Florida had little, if any, effects on student achievement"*
  - *Han & Ryu (EER, 2017): Corée du Sud, "the effects of high school class sizes on test scores are small with tight confidence intervals"*
  - *Argaw & Puhani (EER, 2018): Hesse, "mostly insignificant effect of class size on higher school tracking"*

La méthode ici employée

- Mise en place d'un dispositif longitudinal d'observation à partir de septembre 2017 (entrée en CP) jusqu'en juin 2020 (fin de CE2)
- Echantillon : 204 écoles REP+, 102 écoles REP (comparables du point de vue de 3 caractéristiques), 102 écoles hors REP, soit 15000 élèves de CP
- ***Principale difficulté*** : « pas possible de savoir quelle école REP+ mettra en œuvre le dispositif « CP dédoublé » ou le dispositif PMQC à la rentrée 2017 » (page 22)
- Pourquoi ne rien dire des classes PMQC (1100 élèves) ? Sont-elles exclues ou non des échantillons analysés ? (oui, cf. page 8)
  - Cf. "The Effect of Teacher's Aides in the Classroom: Evidence from a Randomized Trial", Andersen *et alii, JEEA, déc. 2018: "The use of teacher's aides seems to be at least as efficient as class-size reductions if used as a universal instrument. It is a much more flexible intervention that can target specific groups of students for limited periods of time."*

- Restriction du champ d'analyse aux seules écoles comptant au moins 13 élèves en CP (p. 22). Pourquoi ne pas avoir retenu certaines des 7684 écoles ayant moins de 13 élèves en CP ?
- Il aurait été intéressant d'avoir plus d'informations sur la mise en œuvre concrète du dédoublement (par exemple, sur le décloisonnement et le regroupement des classes, p. 18) et de ses effets

Le groupe témoin

- En théorie, le groupe témoin aurait dû être composé de classes non « dédoublées » de CP en REP+, mais *pas possible*
- A la place, classes de CP en REP similaires du point de vue de 3 caractéristiques (proportion d'élèves défavorisés en CE2 dans l'école, taux de retard en CE2 dans l'école, revenu médian de l'IRIS)
- *Suggestion* : si possible, tenir compte des ***valeurs passées de ces trois variables*** (disons, durant les trois ou quatre années passées pour améliorer l'appariement et satisfaire à la condition de tendance commune, cruciale pour la mise en œuvre des doubles différences)

Comment sont appariés les échantillons ?

- Par un score de propension = probabilité d'être dans une classe dédoublée de CP en REP+
- Les données se prêtent-elles à une analyse par appariement? Une condition cruciale (Smith et Todd, *Journal of Econometrics*, 2001) :
  - Les variables permettant de construire le score de propension doivent être suffisamment nombreuses et de qualité (peu de données manquantes, peu d'erreurs de mesure, etc.)
  - Ici **3 variables seulement** dont les observations sont parfois manquantes (pour imputation avec PSM, cf. par ex., Mitra & Reiter, *Statistical methods in medical research*, 2016)
- Pourquoi ne pas tenir compte de la taille des classes de CP du groupe témoin (REP) ? Celle-ci est probablement variable.

Comment améliorer la qualité de l'appariement ?

- Ajouter des covariables : caractéristiques de l'établissement (effectif, nombre de classes, ancienneté des enseignants, etc.) et de l'environnement (distribution des CSP, des âges, structure des ménages dans l'IRIS, notamment familles monoparentales, etc.)
- Accroître la dispersion des scores et le support commun de leurs distributions
- Comparer les méthodes d'appariement (noyau, inverse weighting, etc.)
- Valider les tests d'équilibrage de scores (« *balancing score* »)
- Reporter les estimations des coefficients du score et une mesure de la qualité globale du score (pseudo-R^2)

Utiliser la méthode des plus proches voisins à partir d'un critère de distance (par exemple, distance de Mahalanobis)

Les principaux résultats

- *Comparaison entre élèves des classes « CP dédoublés » en REP+ et élèves des classes de CP en REP*
- *Avec la méthode des **doubles différences**, l'écart en français est égal à 0,04 en français et à 0,11 en mathématiques (ajouter les niveaux de significativité): **élimination des éventuels effets fixes***
- *Les deux écarts augmentent (0,08 et 0.13, respectivement) et sont très significatifs lorsque l'on régresse le résultat en fin de CP sur le résultat à la rentrée en CP et des indicatrices de département*
- *Deux remarques :*
  1. *Régresser sur le score initial peut biaiser les résultats, car ce score initial peut dépendre d'un effet fixe « élève »*
  2. *Ces estimations ne tiennent pas compte des scores de propension qui permettent de comparer (d'apparier) les établissements cibles et témoins*

Endogénéité du résultat à la rentrée 2017

- *Supposons que le « vrai » modèle soit:*

$$y_{i,t_2} = \alpha + \beta \times \mathbf{1}(CP12_i = 1) + \gamma \times y_{i,t_1} + \eta_i + \varepsilon_{i,t_2}$$

$$y_{i,t_1} = \eta_i + \varepsilon_{i,t_1}$$

où y_{i,t_1} est le score initial (à l'entrée en CP) et η_i est un effet fixe, spécifique à l'élève, fonction de sa capacité scolaire et des variables omises pertinentes (environnement familial, niveau d'éducation des parents, etc.)

- *L'omission de cet effet fixe peut conduire à une **surestimation** des coefficients β et γ (cf. simulations dans l'annexe 2)*
- *D'où **ma préférence pour le modèle en doubles différences***
- *Comment sont calculés les effets sur les élèves initialement les plus faibles ? Un effet « théorique » dites-vous (p. 13)? Pourquoi ne pas les estimer (au moyen d'une stratification endogène par exemple) ?*

Krueger (1999) ne régresse pas sur les scores de l'année précédente (expérience STAR, 1st grade = CP)

Explanatory variable	OLS: actual class size			
	(1)	(2)	(3)	(4)
			B. First grade	
Small class	8.57	8.43	7.91	7.40
	(1.97)	(1.21)	(1.17)	(1.18)
Regular/aide class	3.44	2.22	2.23	1.78
	(2.05)	(1.00)	(0.98)	(0.98)
White/Asian (1 = yes)	—	—	6.97	6.97
			(1.18)	(1.19)
Girl (1 = yes)	—	—	3.80	3.85
			(.56)	(.56)
Free lunch (1 = yes)	—	—	−13.49	−13.61
			(.87)	(.87)
White teacher	—	—	—	−4.28
				(1.96)
Male teacher	—	—	—	11.82
				(3.33)
Teacher experience	—	—	—	.05
				(0.06)
Master's degree	—	—	—	.48
				(1.07)
School fixed effects	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30

Utiliser le score de propension pour estimer les effets sur les résultats de fin de CP (ou sur leur évolution au cours de l'année de CP)

Suggestion:

- ***1ère étape : estimer les scores de propension sur les établissements et les élèves (matching multi-niveaux)***
- ***2ème étape : utiliser les scores pour estimer non-paramétriquement l'effet moyen du dédoublement sur les résultats de fin d'année (ou leur évolution)***

$$\mathrm{ATET}_{km} = \frac{1}{N_T} \sum_{i \in \{T=1\}} \left(y_{1i} - \sum_{j \in \{T=0\}} \frac{K\left(\frac{p_j - p_i}{h_{nc}}\right)}{\sum_{k \in \{T=0\}} K\left(\frac{p_k - p_i}{h_{nc}}\right)} y_{0j} \right)$$

Enquête sur les pratiques d'enseignement

- ***Partie très intéressante***
- ***Résultats principaux :***
  - *Classes dédoublées plus favorables aux apprentissages scolaires (attention régulière, concentration, comportement, exécution de tâches)*
  - *Pratiques pédagogiques davantage orientées vers l'activation cognitive et vers la différenciation (pratiques individualisées)*
- *Mais « les écarts de pratiques entre les enseignants des classes de CP dédoublées en REP+ et les enseignants du groupe de contrôle en REP sont **modestes** » (page 20)*
- ***Chaîne causale:***

Dédoublement (coûteux) ⇒ Modification des pratiques d'enseignement ⇒ Amélioration des résultats scolaires (modestes)

- ***Il serait intéressant de savoir si les contextes (d'établissement, de quartier, etc.) favorisent ou non ces modifications de pratiques***

ANNEXE 1

Les effets des changements de classes au cours de l'expérience STAR

Exemple: Transitions entre Grade 1 (CP) et Grade 2 (CE1)

	Second grade			
First grade	Small	Regular	Reg/aide	All
Small	1435	23	24	1482
Regular	152	1498	202	1852
Aide	40	115	1560	1715
All	1627	1636	1786	5049

Subjects moved between treatment and control groups

Krueger (1999) reports reduced form results where he uses ***initial assignment and not current status*** as explanatory variable

---

In Kindergarten, OLS and reduced form estimates are the same because students remained in their initial class for at least one year

In 1st grade, OLS (column 1-4) and reduced form (columns 5-8) are different

Explanatory variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				B. First grade				
Small class	8.57	8.43	7.91	7.40	7.54	7.17	6.79	6.37
	(1.97)	(1.21)	(1.17)	(1.18)	(1.76)	(1.14)	(1.10)	(1.11)
Regular/aide class	3.44	2.22	2.23	1.78	1.92	1.69	1.64	1.48
	(2.05)	(1.00)	(0.98)	(0.98)	(1.12)	(0.80)	(0.76)	(0.76)
White/Asian (1 = yes)	—	—	6.97	6.97	—	—	6.86	6.85
			(1.18)	(1.19)			(1.18)	(1.18)
Girl (1 = yes)	—	—	3.80	3.85	—	—	3.76	3.82
			(.56)	(.56)			(.56)	(.56)
Free lunch (1 = yes)	—	—	−13.49	−13.61	—	—	−13.65	−13.77
			(.87)	(.87)			(.88)	(.87)
White teacher	—	—	—	−4.28	—	—	—	−4.40
				(1.96)				(1.97)
Male teacher	—	—	—	11.82	—	—	—	13.06
				(3.33)				(3.38)
Teacher experience	—	—	—	.05	—	—	—	.06
				(0.06)				(.06)
Master's degree	—	—	—	.48	—	—	—	.63
				(1.07)				(1.09)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30	.01	.23	.29	.30

ANNEXE 2

Les effets de l'endogénéité potentielle de la réussite au temps 1 (rentrée 2017):

Simulations avec Stata

```
clear all
set obs 10000
matrix P = (1, .7 \.7 , 1)
mat A = cholesky(P)
mat list A
A[2,2]
          c1         c2
r1         1          0
r2        .7  .71414284
```

```
gen c1= invnorm(uniform())
gen c2= invnorm(uniform())
gen y1 = c1
gen y2 = .7*c1 + .71414284*c2
corr y1 y2 (obs=10,000)
             |       y1       y2
-------------+------------------
          y1 |   1.0000
          y2 |   0.6958   1.0000
```

```
gen byte reduc=uniform()<=0.36
summarize reduc
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       reduc |     10,000       .3592    .4797899         0          1
```

```
gen presco= y1
gen u=rnormal(0,0.3)
gen postsco= 0.10 + 0.10 * reduc + 0.56 *presco + u
reg postsco reduc presco, robust
```

. reg postsco reduc presco, robust

Linear regression

Number of obs	= 10,000
F(2, 9997)	= 16310.27
Prob > F	= 0.0000
R-squared	= 0.7669
Root MSE	= .30642

postsco	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
reduc	.0964624	.0063701	15.14	0.000	.0839756	.1089491
presco	.5555459	.003085	180.08	0.000	.5494987	.5615931
_cons	.1016407	.0038409	26.46	0.000	.0941119	.1091696

```
gen postsco= 0.10 + 0.10 * reduc + 0.56 *presco + y2 + u
reg postsco reduc presco, robust
```

```
Linear regression                       Number of obs     =     10,000
                                        F(2, 9997)        =   12390.30
                                        Prob > F          =     0.0000
                                        R-squared         =     0.7201
                                        Root MSE          =     .7813
------------------------------------------------------------------------------
             |               Robust
     postsco |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       reduc |   .1224491   .0162571     7.53   0.000     .0905818    .1543164
      presco |   1.254987   .0079743   157.38   0.000     1.239356    1.270618
       _cons |   .0990037   .0097842    10.12   0.000     .0798248    .1181827
------------------------------------------------------------------------------
```

For simplicity assume that $y_{it} = \gamma\, y_{i,t-1} + \alpha_i + \varepsilon_{it}$. Then

$$E[y_{it} | y_{i,t-1}, \alpha_i] = \gamma\, y_{i,t-1} + \alpha_i$$

and

$$Corr[y_{it}, y_{i,t-1} | \alpha_i] = \gamma$$

However α_i is unknown and we actually observe

$$E[y_{it} | y_{i,t-1}] = \gamma\, y_{i,t-1} + E[\alpha_i | y_{i,t-1}]$$

and

$$Corr[y_{it}, y_{i,t-1}] \neq \gamma$$

Specifically,

$$Corr[y_{it}, y_{i,t-1}] = Corr[\gamma\, y_{i,t-1} + \alpha_i + \varepsilon_{it}, y_{i,t-1}]$$
$$= \gamma + Corr[\alpha_i, y_{i,t-1}]$$

since $Corr[\varepsilon_{it}, y_{i,t-1}] = 0$. After some algebra for the special case of random effects with ε_{it} iid $[0, \sigma_\varepsilon^2]$ and α_i iid $[0, \sigma_\alpha^2]$, we get

$$Corr[y_{it}, y_{i,t-1}] = \gamma + \frac{(1-\gamma)}{1 + (1-\gamma)\sigma_\varepsilon^2 / (1+\gamma)\sigma_\alpha^2}$$

This result makes it clear that there are two possible reasons for correlation between y_{it} and $y_{i,t-1}$

True state dependence occurs when correlation over time is due to the causal effect of $y_{i,t-1}$ on y_{it}

This dependence is relatively large if the individual effect $\alpha_i \approx 0$ as then Corr[y_{it} , $y_{i,t-1}$] $\approx \gamma$

More generally, this happens when σ_α^2 is very small relative to σ_ε^2

Correlation due to **unobserved heterogeneity** arises even if there is no causal effect of $y_{i,t-1}$ on y_{it} , so $\gamma = 0$, but nonetheless there is correlation since Corr[y_{it} , $y_{i,t-1}$] simplifies to $\sigma_\alpha^2 / (\sigma_\varepsilon^2 + \sigma_\alpha^2)$ if $\gamma = 0$