Denis Meuret

Audition par le Comité d'évaluation et de contrôle des politiques publiques[1]

Assemblée Nationale, 16 05 18

Développer les évaluations standardisées, et qu'en faire

1. Faut -il développer les évaluations standardisées ?

Oui.

a) Elles sont possibles. Les adversaires de l'évaluation présentent cette pratique comme quelque chose qu'on s'était abstenu de faire jusqu'ici et qu'on veut mettre en œuvre maintenant. Or, ce qui se passe en réalité est que de multiples progrès, en informatique évidemment mais aussi en statistiques ou en dans la mise au point d'épreuves, ont rendu possible ce qui ne l'était pas auparavant : disposer pour un cout raisonnable de données de qualité sur les compétences et les connaissances des élèves permettant de les comparer dans le temps ou dans l'espace et de s'interroger sur les facteurs qui les expliquent. A tel point que dans les pays anglo-saxons ou scandinaves, les plus en pointe dans ce domaine, on valorise une « *data driven education* » et on accorde une grande importance à former les enseignants à l'analyse de ces « *data* ».
Le bon sens indique que n'importe quelle organisation, n'importe quel agent, doit tenter de mesurer les effets de son action pour vérifier qu'ils sont conformes à ce qu'elle souhaite. On peut raisonner par l'absurde : si on s'en abstient, alors on ne peut se guider que sur des croyances ou des convictions, sans les mettre à l'épreuve des faits. Or, le rapport de la Cour des Comptes a raison de souligner les limites des indicateurs fondés seulement sur la réussite aux examens, que ce soit dans la LOLF ou pour

[1] Ce texte est une version longue de l'exposé prononcé devant cette Commission.

la performance des lycées (IVAL). Bref, on peut parier que, si Jules Ferry en avait eu les moyens informatiques, il aurait mis en place des évaluations nationales.

b) Les épreuves standardisées (dont les modalités de passation et de corrections sont standardisées) donnent une vision pertinente des performances des élèves. Une bonne éducation est une éducation qui produit de meilleurs personnes, qui a des effets externes positifs. Personne ne prétend que la maitrise des conjugaisons latines a une vertu en elle-même. Sa vertu (prétendue) est de former l'esprit d'individus plus rationnels, plus rigoureux. C'est donc en s'intéressant aux corrélations entre les scores aux tests standardisés obtenus par les élèves et ce qu'ils deviennent ensuite qu'on peut juger de leur validité.

   Or, les élèves qui ont des scores plus élevés aux évaluations standardisées ont un meilleur avenir. Par exemple, un suivi au Canada des élèves de l'échantillon PISA a montré que, à milieu social égal, ceux qui avaient de meilleurs scores PISA faisaient une meilleure carrière scolaire ensuite, réussissaient mieux dans l'enseignement supérieur. Cela montre au moins que PISA mesure de façon correcte des compétences utiles.

   Par exemple aussi, une étude sur 2,5 millions d'élèves américains (Chetty et al., 2012), a montré que parmi des élèves identiques, en particulier de même milieu social, ceux qui avaient eu, à l'école primaire ou au collège, des professeurs plus efficaces d'après des mesures de valeur ajoutée, obtenues en comparant les résultats de leurs élèves en début et en fin d'année avec des épreuves standardisées, faisaient plus souvent des études supérieures, plus souvent dans des universités plus cotées, avaient à 28 ans de meilleurs salaires, habitaient des quartiers dont les habitants sont plus éduqués, les filles étaient moins souvent enceintes à l'adolescence.

Elles donnent aussi une représentation juste des performances des élèves, mais ceci est l'objet de votre deuxième question, à laquelle je répondrai tout à l'heure.

Donc, oui, il est, surtout aujourd'hui, possible de mesurer les résultats de l'éducation et les épreuves standardisées – qui ne sont pas forcément composées exclusivement de QCM, peuvent comporter des réponses à des questions ouvertes - sont le moyen le plus expédient pour le faire. Aux Etats-Unis, nombreux sont qui reprochent aux épreuves standardisées de « rétrécir le curriculum » ou de favoriser le « teaching to the test ». Les Etats font donc des tentatives régulières pour remplacer ou compléter ces épreuves par d'autres, plus « qualitatives », des port-folio par exemple. Toujours aussi régulièrement, on

abandonne ces autres méthodes après un certain temps, leur fidélité s'avérant insuffisante, leur coût trop élevé. A quoi il faut ajouter que l'avantage des épreuves standardisés sur les épreuves élaborées par les enseignants pour leurs propres classes, disons les évaluations artisanales, est que les premières font l'objet de contrôles de qualité, de leurs propriétés psychométriques par exemple (*Educational Testing Service*, qui fournit aux Etats américains nombre de leurs tests, est une organisation à but non lucratif dont le budget annuel est d'environ 900 millions de $ et qui emploie des centaines de chercheurs). On est par conséquent capable aujourd'hui de faire des épreuves standardisées de meilleure qualité, ce qui ne se serait pas produit si on avait renoncé à les développer à cause des insuffisances de leurs premières versions.

c) Beaucoup de pays le font. PISA 2015 demande aux chefs des établissements de l'échantillon (voir PISA 2015 results, vol2, table II.4.27) à quelle fréquence les élèves de leur établissement passent des « tests standardisés obligatoires », ceux qui nous occupent aujourd'hui. En moyenne dans les pays de l'OCDE, 77% des élèves appartiennent à des établissements où, selon le chef d'établissement, cela a lieu au moins une fois par an. En France, selon les réponses des chefs d'établissements, ce pourcentage est, d'une façon d'ailleurs étonnamment forte, de 67%[2]. Cependant, certains systèmes scolaires, y compris parmi les plus efficaces selon PISA (Suisse, Japon, Finlande, par exemple), n'y ont pas recours. Dans 20 pays sur les 32 pays sur lesquels a porté une étude de la Commission Européenne (Eurydice, Les évaluations standardisées des élèves en Europe : objectifs, organisation et utilisation des résultats, 2009, p. 25) des tests nationaux sont utilisés pour le pilotage des établissements ou des systèmes d'éducation.

Le problème -technique, politique- des épreuves standardisées est moins leur justesse ou leur pertinence que ce qu'on fait de leurs résultats. De ce point de vue, on peut estimer que le rapport de la cour des Comptes a tendance à confondre « assessment » (la mesure des performances des élèves), « evaluation » (mesurer les performances des élèves et comprendre leurs causes) et « regulation » (faire en sorte que ces performances soient conformes aux objectifs fixés par la loi ou l'administration). La Cour semble supposer que, si on a le diagnostic, on a le traitement.

[2] On peut de ce fait s'interroger sur la validité de la réponse dans les autres pays, il faut sans doute considérer ces pourcentages avec prudence.

Or, à diagnostic donné, la mise en œuvre d'un traitement rencontre justement ces difficultés techniques et politiques. Considérons tour à tour ces deux dimensions – technique, politique- des problèmes que rencontre l'aval de la mesure es performances.

Ces problèmes sont « **techniques »** parce qu'il n'est pas simple de remonter d'un résultat à ses causes, d'imaginer des pratiques ou des dispositifs plus efficaces et de les mettre en œuvre. L'OCDE ne trouve pas de corrélation, au niveau des pays, entre la fréquence des évaluations standardisées et la performance, probablement à cause de cette difficulté-là, peut être aussi à cause de la faible validité des réponses sur la fréquence de l'usage des évaluations. Un article récent (Harris, A.), consacré aux effets sur les établissement de la recherche sur leur efficacité a été intitulé *« Lost in translation ? »*.

La résolution de ces problèmes techniques, plus complexes que ne l'imaginaient les promoteurs des évaluations standardisées, y compris l'auteur de ces lignes, prendra du temps. Elle viendra de la recherche et des expérimentations de terrain. Le deux réclament justement que l'on puisse mesurer les performances des élèves. En effet, disposer de mesures générales et routinières des performances permet de dégager du temps pour ce qui est vraiment difficile (comprendre comment on en est arrivé là, dans chaque classe pour chaque établissement) ce qui explique que recherche et expérimentations se concentrent dans les pays qui ont développé ces évaluations à large échelle. Dans ces pays, un nombre impressionnant de recherches de grande qualité portent précisément sur l'aval de l'*assessment*.

Le tableau suivant donne quelques exemples d'utilisations possibles des *assessments*, sachant bien sûr que les dispositifs évoqués peuvent se passer ce ces épreuves (par exemple, le choix de l'école peut se faire sur la base de simple réputation des établissements, l'évaluation des enseignants peut reposer sur d'autres procédures, etc.). En gros, trois modalités sont possibles au niveau des enseignants ou des établissements. C'est une décision stratégique du système que de décider lesquels privilégier : information des usagers ou des acteurs, régulation administrative jouant sur des incitations, engagement dans des processus d'amélioration.

Utilisations possibles des résultats des épreuves standardisées au niveau			
Elèves	Enseignants	Etablissements	Système
Déclencher une aide spécifique à un élève (ex :	Incitations positives (primes, évolutions de	Incitations positives ou négatives (ex :	Régulation

PRE)	carrière, etc) ou négatives (mise sous tutelle de pairs, licenciements)	NCLB[3])	
Adapter l'enseignement aux forces et faiblesses réelles des élèves	Amélioration par la formation ou l'aide par des pairs.	Entrer dans un processus d'amélioration[4] *(improvement*)	Amélioration
Rendre compte des progrès de l'élève à lui-	Feedback de leur valeur ajoutée aux enseignants.	Feedback de leur valeur ajoutée aux	Information

[33] Sur No Child Left Behind et ses effets, voir Meuret, D., 2012, *Les effets de l'accountability sur les politiques d'éducation aux Etats-Unis*, in Education et Société, n°30 ou , *Faut-il se réjouir de la fin de NCLB ?,* Café Pédagogique, 15.05. 15 ou, *Plus équitable, l'école retrouverait la voie de la réussite*, Esprit, décembre 2012.

[4] Il y a deux moyens de concevoir cette amélioration. On peut insister sur l'inutilité d'inventer la roue dans chaque établissement, proposer aux établissements d'adopter tel ou tel mode de fonctionnement (*school design)* et organiser une sorte de marché de ces designs en les évaluant. Cela est fait aux Etats-Unis sous la forme de *Comprehensive school programs. (Borman et al. , 2003, CSR achievement, a meta analysis, Review of Educational Resarch)* compilent les évaluations de ces designs, trouvent un effet petit mais positif (0,15 sigmas) sur les performances, plus important lorsqu'une école est resté plus de 4 ans sous ce dispositif(0,25 Sigma), ils repèrent et indiquent les designs les plus efficaces. On peut à l'inverse insister sur l'intérêt pour chaque école de concevoir les stratégies d'amélioration adaptées à sa situation. C'est la stratégie du *shool improvement,* dont l'évolution offre un bon exemple des progrès accomplis en matière d'utilisation des mesures de performances des élèves. Au début des années 90, deux courants de recherche sur les établissements se regardaient en chiens de faïence au sein d'une même association. L'un (*school effectiveness*) utilisait des méthodes quantitatives pour comprendre ce qui distinguait les écoles efficaces des autres (approche transversale, dont (Grisay, 1997, Dossiers Education et Formations, n° 88, disponible sur le web in « Archives d'Aletta Grisay/ sciences de l'éducation/ rapports de recherche/livres »)) demeure l'exemple le plus abouti en France). L'autre utilisait des méthodes qualitatives, fondées sur l'expérience et la participation des acteurs, pour comprendre ce qui aidait les écoles à s'améliorer (approche longitudinale). Des chercheurs néerlandais ont proposé de coupler ces deux approches au sein d'un dispositif qu'ils ont appelé DASI (*Dynamic Approach to school improvement*) dans lequel les connaissances sur les facteurs d'efficacité des écoles sont mobilisées pour leur amélioration, dans lequel chaque établissement développe sa propre stratégie d'amélioration, mais avec l'aide de chercheurs familiers des études *school effectiveness* . Une évaluation expérimentale de DASI a comparé l'amélioration de deux groupes de 20 écoles défavorisées tirés au sort, le groupe expérimental utilisant DASI, le groupe ce contrôle utilisant l'approche classique *du school improvement*. Ils ont mis en évidence que, dans le groupe utilisant DASI, les performances (en maths) des élèves s'étaient accrues un peu plus que dans le groupe témoin, mais surtout que l'influence de l'origine sociale sur le score de maths s'y était davantage réduite *(Charalambous, E. et al. 2018, Promoting quality and equity in socially disadvantaged schools, a group randomization study, Studies in Educational Evaluation, n° 57).* Anthony Bryk *(Education et Didactique, 2017, vol2, 11*) propose, lui, la mise en place de réseaux d'amélioration, regroupant des établissements qui ont décidé de jouer sur un même facteur pour s'améliorer. La recherche (ex : Grisay, 1993, 1997) signale en effet que les établissements efficaces se distinguent des autres sur certains seulement des facteurs d'efficacité, de sorte que choisir sur quoi porter l'effort est une décision stratégique qui dépend de l'analyse que chaque établissement fait de ses forces et de ses faiblesses. Notons pour finir qu'aucune des stratégies évoquées ici n'obtient de résultats spectaculaires. Il faut considérer l'amélioration des écoles comme un processus continu, avec des essais et des erreurs. Mais justement : si les écoles disposent d'évaluations standardisées systématiques, elles peuvent s'apercevoir rapidement de la réussite ou de l'échec d'une stratégie donnée, la modifier ou l'abandonner si besoin.

même ou à ses parents. Informer des performances des établissements		établissements.	

Il est **politique** parce que les acteurs n'aiment pas la pression que cela peut mettre sur eux. *Educational Researcher* a consacré un n° spécial (2015, n° 44.2) à la perception par les enseignants américains de l'utilisation de mesures de valeur ajoutée (*value adde*) pour les évaluer. Il en ressort que les enseignants américains font davantage confiance au jugement de leur chef d'établissement (le plus souvent, ce genre d'observation par le chef d'établissement est là-bas très normé, doit reposer sur une liste de critères connus) qu'aux mesures de valeur ajoutée, calculées à partir de l'écart entre deux épreuves standardisées, en début et en fin d'année. Or, il y a toutes chances pour que les secondes soient plus fiables que les premières (cf. ci-dessus l'étude de Chetty et al.). Une part de l'explication est une résistance culturelle à ce qui est chiffré, résistance dont le rapport de la Cour des Comps donne plusieurs croustillants exemples. Probablement, cette réticence aux chiffres est particulièrement forte en France, où les chiffres sont associés à la technique, qui est basse, alors que les jugements qualitatifs relèvent de l'esprit, qui, lui, est élevé : les chiffres sont réducteurs, ils font fi des intentions, ne rendent pas justice à la subtilité, à la multidimensionnalité de ce qui est transmis. On peut ironiser sur le sujet : les enseignants ne trouvent pas que les chiffres sont réducteurs quand ils notent les devoirs des élèves ou leurs examens. Quand il s'agit d'utiliser les évaluations pour noter les élèves, et non pour évaluer les enseignants ou les établissements, on n'entend pas les critiques selon lesquelles mesurer les effets à court terme ne prend pas en compte les effets à long terme, mesurer les connaissances ne prend pas en compte le développement personnel et moral et autres critiques *ejusdem farinae*.

Cette réticence, qui n'est pas spécifique à la France, doit être prise au sérieux. A mon sens, elle ne peut l'être en catéchisant les enseignants sur les vertus des épreuves standardisées. Elle pourrait l'être en comparant la justesse des épreuves standardisées avec celle des notes, ou des appréciations qualitatives, des enseignants. Autrement dit, il faut faire fond sur la rationalité des enseignants et comparer le pouvoir prédictif des notes ou des jugements qualitatifs avec celui des résultats à des épreuves standardisées sur la suite de la carrière scolaire, voire sur la vie, des élèves. Par exemple, quelle corrélation y a-

t-il entre le score PISA ou CEDRE d'un élève et la moyenne de ses notes ? Lorsqu'il n'y a pas corrélation, quel est le meilleur prédicteur ? On peut aussi, dans la veine de la recherche de Chetty et al. , vérifier qu'un élève tire des avantages à moyen ou long terme d'avoir un professeur performant selon des épreuves standardisées.

Elle pourrait l'être aussi en amenant les acteurs à l'idée à rompre avec l'idée que l'inefficacité (dans une classe, une école) est honteuse, procède forcément de fautes criantes, auxquelles il suffirait de remédier pour accéder immédiatement à l'excellence. Ce n'est pas l'histoire que racontent les multiples évaluations de multiples dispositifs aux Etats-Unis. Elles diagnostiquent souvent des effets relativement faibles, parfois des échecs (certains *school designs,* par exemple), mais ce qui frappe dans ce pays est la volonté d'essayer, sur la base de recherches toujours plus sûres, de se remettre en cause, une approche dont, à mon sens, ce système scolaire tirera des bénéfices toujours plus nets.

2. Quels sont les biais des tests diagnostic ou bilan ?

Cette question, à vrai dire, est un peu biaisée. D'une part elle emploie une distinction (diagnostic/bilan) que je n'ai vue évoquer qu'en France. Elle a été , je crois, inventée pour persuader les acteurs que les évaluations nationales (dites diagnostic) ne pourraient valablement être utilisées pour faire le bilan des effets au cours d'une année scolaire d'un enseignant ou d'un établissement (ou pour prévenir l'idée que les évaluations généralisées de début d'année scolaire pourraient, *horresco referens,* être utilisées comme tests initial pour le calcul de mesures de valeur ajoutée) D'autre part, la question semble supposer que, par nature, les évaluations artisanales seraient exemptes de biais. Or, ce n'est pas vrai. Il faut distinguer les biais des épreuves elles-mêmes (alors qu'ils sont de compétences égales sur la dimension que l'on veut mesurer (les maths, par exemple) certains élèves les réussissent mieux que d'autres) et les biais de correction (Depuis longtemps, la docimologie a comparé les notes mises par des correcteurs différents aux mêmes copies, et l'on a montré par exemple qu'il fallait, si je me souviens bien, 120 corrections pour que la note d'une dissertation de philosophie soit stabilisée et une quarantaine pour qu'une copie de maths le soit). On sait aussi que certains de ces biais sont systématiques (les enfants beaux sont surnotés, comme les enfants de milieu social favorisé), Une étude ancienne d'Aletta Grisay sur les écoles primaires belges a montré que le jugement des inspecteurs, autre exemple d'évaluation artisanale, surévaluait, par rapport aux évaluations fondées sur des épreuves standardisées, les écoles des beaux quartiers par rapport à celles des quartiers populaires.

Cependant, comme toutes les mesures, y compris celle de la longueur d'une table, la mesure des performances scolaires par les épreuves standardisées est affectée d'un certain degré d'imprécision. Celle-ci se double d'une erreur d'échantillonnage lorsqu'il s'agit d'une évaluation sur échantillon, comme le sont CEDRE ou PISA. L'avantage des épreuves standardisées est que l'on peut mesurer cette imprécision, en calculant un intervalle de confiance. Par exemple, le score moyen des élèves français de l'échantillon en sciences à PISA 2015 est de 495, mais le rapport PISA donne aussi l'intervalle de confiance, soit que cette valeur signifie en réalité qu'il y a 95% de chances que le score de la population des élèves français de 15 ans soit compris entre 491 et 499 (PISA 2015 results, vol1, p 69). Ce rapport donne aussi une table où l'on peut lire quels pays ont surement un score supérieur ou inférieur à celui de la France et aussi les pays pour lesquels la différence entre leur score et celui de la France est trop faible pour que l'on puisse être sûr que l'un est supérieur à l'autre.

De même, la DEPP dans son rapport sur les performances en mathématiques à la fin du CM2, indique un score de 248 pour 2014, de 250 pour la précédente mesure (2008) et que cet écart est non significatif (MEN- DEPP, 2015, CEDRE 2014, Mathématiques en fin d'école, Dossiers Education et Formations, n° 208, p.20). On peut cependant regretter que la mention de l'erreur de mesure ne soit pas systématique dans les publications de cette direction. Par exemple, elles sont absentes dans la NI 20 de 2016, consacrée à l'évolution de 2003 à 2015 de la performance en langue en fin d'école.

Cependant, il s'agit là d'erreurs et non de biais, qui sont des erreurs systématiques. Aux Etats-Unis, il existe une longue tradition de critique des tests au motif qu'ils défavoriseraient certains élèves, par exemple parce qu'ils ne feraient pas droit à la culture de certaines minorités (Mac Neil, 2000, *Contradictions of school reform, The educational cost of standardized testing*, Routledge ou au motif que les enseignants sont attachés au « développement intellectuel et social de chaque élèves » et ne sont pas des « maximisateurs de scores académiques ». Ces deux critiques portent peu en France. La seconde parce que la tradition scolaire française se préoccupe assez peu de développement personnel et social et est fort attachée à la maximisation des scores au brevet, au bac ou au concours d'entrée à Normale Sup'. La première parce que le souci d'équité vis-à-vis des différentes cultures serait jugé anathème en France comme signe de multiculturalisme. Comme, aussi, les tests de la DEPP sont sans enjeu pour les enseignants ou les établissements, ces derniers n'ont pas encouru ce type de critique.

En revanche, les épreuves des évaluations internationales ont été accusées de refléter une « culture anglo-saxonne » et de favoriser les élèves de cette aire culturelle ou les élèves anglophones *(Ex : Blum et Guérin-Pace, 2000, Des chiffres et des lettres, Fayard).* En réalité, PISA se soucie fortement d'éviter ces biais, entre autres en demandant à tous les pays de lui proposer des exercices pour construire ses épreuves et en éliminant lors du pré test les exercices qui pourraient présenter un biais de ce type. Certains faits évidents contredisent l'existence d'un tel bais, par exemple, le fait que des pays parlant la même langue (France vs Québec) ou appartenant à la même aire culturelle (Canada vs Etats-Unis) aient à PISA des résultats fort différents ou encore que les pays les plus performants à PISA appartiennent à des aires culturelles différentes (Finlande vs Singapour, Japon). De façon plus systématique, lors de l'édition 2009, PISA a demandé aux pays de noter les items de l'épreuve de lecture quant à leur capacité de refléter le degré de « préparation pour la vie » des élèves, quant à leur authenticité et à leur pertinence pour des élèves de 15 ans (PISA 2009 results, vol1, p 36). Ils ont ensuite comparé le classement des pays selon qu'il était établi à partir de ces items (en ordonnée) ou de l'ensemble de l'épreuve PISA (en abscisse). Pour la grande majorité des pays, le classement est le même. Parmi ceux pour lesquels il diffère, ceux qui sont mieux classés selon l'ensemble de l'épreuve sont aussi nombreux que ceux qui sont mieux classés selon leurs items préférés. Pour un seul pays, la différence de rang est forte et c'est un pays mieux classé selon l'ensemble du test PISA que selon ses items préférés.

3. Evaluer un système éducatif de manière rigoureuse et contextualisée.

Je ne suis pas sûr que la priorité soit celle qu'indique le rapport de la Cour des Comptes, d'élaborer un système national d'évaluation plus systématique. Nous connaissons les performances de notre système scolaire par les études sur échantillon, les évaluations internationales (PISA , PIRLS, TIMSS4,TIMSS 8, TIMSS advanced) et les évaluations de la DEPP (CEDRE).

Nous savons par ailleurs que les réformes de structure favorables à l'efficacité ont été faites (tronc commun au collège, limitation du redoublement) ou sont en cours (atténuation de la logique de filières au lycée) , et que donc l'origine de nos difficultés est à chercher dans l'ordinaire des classes (Pour le secondaire, cf. *Le Mener et al., 2017, L'accroissement de l'effet de l'origine sociale sur les performances scolaires, par où est-il passé ?, Revue Française de Sociologie, 58(2)).*

C'est donc à améliorer cet ordinaire que doit servir l'évaluation. Ce sont donc les classes qu'il faut évaluer[5].

S'agissant de l'empan de cette évaluation, je rejoins le rapport de la Cour, et soutiens (*Meuret, in Esprit, décembre 2012*) une évaluation rigoureuse de la progression des élèves vers le socle commun.

Les évaluations nous le disent : le secondaire doit diminuer les inégalités sociales, mais le vrai maillon faible du système est le primaire, qui a un très sérieux problème d'efficacité : A PIRLS 2016 (lecture CM1) le score moyen des élèves en France est 30 points en dessous de la moyenne de l'UE ou de l'OCDE (sur une moyenne OCDE de 500,cela représente presque ce qu'un éèvel moyen apprend en un an). Ce score est, avec celui de la Belgique (fr), le plus faible de l'OCDE. Il a baissé de 14 points depuis 2001 (DEPP, NI 17.24) alors que la tendance internationale est à la hausse[6]. TIMSS 2015 (CM1) , avec un score de 488 en maths et de 487 en sciences, la France a les performances les plus faibles de toute l'UE en maths et de toute l'UE sauf Chypre en sciences (DEPP, NI 16.33).

Il faut d'autant plus évaluer le trajet vers le socle commun qu'il y a quelques indices de ce que le socle commun a peut être produit de bons effets (le score des plus faibles, à PISA, a cessé de baisser depuis 2009 en maths et en sciences[7]).

Il faut l'évaluer dans toutes les classes, tous les ans et que les résultats aient des conséquences pour les élèves et les enseignants. L'objectif est de repérer les classes ou les élèves progressent moins qu'attendu vers la maitrise du socle (ce qui demandera d'imaginer des objectifs propres à chaque niveau) et à chaque compétence, et de mettre en place aide et stimulation (aide au diagnostic et à l'utilisation des données de l'évaluation, , aide de pairs, formations ciblées) pour les enseignants dont, au moins deux ans de suite, un trop grand nombre d'élèves n'auraient pas faits, dans un domaine ou l'autre, ou dans les domaines jugés prioritaires , les progrès nécessaires pour cela.

[5] Les classes, et pas seulement les enseignants. Cf. les problèmes de discipline dans les classes françaises, voir in PISA 2015 et Meuret,D., 2017, La mauvaise discipline dans les classes française, notes du Conseil scientifique de la FCPE.

[6] Cette hausse devrait nous épargner les considérations habituelles sur les ravages éducatifs de la modernité et sur l'évolution délétère de l'environnement de l'acte éducatif (par ex : Ferry, L., 2003, Lettre ouverte à ceux qui aiment l'école ; Gauchet, M. et al., 2008, Conditions de l'éducation).

[7] En maths, le premier décile est de 389 en 2003, il baisse fortement à 369 en 2006, 361 en 2009, puis reste à peu près stable (361 en 2009, 365 en 2012, 364 en 2015). Il est à peu près stable depuis 2006 en sciences. Il est un peu plus erratique en compréhension de l'écrit : forte baisse de 2000 (381) à 2006 (346), puis 352en 2009, 358 en 2012, rechute à 344 en 2015.

Une question délicate est de savoir si cette évaluation devrait porter sur les cinq domaines du socle (Langages pour penser et communiquer, Outils pour apprendre, Formation de la personne et du citoyen, systèmes naturels et systèmes techniques, représentation du monde et de l'activité humaine), ce qui serait logique, ou sur les trois domaines qui sont souvent considérés comme les trois fondamentaux (par PISA, par le NAEP aux Etats-Unis, en Grande Bretagne par les évaluations aux *key-stages*) : compréhension de l'écrit, maths, sciences, ce qui serait plus facile

Un tel dispositif est précisément celui dont on a voulu prévenir l'existence en supprimant les évaluations nationales standardisées nationales exhaustives, après qu'on les ait situées en début et non en fin d'année scolaire. Il rencontrerait évidemment des oppositions fortes, que pourraient prévenir une ou deux années blanches (sans conséquences) pour laisser le temps aux acteurs de se familiariser avec ce type de données et apprendre à les utiliser.