

Les effets de l'*accountability* sur les politiques d'éducation aux Etats-Unis¹.

Denis Meuret, Université de Bourgogne (IREDU), Institut Universitaire de France.

Le mot *accountability*, au sens large, évoque toute procédure qui manifeste qu'un individu ou une institution est comptable de son action vis-à-vis d'un autre individu ou d'une autre institution. En ce sens là, on peut trouver exprimée en France l'idée que l'Ecole en général, ou un établissement scolaire en particulier, n'est pas comptable de ses résultats vis-à-vis de ses élèves ou vis-à-vis de sa communauté, par exemple parce qu'un enseignant « se doit d'abord à sa discipline » (Debray, 1998). Il est rare de trouver quelqu'un qui émette une idée de ce genre aux Etats-Unis (Meuret, 2007).

Au sens étroit, ce mot désigne, aux États-Unis, des procédures de régulation de l'éducation dont des éléments précurseurs sont apparus au début des années 70, qui ont été systématisées dans les années quatre-vingt-dix dans plusieurs états (Floride, Kentucky, Texas, notamment), puis ont été étendues au pays entier par la loi No Child Left Behind (2002), loi dont l'administration Obama prépare actuellement la réautorisation par le congrès. Grossièrement, ces procédures, que l'on appelle en France «régulation par les résultats», comprennent trois volets : des standards (ce que les élèves doivent savoir, l'objectif), des épreuves qui mesurent ce qu'ils savent, et des incitations qui résultent de ce que les épreuves révèlent sur l'atteinte des objectifs. En ce sens là, l'*accountability* trouve aux Etats-Unis des opposants aussi résolus (McNeil, 2000) que la régulation par les résultats en trouve en France. Cette opposition peut être de principe, portée souvent par des disciples de John Dewey, qui critiquent le fait même que l'on prétende mesurer les effets de l'éducation. Dans le cadre d'une démocratie pluraliste, disent-ils, les éducateurs doivent respecter le caractère unique de chaque individu et cultiver ce qui fera la spécificité de sa contribution à la société, une exigence en effet deweyenne. Cela, poursuivent-ils, est contradictoire avec l'idée de classer les individus sur la même échelle, de leur appliquer la même norme. «*Norm referenced testing is anathema to the pluralist democrat because it ignores the « passionate inner meaning » (William James) that make democratic individuals unique. Norms have little value to pluralist democrats*» (Garrison and Neiman, 2003, p.26). Une intéressante discussion peut s'engager à ce sujet² mais ici nous accepterons l'idée que les épreuves des procédures d'*accountability* ont quelque valeur humaine et que la question de savoir si ces procédures

¹ Ce texte est une version actualisée d'une intervention au colloque "La sociologie de l'éducation et les recompositions de l'Etat à l'heure de la globalisation et de la construction européenne", INRP, Lyon, novembre 2010.

² Je me permets de renvoyer à D. Meuret, (2011), Education, démocratie, espérance, introduction à Dewey, J. Démocratie et Education, Armand Colin. Il va de soi que ce type d'argumentation pluraliste ne peut être utilisé pour défendre le statu quo en France, puisque ce dernier est tout sauf pluraliste.

améliorent ou non la performance des élèves à ces épreuves est donc importante pour juger de leur valeur. Toutefois, une sorte de discussion intermédiaire existe qui porte sur les effets de l'*accountability* sur l'enseignement (Sur quoi, sur quoi concentre-t-on les efforts ?) et sur les enseignants (leur moral, leur professionnalité), que nous aborderons aussi.

Cette contribution présente d'abord les procédures américaines d'*accountability* puis les résultats de recherches empiriques récentes sur leurs effets : effets sur l'enseignement, sur les enseignants, et enfin sur les performances des élèves.

Théorie et histoire de l'*accountability*

Théories

La régulation de l'éducation peut se faire par l'offre (par des procédures administratives destinées à assurer la qualité de l'éducation) ou par la demande (les clients choisissent leur école) ou encore par une combinaison des deux, comme c'est le cas en Angleterre actuellement. On peut considérer qu'il existe trois types de régulation par l'offre, lesquelles peuvent aussi se combiner : par les normes (programmes, horaires et inspections), par les projets (une unité soumet un plan au niveau supérieur, qui le valide ou non) ou par les résultats (le niveau supérieur fixe des objectifs, surveille leur atteinte et y réagit). L'*accountability* relève de la dernière catégorie.

On peut justifier l'*accountability* par une théorie proposée par l'économie des organisations, dite « de l'agence » ou encore « principal-agent » : L'école, comme l'entreprise, sont des organisations dont le responsable (le « Principal ») doit laisser à chaque « agent » une certaine autonomie, mais donc aussi la latitude de poursuivre d'autres objectifs que ceux de l'organisation ou de les poursuivre avec un enthousiasme mesuré, le problème pour l'organisation étant que l'agent bénéficie d'une asymétrie d'information qui lui permet de prétendre, même si ce n'est pas le cas, qu'il poursuit les objectifs de l'organisation avec la dernière énergie mais qu'il est victime de circonstances qui entravent son action. En outre, l'école, plus que l'entreprise, poursuit plusieurs objectifs et met en œuvre une technologie incertaine de sorte qu'il est possible de se réclamer d'un objectif quand des comptes sont demandés sur un autre, ou d'alléguer – en toute bonne foi ou non, peu importe ici - que l'on vise bien l'objectif en question, mais au moyen de détours technologiques (ex : redonner confiance à un élève par la pratique de la danse pour améliorer ses performances en maths). Cette asymétrie d'information existe, dans le domaine scolaire, entre l'élève et son enseignant, entre l'enseignant et son chef d'établissement, entre ce dernier et sa hiérarchie administrative. On voit que préciser les objectifs prioritaires, les formuler d'une façon telle qu'on puisse mesurer leur atteinte, isoler dans les performances ce qui peut être imputé aux circonstances et ce qui peut l'être aux actions des agents permet de diminuer l'asymétrie d'information et permet donc au « principal » une meilleure appréhension de la contribution des agents à l'efficacité de l'organisation et donc aussi de construire un système d'incitations (positives ou négatives) encourageant cette contribution. Bien sûr, dans un domaine comme

l'école, un tel système exige que les enseignants et le public en général adhèrent aux objectifs fixés au système éducatif.

Cependant, d'autres théories, proposées aussi par l'économie des organisations, suggèrent que peuvent exister des effets pervers qui vont limiter ou annuler l'efficacité du dispositif. Les deux principales sont les suivantes. La « loi de Campbell » stipule que « Plus un indicateur quantitatif est utilisé pour prendre des décisions, plus il sera sujet à des pressions corruptrices et plus il sera susceptible d'altérer et de corrompre le processus social qu'il est censé contrôler » (Campbell, 1976, cité par Krieg, 2011). Implicitement, en effet, les promoteurs de l'*accountability* supposent que ses objectifs seront poursuivis dans le respect de certaines limites fixées par les règlements et l'éthique professionnelle, mais la possibilité existe qu'ils le soient à leurs dépens. Par ailleurs, selon Holstrom and Milgrom (1991) « les dispositifs reposant sur des incitations amènent les agents à se concentrer sur les aspects les plus observables des tâches multidimensionnelles ». On verra que beaucoup des critiques américaines à l'*accountability* reposent, explicitement ou non, sur ces deux lois.

Les politiques américaines d'accountability, hier et aujourd'hui

Au début des années 70, les premières réformes de la gouvernance des systèmes scolaires américains ont consisté, comme en France, à proposer des indicateurs pour aider les établissements à analyser leurs performances, puis, dans les années 80, à accorder plus d'autonomie aux établissements (*school based management*). L'*accountability* – l'ajout d'incitations aux évaluations – est apparue au début des années 90 après les déceptions générées par ces politiques (Leightwood et Menzies, 1998). Les dispositifs qui en relèvent furent proposées par la technocratie ou par des hommes politiques locaux, à la différence par exemple des dispositifs de choix de l'école, qui furent d'abord proposés par des chercheurs, économistes (Friedman, 1962) ou politistes (Chubb and Moe, 1990). De 1990 à 2002, plus de la moitié des états américains (Winters et al., 2010) se dotent de procédures d'*accountability*, dites *soft* (l'état se contente de publier les performances des écoles, l'incitation se limite donc à l'effet sur la réputation de l'école) soit *hard* (avec des récompenses ou des sanctions directes).

No Child Left Behind (NCLB) est une loi bipartisane³, adoptée par le Congrès en tant que réautorisation de l'ESEA⁴, promulguée par G. Bush au début de 2002. Elle stipule qu'en 2014,

³ Ted Kennedy en fut le principal artisan du côté démocrate.

⁴ Dans les années 60, la loi qui a marqué le début de l'engagement de l'état fédéral dans le domaine de l'éducation, votée dans le cadre de la lutte contre la pauvreté engagée par L. Johnson, l'ESEA, (Elementary and Secondary Education Act), précisait que, tous les quatre ans, le congrès devrait en faire un bilan et proposer éventuellement des modifications. Cette réautorisation est parfois l'occasion de changements importants, comme en 1994, lorsque l'administration Clinton a changé une logique de «programmes», qui générait une action incohérente, chaque école se portant candidate au maximum de programmes possibles sans trop de souci de pertinence et de cohérence, en une logique d'aide aux établissements scolarisant une forte proportion d'élèves pauvres ou de minorités ethniques. NCLB marque un changement encore plus important, une ingérence sans précédent du niveau fédéral dans le domaine de l'éducation.

tous les élèves américains doivent avoir atteint le niveau *proficient* (compétent) en lecture et en maths. Il, s'agit là d'un objectif très exigeant⁵. A chaque grade scolaire (du 3eme au 8eme, soit du CE2 à la quatrième) correspond une définition précise de ce niveau. L'éducation étant aux USA du ressort des Etats, ce niveau est défini par chacun par rapport à des standards qui lui sont propres. Pour atteindre cet objectif, la loi prévoit un système d'incitations tel que l'attention soit dirigée vers les groupes ethniques et sociaux qui en sont le plus éloigné. Chaque Etat décide pour chaque école une sorte de chemin critique qui doit conduire à l'objectif. Par exemple, si, dans une école 40% des noirs, 30% des hispaniques, 25 % des pauvres et 60% des blancs atteignaient le niveau *proficient* en 2002, mettons au CM1, chacun de ces pourcentages devra chaque année augmenter assez pour que 100% des élèves de chacun de ces groupe soit (au moins) *proficient* en 2014. L'augmentation, dite AYP (*Adequate Yearly Progress*), devra donc être plus forte pour les noirs ou les pauvres que pour les blancs ou les riches. Si, dans une école, les élèves (soit l'ensemble des élèves, soit ceux d'un seul des groupes sociaux⁶) ne font pas les progrès « adéquats », l'école reçoit une aide. Mais, si c'est le cas pendant deux années de suite, elle est désignée comme *in need of improvement* (INI, qui a besoin de s'améliorer). Dans ce cas, l'école et le district doivent s'accorder sur un plan d'amélioration, un élément de régulation par projet dans cette régulation par les résultats. En outre, les parents reçoivent un « droit de fuite » : ils peuvent envoyer leurs enfants dans une autre école publique du district qui, elle, ne soit pas classée INI, l'école de départ devant les aider dans leurs démarches et le district payer le transport.

Si l'école échoue un an de plus, elle doit mettre gratuitement à la disposition de ses élèves des *Supplemental Educational Services* (SES), des cours supplémentaires offerts par elle-même, par des associations, ou même des entreprises privées, agréées à cet effet par le district. Si au bout de quatre ans, l'école échoue encore, elle doit remplacer une partie de son personnel, revoir son programme d'études, remplacer son principal et augmenter la durée journalière et annuelle de la scolarité. Au bout de cinq années d'échec, l'école est prise en charge directement par l'Etat ou est convertie en *Charter School*. Comme telle, elle est destinée à disparaître si elle n'améliore pas ses résultats⁷.

On aura noté que les incitations de NCLB ne jouent pas sur les ressources des écoles, mettant ainsi fin à un problème que les économistes de l'éducation se posaient dans les années 80. Fallait-il diminuer les ressources des écoles inefficaces, en vue de concentrer ces ressources là où elles étaient utilisées efficacement, ou les augmenter, afin d'aider ces écoles à affronter

⁵ Par exemple, Krieg (2011) indique qu'en 2011, dans l'état de Washington, en 2001-2002, 30% seulement des élèves atteignaient ce niveau en maths au grade 4.

⁶ NCLB distingue cinq groupes ethniques : noirs, hispaniques, blancs, indiens américains, asiatiques et trois autres groupes : pauvres, élèves dont la langue maternelle n'est pas celle de l'école, élèves en éducation spécialisée (indiqué par exemple in Krieg, 2011).

⁷ En 2010, selon le Center on Education Policy (2011), 38% des écoles du pays n'ont pas fait les AYP requis pour elles, parmi lesquelles certaines pas pour la première fois, qui sont donc considérées INI. Dans certains états, en particulier ceux qui ont des standards (trop ?) ambitieux, la proportion d'écoles INI est plus élevée, trop élevée pour que les districts puissent se soucier de chacune (en Californie, en 2008, 48% des écoles étaient dans ce cas (Mintrop and Sunderman, 2009). Dans d'autres, il est plus raisonnable : dans le Maryland, 15% environ des écoles sont l'objet de l'une ou l'autre des incitations négatives prévues par NCLB (calcul d'après Helmet, 2011).

leurs difficultés, mais au risque de créer une incitation à l'inefficacité ? NCLB résout ce problème en jouant, non sur les ressources des écoles, mais, comme d'ailleurs la réglementation anglaise, sur leur autonomie, en diminuant celle des écoles pas assez efficaces.

L'administration Obama, qui doit soumettre au Congrès une nouvelle ré-autorisation de l'ESEA, a indiqué les infléchissements qu'elle entend à cette occasion apporter à NCLB (USDE, 2010). La logique *standards/ accountability* est conservée, ainsi que la désagrégation des résultats aux tests par groupes sociaux ou ethniques. En revanche :

- Sous NCLB les Etats avaient intérêt à proposer des standards faibles pour augmenter leurs chances d'atteindre l'objectif de 100% *proficient*, même si cette tentation était atténuée par le fait que les évaluations nationales du NAEP (*National Assessment of Educational Progress*) permettaient de repérer les états à standards faibles en comparant les résultats de leurs élèves au NAEP et à leurs tests (Jacob, 2007). L'administration actuelle entend que les états se dotent de standards plus ambitieux, elle a lancé à cet effet le programme *Race to the Top*, un concours qui attribue des subventions fédérales aux états qui proposent des standards exigeants et des politiques crédibles pour y parvenir.
- NCLB comportait des incitations négatives, mais pas d'incitations positives. Cela lui a été reproché (Mintrop & Sunderman, 2008). Dans le nouveau système, les écoles qui améliorent leurs performances moyennes ou réduisent l'écart entre les scores des groupes sociaux seront « reconnues et récompensées ». Seront aussi récompensés les districts qui auront réussi à redresser la situation dans leurs écoles les plus en difficulté.
- Sous NCLB, une école qui améliorait son efficacité et sa proportion d'élèves *proficient* pouvait néanmoins être déclarée INI si ses performances n'augmentaient pas assez pour atteindre les AYP. Cela deviendra impossible parce que les écoles seront jugées, non sur la proportion de leurs élèves qui atteignent un seuil donné, mais en fonction des progrès réalisés par les élèves. Ce qui est prévu ici est l'usage d'indicateurs de « valeur ajoutée », soit le fait de savoir si, dans une école donnée, les élèves progressent plus qu'attendu compte tenu de leur niveau initial. Cela est certainement préférable d'un point de vue politique et éthique, mais un problème technique se pose : Parce que ces indicateurs mesurent l'écart entre deux données (le résultat observé et le résultat attendu), ils sont affectés d'une erreur de mesure importante ; si, en plus, l'indicateur est l'écart entre deux valeurs successives de la valeur ajoutée, le problème redouble et le nombre d'établissements dont on pourra dire de façon sûre que leurs élèves ont davantage (ou moins) progressés que l'année précédente risque d'être très faible.
- On l'a vu, la liste des incitations prévues était dans NCLB limitée et fixée au niveau fédéral. Dans la nouvelle législation, les états et les districts auront davantage de latitude pour inventer des incitations adaptées aux situations locales.
- Les états devront investir de façon spécifique et massive auprès des écoles faisant partie des 5% les plus systématiquement INI. Cette décision rejoint les résultats d'une des études qui sera présentée ci-dessous (Hemelt, 2011).

Il semble possible d'analyser cette réforme de NCLB de la façon suivante. L'Etat fédéral se recentre sur ses missions principales (veiller au niveau des standards) et laisse davantage aux états le détail des incitations à mettre en œuvre.

Les effets de l'*accountability* aux Etats-Unis

Ces effets sont l'objet d'une controverse assez vive. Les premières réactions, des chercheurs en éducation en particulier, furent d'une violence qui évoquait davantage les controverses françaises entre républicains et pédagogues que les controverses académiques. Mc Neil (2000) en est un bon exemple : L'*accountability* signifie, selon cet ouvrage écrit contre le dispositif d'*accountability* mis en place au Texas dans les années 90, que l'on renonce à tout enseignement authentique pour enseigner des compétences mécaniques, que les élèves faibles et pauvres seront sacrifiés sur l'autel de la performance chiffrée, etc. Cette opposition d'ordre culturel, dont on trouve un écho dans l'idée que la régulation par les résultats impose aux écoles une logique entrepreneuriale étrangère à leur nature⁸ (Lessard & Meirieu, 2005), a fait place plus récemment à un débat qui s'appuie davantage sur des données : quels effets les enquêtes, les évaluations, permettent-elles d'observer ?

Nous présentons ici les résultats de quelques unes de ces recherches, parmi celles qui présentent un niveau minimum de rigueur (nombre suffisant d'observations, utilisation d'une situation de référence et contrôle des biais potentiels dans la comparaison), mais sans prétendre à l'exhaustivité⁹, qui demanderait, vue l'ampleur de la littérature sur le sujet, un travail d'une toute autre ampleur.

Ces recherches abordent trois questions : Quels sont les effets des dispositifs d'*accountability* sur le contenu de l'enseignement (1), sur le travail et le moral des enseignants (2), sur la triche et, c'est le plus important et nous y consacrerons plus de place, sur les performances scolaires des élèves, performances moyennes d'une part (efficacité), performances des élèves les plus faibles ou des élèves défavorisés d'autre part (équité).

Contenu de l'enseignement

L'accord existe sur le fait que NCLB a vraiment modifié le quotidien des écoles (Gros & Goertz, 2005, Mintrop & Sunderman, 2009) mais les auteurs divergent sur la valeur de ces changements. Les reproches principaux peuvent être regroupés en quatre rubriques : Un curriculum rétréci, un enseignement stratégique, l'attention à certains élèves au détriment des autres, la triche. Considérons ce que disent les recherches sur chacun d'eux.

Un curriculum rétréci

⁸ « N'ayant d'autre projet que le pilotage par les résultats, l'institution, autrefois structurée autour de principes forts, se disloque » (Ph. Meirieu, in Lettre de l'éducation n°709, 27.6.2011). Sur ce qu'il faut penser des principes, voir Dewey, J. Le public et ses problèmes, trad. fcse, 2003, Publications de l'Université de Pau.

⁹ Ont été sélectionnées les recherches sur l'*accountability* américaine publiées dans la *Economics of Education Review* depuis 2003, auxquelles ont été ajoutées les recherches sur le même sujet publiées dans d'autres revues d'économie américaines et citées dans les premières, ainsi que les plus connus des articles publiés sur le sujet dans des revues américaines de sciences de l'éducation.

Le rétrécissement du curriculum a commencé aux Etats –Unis avec le « *back to basics* » des années 80, donc avant l'*accountability*, mais il est vrai que, en mesurant les performances des élèves dans certains domaines seulement, celle-ci peut amener les écoles à concentrer horaires et moyens sur ces derniers au détriment des autres. De fait, on observe bien que des ressources passent des matières non testées aux matières testées (Jacob, 2005, par exemple), mais les effets sur la performance des élèves dans les matières non testées est moins net. (Winters et al., 2010) trouvent que les progrès des élèves sont plus forts en Floride dans les écoles mises sous pression par l'*accountability* que dans les autres, ceci surtout dans les disciplines prises en compte par le dispositif, lecture et maths, mais aussi, bien que moins nettement, en sciences, une discipline non prise en compte¹⁰. Ils avancent deux hypothèses explicatives : de plus grands progrès en lecture et maths peuvent favoriser les progrès en sciences ; l'école peut favoriser l'apprentissage des maths et de la lecture par des politiques générales (lutte contre l'absentéisme et les retards, valorisation de la réussite en général) qui profitent aussi aux autres disciplines. Figlio & Rouse, 2006, à nouveau en Floride, observent le même pattern : l'amélioration des performances est plus grande dans les disciplines testées que dans les disciplines non testées mais elle existe aussi dans ces dernières. Jacob (2005) observe à Chicago que, après la mise en œuvre du dispositif, les performances des élèves augmentent deux à quatre fois plus dans les disciplines testées (maths et lecture) qu'en sciences et en *social studies*, ce qui signifie qu'elles s'améliorent aussi, bien que faiblement, dans ces dernières.

L'enseignement stratégique

L'enseignement stratégique vérifie la loi de Holmstrom et Milgrom. Il a deux aspects essentiellement : Les efforts peuvent être consacrés aux élèves ou groupes d'élèves qui menacent le plus le résultat de l'école, au détriment des autres élèves ; l'enseignement peut viser à faire réussir les élèves aux tests et non à leur faire acquérir des compétences plus complexes, importantes pour la suite de leurs études et pour leur vie. Ce dernier reproche (*teaching to the test*) vaut moins si les tests sont de bonne qualité, la question est celle de sa validité compte tenu de la qualité actuelle des tests.

Il semble bien que des efforts particuliers soient en effet déployés pour les élèves stratégiques, cependant pas tout à fait comme le craignaient les critiques. Pour ceux-ci, les écoles privilégieraient les *bubble kids*, les élèves situés près du seuil, au détriment des élèves les plus faibles. Springer (2008) observe au contraire que les effets de la mise d'une école en *INI* sont particulièrement positifs pour les élèves les plus faibles. D'autres études aussi témoignent d'une attention plus grande aux élèves faibles. A l'inverse, les résultats sont plus contrastés sur le point de savoir si l'attention aux élèves faibles organisée par NCLB se fait au détriment des élèves les plus forts. Springer (2008) observe que ce n'est pas le cas, mais Krieg (2011) propose une réponse plus nuancée. Il observe que, dans les écoles *INI*, les élèves d'un groupe social qui a fait les AYP dans une école qui globalement ne les a pas fait progressent moins que des élèves similaires d'une école qui, elle, a fait ses AYP. Autrement dit, une école mise

¹⁰Dans le cadre de NCLB, les élèves sont testés en maths et langage de puis 2003, mais aussi en sciences depuis 2007. L'étude de Winters et al. Porte sur le A+ Plan, où les élèves sont testés en maths et langage seulement.

sous pression à cause d'un ou plusieurs groupes déplace ses efforts vers ces groupes au détriment des autres groupes

Les enseignants enseignent-ils pour le test ? Il est vrai que certaines écoles réagissent à l'*accountability* par des séances de *tests preparation* (Gross & Goertz, 2005). J'ai entendu, d'adversaires américains de l'*accountability*, des descriptions apocalyptiques de telles séances. Sims, 2008 a observé que les districts ont parfois avancé le début de l'année scolaire pour avoir plus de temps d'enseignement avant les tests. Cependant, Gross & Goertz (2005) ont observé que les lycées réagissaient aussi par des politiques susceptibles de favoriser des progrès réels chez leurs élèves : *tutoring*, refonte du curriculum de l'école pour le rapprocher des standards de l'Etat, écoles d'été, programmes d'apprentissage de la lecture pour les élèves les plus faibles, implantation de cours nouveaux, soit plutôt basiques pour les élèves faibles, soit plutôt avancés pour les élèves forts. Une méta-analyse des études sur le *teaching to the test* (Au, 2007) observe comme Gross et Goertz des réponses fort différentes d'une école à l'autre. Elle conclut ainsi son analyse " Le premier effet des tests à fort enjeu est que le curriculum est rétréci aux sujets testés¹¹, que les domaines enseignés sont fragmentés en morceaux reliés aux tests et que les enseignants pratiquent une pédagogie frontale (*teacher centered*). Cependant, cette étude trouve aussi que, dans une minorité significative de cas, certains types de tests à fort enjeu ont conduit à un curriculum plus vaste, à un enseignement plus intégré (des différentes disciplines, DM), à un enseignement davantage centre sur l'élève et à une pédagogie plus coopérative».

La réalité du « *teaching to the test* » peut aussi être testée en observant si les élèves des écoles désignées comme en difficulté progressent dans les compétences les plus élémentaires (*low skills*) au détriment de leurs progrès dans les compétences plus élevées (*high skills*). Jacob (2005) observe à Chicago que les élèves de ces écoles progressent plus que ceux des autres écoles dans les *low skills* (fractions, par exemple) mais non dans les *high skills* (problèmes complexes) sans observer cependant de baisse de ces dernières, comme l'anticipaient les critiques (Mc Neil, 2000). Ceci dit, ils observent que les gains sont obtenus pour l'essentiel sur des habiletés spécifiques au test, ce qui appuie la thèse du *teaching to the test*. West et Peterson (2006) comparent en Floride les effets de l'*accountability* sur le test utilisé par le dispositif (le FCAT) et sur un test plus général. L'effet positif qu'ils observent sur le premier est nettement moins prononcé pour le second. Ils concluent cependant que le gain observé sur le premier « n'est pas obtenu aux dépens d'un apprentissage plus général ». De leur côté, Carnoy et Loeb (2002), une étude pré-NCLB, établit que le pourcentage d'élèves *proficient*, et non seulement le pourcentage d'élèves *basic*¹², augmente davantage dans les Etats ayant mis en place des dispositifs d'*accountability* que dans les autres, ce qui contredit, selon eux, la thèse que ces dispositifs favoriseraient uniquement l'acquisition de *low skills*.

¹¹ Une formulation un peu forte au vu des recherches présentées ici en ce qu'elle suggère que les écoles n'enseignent plus que les matières testées, ce qui est évidemment faux.

¹² La plupart des tests utilisés aux Etats-Unis distinguent trois niveaux de compétences (*basic*, *proficient*, *outstanding*), ce qui génère quatre catégories d'élèves : *below basic*, *basic*, *proficient*, *outstanding*.

Ces études confirment donc l'existence d'un enseignement stratégique au sens où une attention plus forte est souvent portée aux élèves en difficulté et davantage de ressources consacrées aux disciplines testées mais il ne semble pas que les autres élèves ou les autres disciplines en soient affectés significativement. Le schéma le plus probable semble être qu'au pire, on n'observe pas d'amélioration pour ces élèves ou ces disciplines par rapport à la situation pré-*accountability*. Il faut cependant noter celle de ces études qui porte sur les effets de NCLB et non des dispositifs précédents, (Krieg, 2011), diagnostique un effet négatif pour certains groupes d'élèves de la mise en INI.

La triche

La responsable des tests pour l'état de Floride m'a indiqué en 2004 que des cas de triche caractérisée se rencontraient en effet, que certains principaux avaient été sanctionnés pour cela et que des techniques statistiques (mise en évidence de progrès improbables d'une année sur l'autre, par exemple) permettaient de surveiller ce phénomène. Selon Jacob & Levitt (2003), 4 à 5 % des enseignants de Chicago, sous le régime d'*accountability* précédant NCLB, ont aidé leurs élèves à tricher lors du test, surtout dans écoles en difficulté. Hemelt (2011) remarque, en revanche, qu'il n'observe pas de concentration d'écoles juste au dessus des seuils, ce qui lui fait penser que les écoles ne manipulent pas leurs scores.

On peut, par ailleurs, rapprocher de la triche certains comportements stratégiques que Jacob (2005) a repéré dans les écoles de Chicago : faire redoubler à certains élèves faibles le *grade* précédant celui où sont passés les tests d'*accountability*¹³, augmenter la proportion d'élèves admis en éducation spécialisée (dispensés de participer aux tests dans le dispositif de Chicago). Il faut donc observer que l'*accountability* génère des formes spécifiques de tricherie, mais il convient aussi de se rappeler que les autres systèmes de gouvernement de l'école n'en sont pas exempts : certaines écoles américaines trichent quand elles sont jugées par les tests comme certains élèves ou étudiants français trichent quand ils sont jugés par les examens. Jacob et Levitt (2003) montrent par ailleurs que, pour le dispositif de Chicago, la tricherie qu'ils ont mesurée n'explique qu'une fraction infinitésimale des effets positifs de cette politique sur la performance des élèves.

Les enseignants

Le débat est le suivant : à l'accusation de démoraliser les enseignants en leur fixant des objectifs trop ambitieux et de les déprofessionnaliser en les mettant sous la coupe de la bureaucratie, les partisans de l'*accountability* répondent que celle-ci tend au contraire à professionnaliser les enseignants en les amenant à viser des objectifs plus exigeants et à un exercice du métier plus collectif et davantage fondé sur l'analyse rigoureuse des compétences des élèves (*data driven*¹⁴).

¹³ Cette pratique ne vaut plus sous NCLB puisque tous les *grades* de 3 à 8 sont testés.

¹⁴ Un programme de « data driven education » proposé aux districts par la Johns Hopkins University, qui consiste à implanter dans les écoles, outre les évaluations annuelles de NCLB, des évaluations trimestrielles et

Les entretiens que j'ai pu avoir avec des enseignants syndicalistes de Floride me les ont montré favorables aux standards, mais critiques sur la mesure des performances des élèves par les seuls tests¹⁵.

West & Paterson (2006) confirment, par une enquête plus rigoureuse, que les enseignants de Floride pensent que les standards fixés au niveau de l'Etat induisent un meilleur enseignement. Selon leur enquête, ce jugement sur les bons effets des standards prend le pas sur le stress que leur cause le fait de pouvoir être jugés à travers eux. Au total, ces auteurs estiment que l'introduction des standards a donné aux enseignants une pratique « davantage orientée vers les buts » et a accru la collaboration entre eux.

Louis (2005) a mené des entretiens approfondis avec les enseignants de trois lycées, situés dans des états différents, Elle écrit « avoir trouvé moins de résistance aux standards d'états que nous ne l'attendions », ce qui va dans le sens des résultats précédents.

Gross & Goertz, (2005) dans leur étude de 48 lycées dans six états, observent que les enseignants « n'ont pas seulement les tests pour objectif, mais prennent les tests au sérieux ». Ils sont tout à fait d'accord avec la logique des standards, avec une école qui se fixe des objectifs académiques élevés, et, écrivent elles, « ont fini par accepter le principe général d'une *accountability* basée sur la performance ». Elles concluent « La pression, la fierté professionnelle et l'acceptation de l'intérêt de la réforme pour les élèves ont ensemble conduit à ce que les enseignants se montrent réactifs à la réforme (« *contributed to a substantial amount of response* »). D'après leur étude, les lycées qui se sont montrés le plus réactifs à la réforme ne se sont pas contentés d'idées générées en interne, les réponses y ont été décidées au niveau des départements et pas de chaque enseignant séparément, les enseignants y étaient capables d'analyser les données issues des tests.

Ces diagnostics, on le voit, ne font pas état d'une adhésion enthousiaste, mais pas non plus du refus résolu que prédisaient beaucoup de critiques. Au total, ce qui se dégage, me semble-t-il, de ces recherches est que le discours critique vis-à-vis de l'*accountability* est moins faux qu'exagéré (la triche existe mais reste marginale, les enseignants réagissent mieux qu'annoncé à ces politiques, le *teaching to the test* existe mais est loin de prendre toujours des formes caricaturales) et unilatéral (les *low skills* progressent mais, au pire, les *high skills* ne régressent pas, les progrès sont plus grands dans les disciplines testées que dans les autres, mais les performances des élèves dans ces dernières ne sont, au minimum, pas pires qu'avant la mise en œuvre des dispositifs). Cependant, le test le plus important de l'*accountability* en

d'accompagner ces écoles dans l'analyse de ces données et dans le choix des politiques à mettre en œuvre pour remédier aux lacunes qu'elles révèlent, vient d'être évalué de façon très rigoureuse (*random sampling design*) par des chercheurs de l'université de Wisconsin- Madison. Les effets de ce programme sur les progrès en maths des élèves sont positifs, importants et significatifs, tandis que ceux sur la lecture sont positifs mais faibles et non significatifs (Carlson & al., 2011).

¹⁵ Il semble a priori facile de répondre à cette critique par des tests plus sophistiqués, multidimensionnels. Il faut noter cependant, dans la quasi totalité du nombre significatif d'Etats où l'on avait mis en place de tels tests, on est revenu en arrière, vers des QCM, pour des raisons de coût mais aussi de fiabilité et d'acceptabilité politiques de ces tests (Beatty et al., 2009).

général et de NCLB en particulier est leur effet sur l'efficacité et l'équité du système, soit sur l'évolution des performances moyennes des élèves, sur celles des plus faibles, sur l'écart de performances entre groupes sociaux ou ethniques (*the achievement gap*).

L'effet sur les performances des élèves.

Nous présenterons ces effets en évoquant surtout des recherches menées par des économistes. Ces derniers ont développé une expertise particulièrement forte dans l'évaluation des effets des politiques publiques, un exercice particulièrement difficile lorsque les politiques en cause ne jouent pas sur un seul facteur (diminuer la taille des classes, par exemple) mais sur l'ensemble du fonctionnement du système, lorsqu'il s'agit de réformes « systémiques ». Il faut en effet comparer les effets du système réformé avec ceux que l'on aurait observé sans la réforme, le contrefactuel. Par définition, ce dernier n'est pas observable. Il faut donc trouver des comparaisons qui s'approchent le plus possibles de celle là. Elles sont de deux types : avec/sans ou avant/après la mise en place de la politique, le premier type étant possible seulement lorsque la politique n'est pas mise en œuvre partout. Les études de ce type portent par conséquent uniquement sur la période pré NCLB, lorsque certains états ou districts avaient mis en œuvre des politiques d'*accountability* et d'autres non. Dans les deux cas, il faut veiller à ce que la situation de référence (« avant » ou « sans »), celle qui mime le contrefactuel, ne diffère de la situation évaluée que par l'absence de la politique qui est l'objet de l'évaluation. Deux difficultés se présentent à cet égard. D'abord, il n'est pas rare que plusieurs politiques soient mises en œuvre en même temps. Par exemple, le Texas a mis en œuvre dans les années 90 une politique d'*accountability* qui a été suivie d'une diminution importante des inégalités (Skrla, Johnson, Scheurich & Koschorek, 2004). Mais, en même temps, cet état avait réformé le financement des écoles de façon à égaliser les ressources des écoles des différents districts. A laquelle des deux réformes fallait-il attribuer la diminution des inégalités ? Ensuite, on mesure de mieux en mieux qu'il est difficile de construire une situation de référence adéquate. On a longtemps cru qu'il suffisait pour cela de tenir sous contrôle les principaux déterminants de la performance scolaire (la composition sociale des écoles, les performances initiales des élèves), mais on sait mieux aujourd'hui que des variables inobservables peuvent venir perturber la comparaison. En outre, les comparaisons avant/après doivent prendre garde à deux choses : d'une part, il faut comparer des tendances plus que des moments : si le niveau des élèves s'améliore après la mise en œuvre de la politique X, cela n'a pas le même sens selon que leurs performances étaient ou non déjà croissantes avant cette mise en œuvre. D'autre part, on n'est jamais sûr non plus que la situation de référence soit la poursuite linéaire des tendances observées avant la mise en œuvre (Harris, 2010). Personne ne prévoyait en 1998 en France que les performances des élèves de CM2 baisseraient fortement (MEN-DEPP, 2008) de sorte qu'une politique qui aurait stabilisé ces performances aurait été jugée inefficace sur la base d'une extrapolation des tendances antérieures, alors qu'elle aurait été en réalité efficace.

Commençons par les études avec/sans.

Amrein et Berliner (2002) comparent l'évolution des scores NAEP au cours des années 90 dans les états avec *accountability* avec cette évolution au niveau national. Si elle est supérieure, ils concluent que la politique est efficace. Ils identifient 26 états avec *accountability*, mais travaillent seulement sur ceux où il n'y a pas d'accroissement du nombre d'élèves exclus du test, soit une douzaine. Ils observent que, sur ces douze, une minorité seulement a progressé plus que le niveau national et concluent donc à l'inefficacité de l'*accountability*. Cette étude s'est attiré des critiques inhabituellement sévères de la part de deux économistes.

Hanuskek et Raymond (2003) leur reprochent d'avoir comparé les Etats avec *accountability* aux données nationales plutôt qu'aux Etats sans *accountability*, de généraliser aux 50 Etats ce qu'ils observent sur 12 seulement, de ne pas tenir compte de l'ampleur des écarts avec la moyenne nationale. Eux-mêmes utilisent une approche longitudinale, ils comparent le score NAEP du grade 4 en 1996 et le score du grade 8 en 2000 entre des Etats qui, dans cette période, avaient implanté un dispositif d'*accountability* ou non. Ils trouvent que, en moyenne, le niveau des élèves augmente plus entre les grades 4 et 8 dans les Etats « avec » que dans les états « sans ». En maths, l'élève moyen dans un état « avec » voit son score NAEP augmenter de 1,6 % par an contre 0,7% dans un Etat « sans » et 1,2 % dans des Etats où les résultats des écoles sont rendus publics mais sans qu'y soient attachées des incitations (*soft accountability*).

Bishop et al. (2001) (cité par Harris & Herrington, 2006) utilisent aussi les scores NAEP dans une approche semblable : ils trouvent des effets positifs aux dispositifs qui comportent des sanctions (*hard accountability*), mais des effets pas très nets lorsque le dispositif comporte seulement la publication des performances (*soft accountability*).

Carnoy et Loeb (2002) est une des études « avec/sans » les plus rigoureuses et les plus connues. Elle tient sous contrôle les facteurs qui pourraient fonctionner comme cause commune à un accroissement des scores et à l'implantation de l'*accountability* (la tendance des scores avant l'introduction de l'*accountability*. par exemple, est tenue sous contrôle, de même que la proportion d'élèves noirs et hispaniques, la proportion de ressources venant de l'Etat plutôt que du district, etc.) . Les auteurs établissent une échelle 0-5 du degré d'*accountability* selon une série de critères (nombre de grades avec tests de l'Etat en 1999-2000; publication des résultats des écoles; incitations; existence d'examens de sortie en fin de lycée) et ils étudient les relations entre la place d'un état sur cette échelle et l'évolution de ses résultats au NAEP 4 et 8, le taux de redoublement au grade 9, et enfin la proportion d'élèves atteignant le grade 12 (notre classe de terminale). En effet, certains craignaient que l'*accountability* ait pour effet d'augmenter le taux de redoublement et de sorties précoces (*dropouts*).

Ils observent ceci. D'abord sur la proportion d'élèves qui atteignent au moins le niveau "basic" aux grades 4(CM1) et 8 (4ème) (son évolution entre 96 et 2000, sous contrôle de plusieurs facteurs) : « Une augmentation de deux degrés dans l'échelle d'*accountability* correspond à peu près à un gain plus grand d'environ un demi écart-type. Ceci est vrai pour les blancs, pour les noirs et pour les hispaniques (p322). Le pourcentage de « au moins proficient » augmente encore plus, comme nous l'avons déjà évoqué. Ces gains sont plus

forts pour les élèves du *grade* 8 que pour ceux du *grade* 4. Ensuite, le degré d'*accountability* n'a pas d'effet significatif sur le taux de redoublement et ni sur le taux d'achèvement des études secondaires. Enfin, comme les effets sont parallèles pour les différents groupes ethniques, les écarts entre eux ne sont pas atténués.

Lee et Wong (2004) est une étude du même type que la précédente, comparant l'évolution de l'équité de l'éducation au cours des années 90 dans les états à *accountability* forte et dans ceux à *accountability* faible ou nulle. Ils observent que les inégalités entre groupes sociaux évoluent de façon semblable dans les deux groupes d'états, à la fois quant aux ressources des écoles qu'ils fréquentent majoritairement (enseignants qualifiés, taille des classes, dépenses totales) et quant aux scores.

Ces études convergent vers un effet positif des dispositifs d'*accountability* implantés dans les années 90 sur les performances moyennes des élèves et un effet nul sur l'*achievement gap*, sauf peut être au Texas où une diminution de cet écart a été observée (Skrla et al., 2004)¹⁶.

Une forme particulière d'études avec/sans est celle qui compare, dans le cadre de dispositifs d'*accountability*, des écoles qui ont été mises sous pression par le dispositif et les autres : si le dispositif produit les effets attendus, les premières doivent faire, toutes choses égales par ailleurs, davantage de progrès que les secondes pour échapper à ce statut, alors que les critiques avancent que les écoles ainsi désignées sont précisément celles qui manquent des ressources nécessaires à améliorer leurs performances¹⁷.

Plusieurs de ces recherches portent sur le dispositif floridien, le A+plan¹⁸ et donnent une appréciation positive de l'effet de la mise sous pression.

Figlio et Rouse (2006) observent que les scores des écoles F augmentent nettement plus que ceux des autres écoles. Ces résultats sont atténués parce qu'ils sont en partie dus, disent-ils, à (i) la régression à la moyenne (les écoles à bas scores peuvent l'être par effet d'une erreur de mesure et donc leurs progrès seront d'autant plus grands l'année suivante, exagérant l'effet positif de l'*accountability* sur les écoles en difficulté) (ii) au fait que les écoles se concentrent sur les grades qui font l'objet d'une évaluation à enjeu (iii) au fait que certains districts ont changé les zones de recrutement des écoles F pour améliorer leur population et donc leurs scores. L'écart F/non F existe aussi, on l'a vu, sur les tests sans enjeu, même si ils sont moins

¹⁶ Au Texas en général (Skrla et al.,2004) mais pas dans le district de Dallas, où Ladd(1999) (citée par Harris & Herrington, 2006) observe un effet très important du dispositif pour les élèves blancs, plus faible pour les hispaniques et nul pour les élèves africains- américains, et donc un creusement de l'*achievement gap*.

¹⁷ Cette question est évidemment importante pour apprécier les effets de l'*accountability*. Ces études ne nous apprennent rien, cependant, sur un autre effet attendu, ceux des actions entreprises par les écoles pour *ne pas* tomber sous ce statut..

¹⁸Dans ce dispositif, les écoles reçoivent une note de A (la meilleure) à F, en fonction d'une série d'indicateurs portant sur les performances des élèves, mais aussi leur taux d'absentéisme par exemple. Les écoles notées F font l'objet d'une évaluation approfondie par une équipe externe qui propose un plan d'amélioration. Celles qui sont notées F deux ans de suite voient leurs élèves recevoir un bon qui leur permet d'aller dans une autre école, publique ou privée (cette dernière possibilité ayant fait l'objet de procès devant les tribunaux, comme attentatoire à la séparation entre l'église et l'Etat).

importants que sur les tests liés au dispositif . Selon leurs observations, le reproche de privilégier les grades évalués par le dispositif n'est pas justifié : Les écoles consacrent plus d'énergie à ces grades mais pas au détriment des autres grades. Enfin, ils montrent que les progrès des écoles en difficulté étaient aussi grands lorsque la sanction était seulement la publication des tests (le *stigma*) et ne comportait pas la menace d'autoriser les élèves à quitter leur école pour une autre.

Rouse, Hannaway, Goldhaber&Figlio (2007) comparent les écoles situées juste de part et d'autre du label F. Cette méthode (*régression discontinuity design*) utilise le fait que, compte tenu des erreurs de mesure, ces écoles ont toute chance d'être en réalité d'efficacité égale, et qu'en les comparant, on observe donc l'effet « net » de la mise sous pression. Son inconvénient est de ne rien nous apprendre sur l'effet de la politique sur les écoles les plus en difficulté. Ils mesurent que l'attribution du label F se traduit par une augmentation nette de la performance des élèves.

Winters, Trivitt et Greene (2010) utilisent la même technique. Ils observent eux aussi un effet positif de la mise sous pression, en maths et lecture surtout, mais aussi, on l'a vu, en sciences, bien que cette discipline ne soit pas évaluée par le dispositif.

Le grand intérêt de l'étude de West et Peterson (2006) est de comparer l'effet des *choice threats* (la menace d'autoriser les élèves à désertir leur école si elle ne s'améliore pas) sous le A+ Plan d'une part et sous NCLB d'autre part. Eux aussi comparent des écoles situées immédiatement de part et d'autre du seuil. Comme les études précédentes, ils observent que la menace est efficace dans le cadre du A+ plan, mais ils observent aussi que, dans le cadre de NCLB, elle ne l'est pas. En cause selon eux : d'une part un effet de stigmatisation beaucoup plus faible dans la mesure où les écoles *INI* de NCLB sont beaucoup plus nombreuses que les écoles F du A+plan¹⁹, d'autre part, une menace moins forte puisque le choix offert par NCLB est limité aux écoles publiques du district²⁰.

Deux autres études étudient l'effet de la mise en *INI* sous NCLB, l'une dans le Maryland (Hemelt, 2011), l'autre dans plusieurs états du Nord Ouest (Springer, 2008).

Springer (2008) observe, lui, un effet positif de la mise en *INI* sur les performances des élèves. On l'a vu, il trouve aussi que, plus un élève est faible (loin du seuil), plus ses progrès sont importants par rapport à ceux d'un élève identique dans une école non *INI*... et que cet effet positif ne s'obtient pas au détriment des bons élèves de ces écoles.

L'approche de Hemelt (2011) est aussi une régression de discontinuité. Les élèves d'une école mise sous *INI* progressent-ils davantage que ceux d'une école semblable, mais ayant échappé de peu à la mise sous *INI* ? Sa réponse est négative lorsqu'il s'agit d'une école mise en *INI* à

¹⁹Ceci surtout parce que le fait que tous les groupes sociaux ou ethniques séparément doivent atteindre les AYP est très exigeante.

²⁰ En effet, Mintrop & Sunderman (2009) le pointent, 1% seulement des familles éligibles à cette possibilité en ont profité. Cette anecdote va dans le même sens : Une directrice d'école primaire de l'Indiana s'est vantée auprès de moi que, la réputation de son école étant bien meilleure que son classement NCLB, une seule famille avait profité de la possibilité de choix et que, l'ayant appris, six autres lui avaient aussitôt demandé d'accueillir leur enfant.

cause de la moyenne d'ensemble de ses élèves : il observe même que les élèves de ces écoles progressent moins que ceux des écoles témoins. En revanche, lorsqu'une école a été désignée *INI* à cause des résultats de certains groupes d'élèves seulement alors que, en moyenne pour tous ses élèves, elle avait fait ses *AYP*, alors, les scores des élèves de ces groupes s'améliorent davantage que ceux des élèves équivalents d'écoles non *INI*. Ceci suggère évidemment qu'une telle mise sous pression est davantage capable d'aider les écoles qui ont des problèmes locaux, susceptibles d'être résolus par un déplacement des efforts et des moyens, que des écoles qui rencontrent des difficultés avec la grande majorité de leurs élèves. On l'a évoqué, ces résultats appuient le projet de l'administration Obama d'engager des actions spécifiques pour les écoles les plus en difficulté.

Que disent les comparaisons avant/après la mise en œuvre des politiques d'*accountability* ?

Jacob (2005) observe que les performances de élèves augmentent davantage après la mise en place d'un dispositif d'*accountability* à Chicago en 1996 que, à la même période, dans d'autres grands districts de l'Illinois qui n'ont pas mis en place de politique d'*accountability* et surtout que l'augmentation des performances des élèves de Chicago est plus forte après qu'avant la mise en œuvre de la politique. Toutefois, on l'a vu, il met en évidence une série de faits qui relativisent ce résultat : Les progrès s'observent sur le test utilisé pour l'*accountability* (l'ITBS) aux grades 3,6 et 8, mais seulement au grade 8 sur un autre test non lié à elle (l'IGAP). Les progrès sont deux fois plus importants pour les *low skills* que pour les *high skills*. Les enseignants auraient déplacé ressources et efforts vers l'enseignement des maths et de la lecture aux élèves faibles: les performances de ces derniers ont augmenté plus que celles des autres élèves en maths et en lecture mais pas plus en science et en *social studies*. Un tel résultat est susceptible de deux lectures. Un adversaire de la politique soulignera le caractère partiel des progrès, un partisan se félicitera de ce que les priorités de la politique (les progrès en maths et lecture des élèves faibles) soient tenues sans dommage pour les autres apprentissages et pour les apprentissages des autres.

Une autre approche avant/après consiste simplement à observer si les performances moyennes des élèves, dans un Etat ou au niveau national, sont meilleures après qu'avant la politique. Les performances des élèves avant et après NCLB peuvent être comparées en utilisant soit les tests des Etats (qui ont un avantage politique : selon la loi, ils sont le critère du succès ou de l'échec) ou le NAEP (qui a un avantage scientifique : les écoles ne sont pas susceptibles d'y préparer leurs élèves de façon mécanique). Mintrop et Sunderman (2009) reconnaissent que les résultats aux tests des Etats se sont améliorés après NCLB mais avancent que ce n'est pas le cas selon le NAEP, ce dont ils déduisent que l'amélioration d'après les tests des Etats, entièrement due au *teaching to the test*, est une amélioration en trompe l'œil. Cela n'est pas tout à fait exact. La divergence entre les résultats aux tests des Etats et à ceux du NAEP, en effet, ne doit pas être surestimée. Sur 33 Etats sur lesquels la comparaison est possible pour les années 2002-2007, le pourcentage d'élèves situés au dessus du seuil *basic* ou *proficient* s'est accru à la fois selon le NAEP et selon les tests des états dans 30 pour la lecture au grade 4, dans 19 pour la lecture au grade 8, dans 31 pour les maths au grade4 et dans 28 pour les maths au grade 8. S'agissant des écarts entre les groupes sociaux, sur 280 comparaisons possible (groupes/ matières/grades) 232 écarts se sont réduits et 41 se sont creusés selon les

tests des Etats tandis que, selon le NAEP, 178 se sont réduits et 104 se sont creusés. L'écart entre les élèves dont les parents ont des revenus élevés vs faibles se sont réduits à la fois selon le NAEP et selon les tests des états dans 20 états sur les 30 où les données sont disponibles (CPE, 2008).

Toutefois, on l'a dit, un test plus authentique est de savoir si les performances des élèves ont augmenté *davantage* après qu'avant NCLB. La réponse du NAEP est ici plutôt négative, le taux d'augmentation est semblable dans les deux périodes (NCES, 2007, 2009). Plus généralement, d'ailleurs, Lee (2007, citée par Mintrop & Sunderman, 2009) a étudié l'évolution des scores au NAEP depuis sa création (1971), elle observe une tendance « de petite à modérée » à la hausse du score moyen et à la réduction de *l'achievement gap*, mais cette tendance est linéaire sur la période et elle ne semble affectée par aucune des politiques éducatives nationales mises en œuvre pendant la période, y compris NCLB. La pente est même plus faible dans la période 2007-2009 que dans la période 2003- 2007, ce qui a poussé certaines personnalités auparavant partisans de NCLB à se prononcer contre cette loi (Ravitch, 2009). Par ailleurs, la réduction de l'écart entre élèves blancs et noirs s'opère plutôt au même rythme dans les années précédant 2003 et dans les années suivantes²¹.

Cependant, deux arguments peuvent être avancés pour considérer que la poursuite au même rythme de l'amélioration précédant NCLB n'indique pas forcément un échec de cette politique: D'une part, cela reste une hypothèse que, sans NCLB, l'amélioration des performances des élèves se serait poursuivie au même rythme (Harris, 2010). D'autre part, on peut imaginer que la linéarité observée par Lee (2007) soit le signe, non pas d'une absence d'effet de ces politiques, mais d'un effet semblable de *toutes* ces politiques²² (Goals 2000, Improving America School act (1994),...) et que la diffusion progressive des dispositifs d'*accountability* au cours des années 90 soit, au moins en partie, responsable des améliorations observées durant ces années là. Les recherches citées ici, en général positives pour les dispositifs mis en place par les états dans les années 90, vont dans ce sens, ainsi qu'une autre : Selon Dee et Jacob (2010), les scores des élèves ont augmenté davantage, après NCLB, dans les Etats qui n'avaient pas implanté auparavant de pratiques rigoureuses d'*accountability* que dans ceux qui en avaient déjà implanté.

Au crédit de NCLB, il semble aussi qu'on puisse mettre l'évolution, aux Etats-Unis, des performances à PISA même si, bien sûr, d'autres éléments ont pu jouer. L'évolution des scores moyens n'est pas particulièrement remarquable mais celle des élèves les plus faibles ou les plus défavorisés socialement est nettement positive, surtout si on la compare avec la même évolution en France. De 2003 à 2009, en compréhension de l'écrit, le score du premier décile, des élèves les plus faibles, donc, augmente de 11 points (sur 500) aux Etats-Unis (augmentation non significative) et baisse de 15 points en France (baisse significative) tandis que l'impact du milieu social de l'élève sur son score diminue de 5 points aux Etats-Unis tandis que il augmente de 6 points en France. Il faut noter aussi que, aux Etats-Unis, les scores

²¹ Ceci est vrai pour le grade 8. Au grade 4 les résultats sont contrastés : le rythme de diminution de l'écart s'est accéléré en lecture, fortement ralenti en maths.

²² Toutes ces politiques avaient pour objectif une amélioration des performances des élèves dans les disciplines de base.

PISA des meilleurs élèves ont augmenté, depuis 2003, moins que ceux des plus faibles, ce qui rejoint certains résultats présentés ci-dessus (Krieg, 2011 ; Springer, 2008), et va dans le sens d'une équité accrue du système éducatif américain, du moins pour ceux qui considèrent la réduction de l'écart entre les plus faibles et les plus forts à la fin de la scolarité obligatoire comme un critère d'équité²³.

Même si les interprétations des évolutions que l'on vient d'indiquer divergent, tous les observateurs conviennent que le rythme d'amélioration actuel ne permettra pas d'atteindre l'objectif fixé par la loi pour 2014. Certains en concluent que la date butoir doit être reculée mais d'autres estiment qu'il n'y a pas de sens à fixer un même objectif à tous les élèves. Parmi les seconds, Rothstein et al (2006) estiment qu'un objectif mobilisateur pour les élèves faibles ne peut que démobiliser les élèves forts et réciproquement. Ils se prononcent donc pour l'abandon de tout objectif absolu et son remplacement par des indicateurs permettant de vérifier que l'efficacité des écoles augmente d'une année sur l'autre.

Conclusion

Il serait prudent de tirer de ces recherches une conclusion fréquente quand on mesure les résultats d'une politique qui est l'objet de controverses : Les effets ne sont ni aussi catastrophiques que ceux imaginés par ses adversaires, ni aussi positifs que ceux imaginés par ses partisans.

Je souhaite ici aller plus loin, même si cela amène à quitter le terrain de ce qui est sûr pour celui de ce qui est le plus probable, terrain qui est celui du politique plus que celui du scientifique²⁴.

Ces études montrent, à mon sens, qu'il y a plus de chances d'améliorer l'efficacité et surtout l'équité des systèmes scolaires si l'on s'engage dans une politique de régulation par les résultats soucieuse des élèves les plus faibles que si on ne s'y engage pas. L'idée que la mise en œuvre de dispositifs d'*accountability* s'accompagne plus souvent d'une amélioration que d'une dégradation de l'équité peut aussi s'appuyer sur le constat suivant : Parmi les cinq pays de l'OCDE où l'impact du milieu social sur le score PISA en compréhension de l'écrit a

²³ L'amélioration de l'équité est particulièrement forte en sciences entre 2006 et 2009, alors qu'il s'agit d'une discipline testée par NCLB seulement depuis 2007 (et dont l'évolution ne peut être mesurée par PISA que depuis 2006). Cela ne va pas dans le sens de la thèse qui attribue ces progrès à NCLB, sauf à penser que les effets les plus nets de ces politiques s'observent immédiatement après leur implantation.

²⁴ Le public pourra reprocher à un politique de ne pas s'être engagé dans la politique X à un moment donné, s'il s'avère plus tard que c'était une bonne politique, et ledit politique, qui ne prend que des décisions en univers incertain, ne pourra répondre qu'il attendait des preuves irréfutables pour se décider. Je reprends ici l'argumentation de Harris (2010) : Les chercheurs doivent surtout éviter les erreurs de type 1 (voir un effet où il n'y en a pas), mais, comme le public pourra reprocher aux politiques de ne pas avoir tout tenté pour remédier à une situation insatisfaisante, ces derniers doivent attacher au moins autant d'importance aux risques d'erreurs de type 2 (ne pas voir d'effet où il y en a).

diminué de façon significative entre 2000 et 2009 (OCDE, 2010, p 163), quatre (Canada, Chili, République tchèque, Etats-Unis) ont mis en place une forte *accountability*²⁵. Parmi les quatre pays où cet impact a augmenté de façon significative, cependant, trois (Corée, Islande, Suède) ont aussi implanté de forts dispositifs d'*accountability*. Parmi les six pays où le premier décile des scores a augmenté de façon significative, deux (Portugal, Pologne) ont mis en place une forte *accountability* et deux (Danemark, Norvège) ont mis en place une *accountability* faible. Inversement, parmi les trois pays où les scores ont baissé de façon significative un seul pays (la Suède) a mis en place une *accountability* (forte).

Cependant, la régulation par les résultats peut revêtir des formes diverses et une des leçons de ces recherches (Hemelt, 2011, par exemple) est que, autre leçon fréquente de l'évaluation des politiques systémiques, « le diable est dans les détails » de sorte que l'engagement dans une telle politique ne peut être que prudent, accompagné d'évaluations qui en permettent la réorientation.

Cette prudence est requise à deux niveaux :

- politique, d'abord. Puisqu'il est avéré qu'une telle politique réoriente effectivement l'enseignement, il faut s'assurer que ses objectifs (les standards, mais aussi les objectifs fondés sur eux, la priorité donnée aux élèves faibles par NCLB par exemple), que ses tests, font l'objet d'un accord le plus large possible parmi le public et aussi parmi les enseignants. Cela suppose la mise en place de procédures démocratiques ad hoc.
- technique, ensuite. La réforme de NCLB prévue par l'actuelle administration américaine paraît un bon exemple de pilotage tirant les conclusions des points faibles révélés par les évaluations. Ainsi des politiques spéciales en direction des écoles les plus en difficulté, puisqu'il semble qu'elles ne bénéficient pas du dispositif actuel et peuvent même en pâtir (Hemelt, 2011) ; ainsi de la décision de (re)décentraliser au niveau des états la définition du régime d'incitations, qui fait écho au fait que les évaluations des dispositifs pré-NCLB, en Floride et au Texas notamment, sont plus unanimement positives que celles de NCLB. Mais il reviendra dès lors à chaque état de piloter cette politique à partir d'évaluations rigoureuses.

Septembre 2011.

²⁵ Plus précisément : un pourcentage élevé de leurs établissements répond au questionnaire PISA 2009 que les évaluations des élèves sont utilisées fréquemment, soit pour prendre des décisions sur le fonctionnement des établissements ou pour informer le public ou comparer les établissements (*accountability faible*), soit utilisées fréquemment dans ces deux directions à la fois (*accountability forte*) (OCDE, 2010, *Strong Performers, and successful reformers*, p 51). Ces dispositifs ne visent pas forcément autant l'amélioration des performances des élèves les plus faibles que le dispositif américain.

Bibliographie

- Amrein, A.L. & Berliner, D.C. (2002). High-stakes testing, uncertainty, and student learning *Education Policy Analysis Archives*, 10(18).
- Au, W. (2007) High Stake testing and curricular control : a qualitative meta synthesis, *Educational Researcher*, 36, 258-267.
- Beatty,A. &al., 2009, *Best Practices for State Assessment Systems*, Part1, National Research Council, disponible sur la toile.
- Carlson, D., Borman,G.G. & Robinson, M. , 2011, A Multistate District- Level Cluster Randomized Trial of the Impact of a Data –Driven Reform, *Educational Evaluation and Policy Analysis*, 33 (3) 378-398.
- Center for Education Policy (2008) *Has student achievement increased since 2002?* Washington, DC.
- Center for Education Policy (2011) How many schools have made AYP? (disponible sur www.cep.dc.org)
- Chubb, JE &Moe, T.M (1990) *Politics, Market and America's schools*, Washington : The Brookings Institution.
- Debray, R., 1998, Lettre ouverte à M. le Ministre de l'Education Nationale, *Le Monde*, 3 mars.
- Dee, T. & Jacob, B., 2010, Evaluating NCLB, *Education Next*, summer 2010 10(3).
- Garrison, J. & Neiman, A., 2003, *Pragmatism and Education*, in The Blackwell Guide to the Philosophy of Education, N. Blake et al., eds., Maiden, Ma: Blackwell Publishing.
- Figlio, D.N. & Rouse, C.E., 2006, Do accountability and vouchers threats improve low-performing schools? *Journal of Public Economics*, 90.
- Friedman, M. (1962) *The role of Government in Education*, in Capitalism and Freedom, Chicago: The University of Chicago Press.
- Gross, B. & Goertz,M.E., 2005, Holding high hopes, How High schools respond to State Accountability Policies, CPRE Research Report Series, RR056.
- Hanushek, E & Raymond, M., 2003, High Stakes Research, *Education Next*, vol3 n°3
- Harris, D., 2010, Race to the top, Are the Naysayers right? *Education Week*, 26.3.
- Harris, D; & Herrington, C., 2006, La régulation par les résultats contribue-t-elle à l'amélioration des écoles? in Chapelle, G. et Meuret,D., Améliorer l'école, Paris: PUF.

Hemelt, S.W., 2011, Performance effects of failure to make AYP: evidence from a regression discontinuity framework, *Economics of Education Review*, 30 (4), 702-723.

Holmstrom, B.& Milgrom, P., 1991. Multitask principal-agent analyses: incentive contracts, asset ownership and job design. *Journal of Law, Economics and Organization* 7, 24–52.

Jacob, B. A., 2005. Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, 89(5–6), 761–796.

Jacob, B.A., 2007, *Test Based Accountability and student achievement, an investigation of differential performance on NAEP and State assessments*, NBER; Working Paper 12817.

Jacob, B.&Levitt, S., 2003. Rotten apples: an investigation of the prevalence and predictors of teacher cheating, *Quarterly Journal of Economics*, CXVIII (3), 843– 878.

Krieg, J.M., 2011, Which students are left behind, *Economics of Education Review*, vol. 30(4).654-664.

Lee,J. &Wong, K.K., The impact of accountability on racial and socioeconomic equity: Considering both School Resources and Achievement Outcomes, *American Educational Research Journal* 41(4), 797-832.

Leightwood, K, & Menzies, T. (1998). A review of research on School Based Management. *School Effectiveness and School Improvement*, 9, 3, 233-285.

Lessard, C. & Meirieu, Ph., 2005, *L'obligation de résultats en éducation*, Bruxelles : de Boeck.

Louis, K.S., Febey,K. & Schroeder,R. 2005, State Mandated Accountability in High Schools: Teacher interpretation of a new era, *Educational Evaluation and Policy Analysis*, 27(2), 177-203.

Mc Neil, L., 2000, *Contradictions of School Reform, Educational Cost of Standardized testing*, New York: Routledge.

Meuret, D., 2007, *Gouverner l'Ecole*, Paris: PUF.

Mintrop, H. and Sunderman, G., 2009, « Predictable failure of federal sanctions-driven accountability for school improvement, and why we may retain it anyway » *Educational Researcher*, 38(5).

Ravitch, D., 2009, *The Death and Life of the Great American School System: How testing and Choice are undermining Education*, New York: Basic books.

Rothstein, R., Jacobsen, R. & Wilder, T., 2006, *Proficiency for all, an oxymoron*, paper for the Symposium “Closing the achievement gap: NCLB and its alternatives” New York: Teacher college.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2007). *Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure*. NBER Working Paper Series, N°. 13681.

Skrla, J. Johnson, J.F., Scheurich, J.J. & Koschorek, J.W., 2004, *Accountability for Equity*, in J. Skrla & J.J. Scheurich, *Educational Equity and Accountability*, New York: Routledge & Farmer.

Sims, D.P., 2008, Strategic response to school accountability measures: it’s all in the timing, *Economics of Education Review*, vol. 27(1), 58-68.

Springer, M. G. (2008). The influence of an NCLB accountability plan on the distribution of student test score gains. *Economics of Education Review*, 27, 556–563.

West, M. R., & Peterson, P. E. (2006). The efficacy of choice threats within school accountability systems: Results from legislatively induced experiments. *The Economic Journal*, 116 (510), C46–C62.

Winters, M.A., Trivitt, J.R. & Greene, J.P., 2010, The impact of high stake testing on student proficiency in low stake subjects: Evidence from Florida’s elementary science exam, *Economics of Education Review*, 29(1) pp. 138-146.

MEN-DEP, 2008, Lire, écrire, compter : les performances des élèves de CM2 à vingt ans d'intervalle 1987-2007, NI 0838

NCES, 2009, Nation Report Card, Washington (DC): Department of Education.

NCES, 2007, Nation Report Card, Washington (DC): Department of Education.

OCDE, 2010, PISA 2009 results, vol5: Learning trends, Paris: OCDE.

US Department of Education (2010) A blueprint for reform, the reauthorization of the Elementary and Secondary Education Act (disponible sur le site du US Department of Education).

