



## **Complément à l'étude**

*Les évaluations standardisées des élèves en Europe: objectifs, organisation et utilisation des résultats* EACEA; Eurydice

---

# **LES EFFETS THÉORIQUES ET RÉELS DE L'ÉVALUATION STANDARDISÉE**

Nathalie Mons

Maître de conférences en sciences de l'éducation à l'université Pierre-Mendès-France (Grenoble 2, France) et professeur invité au *London Institute of Education* (Royaume-Uni)

Les opinions exprimées dans cet exposé sont celles de l'auteur et ne reflètent pas nécessairement celles de la Commission européenne.

Août 2009

Réseau Eurydice



## TABLE DES MATIÈRES

---

<b>Table des matières</b>	<b>3</b>
<b>Introduction</b>	<b>5</b>
<b>I. Les effets attendus: cadre théorique politique et pédagogique de l'évaluation standardisée</b>	<b>7</b>
A. Le cadre théorique politique de l'évaluation standardisée: un instrument de pilotage des systèmes éducatifs	7
1. Évaluation standardisée et rénovation de la gestion de l'État: New Public Management et le mouvement de la <i>Policy Evaluation</i>	7
2. La logique économiste de l'évaluation standardisée et le courant pragmatique de la <i>school effectiveness</i>	9
B. Le cadre théorique pédagogique de l'évaluation standardisée: des modèles encore en construction	11
<b>II. Les effets réels de l'évaluation standardisée sur les performances des systèmes éducatifs</b>	<b>17</b>
<b>III. Évaluation standardisée et processus éducatifs: les comportements des enseignants, cadres intermédiaires, élèves et parents face au <i>testing</i></b>	<b>25</b>
A. Enseignants: une tendance à la résistance à la culture de l'évaluation standardisée quantitative	25
B. Les cadres intermédiaires du système éducatif: une prise en main de l'outil	31
C. Les élèves et le fardeau des tests	32
D. Les parents: une réception positive du <i>testing</i> mais des attentes pour l'école qui vont au-delà	34
<b>Références</b>	<b>39</b>



## INTRODUCTION

---

L'évaluation standardisée n'est pas une pratique nouvelle dans les systèmes éducatifs des pays développés. Pour autant, en cette fin de la première décennie 2000, l'outil, qui se caractérise par la conception, l'administration et la correction d'épreuves d'évaluation uniformisées, suscite de nombreuses polémiques, tant scientifiques que médiatiques (Haney, 2000; Hanushek et Raymond, 2003; House of Commons, 2007...). Si l'instrument est fortement questionné, c'est parce que désormais loin d'être un simple outil neutre d'évaluation des compétences et des savoirs des élèves, il s'est imposé comme un instrument majeur d'une nouvelle régulation politique des systèmes éducatif. **Alors qu'hier l'évaluation standardisée centrée sur la mesure des apprentissages s'intéressait principalement à l'élève, aujourd'hui son champ d'intervention est beaucoup plus large et met en lien le pédagogique – sa terre d'élection traditionnelle – et le politique dont il est devenu un outil de pilotage** (Behrens, 2006).

**Ces dispositifs sont ainsi désormais à la croisée des nouvelles tendances qui caractérisent les politiques éducatives développées dans les pays de l'OCDE depuis les années 1980** (Mons, 2007). Pour en affiner la compréhension, l'évaluation standardisée doit, en effet, être mise en relation avec quatre évolutions récentes de nos systèmes éducatifs: a) **la centration sur une mesure quantitative des apprentissages et la priorité donnée à des objectifs cognitifs au détriment d'objectifs de socialisation larges** (Osborn, 2006), en lien avec le développement du concept de compétences dans la veine économiste de la théorie du capital humain; b) le développement **d'un nouveau contrôle social des enseignants et des écoles par les responsables administratifs de l'éducation** au sens large (districts, municipalités, administrations déconcentrées, régions suivant les pays) dans le cadre le plus souvent de réformes de décentralisation et d'autonomie scolaire (Maroy, 2008); c) **l'évolution de la répartition des pouvoirs** entre les acteurs centraux ou fédéraux et les responsables locaux qui voient ainsi leurs marges de manœuvre fortement encadrées (Broadfoot, 2000) et, enfin, d) **le développement de la redevabilité de l'École envers le grand public**, en général, et les parents, en particulier, dans le cadre de nouvelles relations entre le politique, l'État, l'administration, d'un côté, et la société civile, de l'autre, ces nouvelles relations étant sous-tendues par l'avènement d'une «démocratie du public» dans laquelle la définition du bien commun n'est plus le seul monopole des dirigeants légitimes (Manin, 1996).

Au confluent de ces multiples influences – influences auxquelles de nombreux acteurs au sein des systèmes éducatifs essaient de résister: les enseignants contre la nouvelle culture d'évaluation quantitative, les autorités locales contre la main mise des responsables politiques centraux ou fédéraux, l'École-institution forteresse contre l'immixtion des parents –, l'évaluation standardisée, élevée au rang d'outil politique, ne pouvait être qu'une politique fortement questionnée.

Fondée sur la rhétorique de l'efficacité scolaire – le développement des tests doit permettre d'améliorer le fonctionnement général des systèmes éducatifs et les apprentissages des élèves en particulier – cette politique se doit d'être évaluée sur ce terrain pour apporter quelques lumières scientifiques dans ce débat public vif. **Éclairer la question des effets des dispositifs d'évaluation standardisée est donc l'objectif de ce rapport.** Pour ce faire, **une première section s'intéressera aux effets théoriques attendus de l'évaluation standardisée à la fois comme outil de pilotage (ou de régulation) des systèmes éducatifs et comme instrument pédagogique d'une amélioration des acquis individuels des élèves.** Quels cadres théoriques et conceptuels – à la fois politiques et pédagogiques puisque l'instrument possède cette dualité – ont été pensés pour expliquer en quoi l'évaluation standardisée peut améliorer les performances des systèmes éducatifs?

Au-delà de ces cadres théoriques multiples, **dans un second temps, ce rapport s'intéressera à une analyse empirique des effets réels de l'évaluation standardisée tant en termes d'efficacité et d'égalité scolaires que d'efficience.** Car, au-delà de l'amélioration moyenne des performances des élèves – qu'appréhende le concept d'efficacité –, il faut s'interroger sur l'impact de cette politique sur les élèves issus de milieux défavorisés (socialement, ethniquement...) ou souffrant de handicaps. L'évaluation de ces politiques doit également s'orienter vers une analyse coût-efficacité. Combien dépense-t-on pour mettre en œuvre le *testing* pour quels résultats? Pour apporter quelque éclairage sur ces interrogations multiples, nous présenterons une large revue de la littérature scientifique, fondée sur des études de cas nationaux et des comparaisons internationales. **À travers ce champ scientifique extrêmement controversé, nous montrerons que tant en termes d'efficacité que d'égalité scolaires, les effets de l'évaluation standardisée sont difficiles à appréhender car aucun consensus empirique net ne se dessine autour d'un bénéfice de ces réformes.**

Ce résultat ou du moins cette absence de convergence des conclusions des études scientifiques interroge: quels sont les processus en jeu dans les réactions des acteurs locaux du système éducatif qui peuvent expliquer que les résultats des dispositifs d'évaluation standardisée apparaissent suivant les terrains aussi divers <sup>(1)</sup>. **Nous serons ainsi amenés, dans la troisième partie de ce rapport, à décrire les mécanismes associés à l'introduction du *testing* en lien avec les comportements d'une pluralité d'acteurs: les enseignants compris individuellement et collectivement comme équipe pédagogique, les personnels encadrant le système éducatif (chefs d'établissement, responsables d'autorités locales ou supérieures en charge de l'éducation...), les parents d'élèves et les élèves bien sûr... Comment ces acteurs réagissent-ils à la mise en place de dispositifs d'évaluation standardisée? Leurs comportements varient-ils en fonction des caractéristiques des réformes mises en place?**

---

<sup>(1)</sup> Nous aurons déjà montré dans la partie précédente de ce rapport consacrée à la revue des études empiriques que les divergences des résultats des recherches portant sur les effets de cette politique s'expliquent également en partie par des méthodes de recherche variables.

## I. LES EFFETS ATTENDUS: CADRE THÉORIQUE POLITIQUE ET PÉDAGOGIQUE DE L'ÉVALUATION STANDARDISÉE

---

«L'usage des résultats des élèves aux tests dans les systèmes d'*accountability* n'est pas nouveau. Des exemples de rémunération en fonction des résultats, comme l'engouement pour la contractualisation dans les années 1960, apparaissent et disparaissent au gré des décennies. Ce qui est différent dans le mouvement actuel d'*accountability* fondé sur la performance, c'est qu'il s'est généralisé. Comme le notent Elmore, Abelman, and Fuhrman, «ce qui est nouveau, c'est le rôle désormais central que jouent les performances des élèves dans la gouvernance publique» (1996, p. 65)» (Linn, 2000).

Car, au-delà de l'outil pédagogique de mesure des acquis des élèves, l'évaluation standardisée a acquis un nouveau statut politique qui fait d'elle un instrument central dans la régulation des systèmes éducatifs. Cette dualité pédagogique-politique s'appuie sur un substrat théorique développé au double niveau macro et micro. **Dans un premier temps, nous présentons le cadre conceptuel macro-politique qui a pour fondement à la fois, dans le champ général des politiques publiques, le *New Public Management* et le mouvement de la *Policy Evaluation* et, dans le secteur plus restreint de l'éducation, l'économie de l'éducation et le courant de la *school effectiveness*.** L'alliance de ces quatre écoles théoriques et pragmatiques a assigné des objectifs précis macro-politiques à l'évaluation standardisée. **Dans un second temps, nous présentons au niveau micro de l'établissement scolaire et de la classe ce qui est attendu du *testing* comme outil pédagogique d'orientation de l'activité des enseignants et des élèves.**

Il est intéressant de se pencher sur les cadres théoriques des politiques d'évaluation standardisée: en effet la mise en œuvre de ces outils est souvent présentée comme une mesure de bon sens – mesurer quantitativement les acquis des élèves permettrait de les faire progresser. Or cette première section du rapport montrera que ce mouvement du *testing* ne relève en rien de l'évidence mais s'inscrit dans des mouvements de pensée spécifiques liés à un contexte intellectuel mais aussi social et économique particulier, celui de la remise en cause de l'intervention de l'État et de la mise en évidence d'une «crise de l'éducation publique», réelle ou «fabriquée» d'ailleurs (Berliner et Biddle, 1995).

### A. Le cadre théorique politique de l'évaluation standardisée: un instrument de pilotage des systèmes éducatifs

Les rôles et les effets théoriques attendus de l'évaluation standardisée s'inscrivent à la fois dans un mouvement général de rénovation de la gestion de l'État et dans des courants de pensée spécifiquement développés dans le secteur de l'éducation.

#### 1. Évaluation standardisée et rénovation de la gestion de l'État: *New Public Management* et le mouvement de la *Policy Evaluation*

Le *New Public Management* (NPM), né aux États-Unis, s'impose à partir des années 1970 comme un nouveau courant d'analyse du fonctionnement des administrations dans un contexte de remise en cause de l'action de l'État. À cette époque en effet, différents facteurs économiques (stagflation, mondialisation des économies...) viennent ébranler la conception macroéconomique keynésienne de l'intervention de l'État. S'installe progressivement dans les administrations de l'économie et des finances une orthodoxie budgétaire influencée par les thèses monétaristes (Jobert 1994 et Siné 2006, cités par Pons, 2008). Dans ce contexte, l'objectif principal est d'assainir les dépenses publiques, donc de rationaliser l'action de l'administration.

Le secteur de l'éducation n'échappe pas à ce contexte de crise – à la fois crise des ressources mises à disposition et crise de l'image de ce service public battue en brèche par la publication de rapports en dénonçant les défaillances (voir par exemple le rapport américain de 1983 *A nation at risk* ou les *blacks papers* anglais des années 1980).

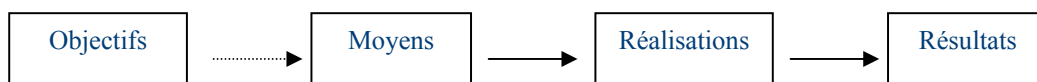
**Face à ce contexte de crise, le *New Public Management* va développer un ensemble de lignes directrices permettant de guider la rénovation de l'administration dans le sens d'un meilleur coût-efficacité.**

**En éducation, parmi ces nouvelles orientations, quatre principaux piliers du *New Public Management* vont donner à l'évaluation standardisée un rôle très spécifique dans les nouveaux agencements institutionnels:**

- les productions des services publics sont mesurables;
- elles doivent donc être mesurées grâce à des outils spécifiques dont la validité doit être scientifiquement assurée;
- les acteurs publics opérationnels qui bénéficient de davantage d'autonomie sont redevables de leurs actions devant les gestionnaires du système (c'est le modèle de l'évaluation gestionnaire) et/ou devant les citoyens (modèle de l'évaluation démocratique);
- l'organisation publique doit être régulée par les résultats (*outputs*) et non plus exclusivement par des contrôles procéduraux orientés sur les ressources (*inputs*).

Dans ce contexte, en éducation, les fonctions et les effets attendus de l'évaluation standardisée sont donc multiples. Elle doit tout d'abord permettre une évaluation des acquis des élèves, indicateur de la «qualité» de la production du service éducatif. Elle sert également à faire le lien entre les acteurs responsables de l'opérationnalisation de ce service (les enseignants, les écoles et suivant les configurations institutionnelles les autorités locales en charge de l'éducation) et les responsables, souvent nationaux, en charge de sa conception. En ce sens, elle s'impose comme un outil directeur à la fois guidant les activités des acteurs opérationnels et informant le niveau hiérarchique supérieur sur les performances de ces agents (Wöessmann, 2007). Enfin, dans le cadre d'une nouvelle redevabilité envers la société civile, il est attendu de l'évaluation qu'elle serve d'instrument d'information en direction des acteurs externes à l'école, en général, et envers les parents, en particulier.

**Plus généralement, la fonction de l'évaluation standardisée s'inscrit également dans le mouvement de la *Policy Evaluation*** (Spencehauer 2003, cité par Mons et Pons, 2009). S'appuyant sur le modèle des sciences expérimentales, ce courant de pensée pragmatique se propose de découper en trois séquences les dispositifs d'action politique. Les décideurs doivent tout d'abord clairement expliciter les objectifs de référence de leurs politiques, afin de pouvoir dans un deuxième temps choisir les méthodes de mesure et les indicateurs pertinents et enfin déterminer, par la comparaison, les effets d'une politique. Ce triptyque «projet-mesure-analyse des effets» est censé fonder toute évaluation de politique.



**L'évaluation standardisée s'inscrit bien dans cette vision d'une action publique séquentielle et rationnelle – l'élaboration et la mise en œuvre des politiques publiques obéissent à un enchaînement d'étapes logiques –, les objectifs des politiques publiques peuvent être atteints**



si les bons moyens sont mis en face d'objectifs clairement définis. Ce courant de pensée reflète une vision *top-down* de l'action publique dans laquelle les politiques publiques seraient mises en œuvre telles que décidées dans leur cadre légal – réglementaire ou législatif – par les acteurs locaux. Le *testing* est ici un outil majeur de l'étape finale («résultats») de l'évaluation des politiques publiques et doit donc permettre la confrontation entre les objectifs initiaux et les performances atteintes au final.

**Ces deux courants de pensée – le *New Public Management* et la *Policy Evaluation* – sont cependant questionnés par certains chercheurs.** Cet intérêt accru pour une approche de l'action publique par les résultats et une centration sur l'évaluation des productions des services publics, dont celui de l'éducation, interrogent en effet. Les études qui s'intéressent au processus mis en œuvre dans le cadre de l'évaluation standardisée, et notamment aux comportements des acteurs locaux (enseignants, chefs d'établissement...) montrent comme nous le verrons dans la suite de ce rapport que cette vision théorique *top-down* s'avère imparfaite: les acteurs locaux développent en effet des stratégies d'adaptation aux injonctions institutionnelles qui peuvent ne pas aller dans le sens des objectifs attendus (par exemple, exclusion des élèves en difficultés des tests pour améliorer artificiellement le niveau moyen de l'école, entraînement intensif aux évaluations au détriment d'un enseignement de fond...).

Au-delà de la rénovation de la gestion publique, les politiques d'évaluation standardisée sont aussi portées par deux mouvements qui se développent spécifiquement en éducation.

## **2. La logique économiste de l'évaluation standardisée et le courant pragmatique de la *school effectiveness***

Pour les économistes, dans la lignée de la théorie du capital humain, l'éducation est non plus considérée comme un coût mais comme un investissement dont va résulter un capital. L'agrégation des compétences individuelles va refléter partiellement la force économique des pays.

Les liaisons entre efficacité économique et acquisitions scolaires ont été mises en lumière, entre autres, par Barro (1999). Même si le questionnement sur la relation entre éducation et croissance économique reste toujours ouvert, le chercheur américain a montré que la mesure du capital humain à travers les scores aux évaluations scolaires explique mieux la progression des richesses nationales qu'une mesure par les carrières scolaires.

Dans cette perspective, les interrogations sur la réalité des apprentissages, auxquelles les évaluations standardisées cherchent à répondre, sont donc fondamentales. «Dans la logique d'une économie centrée sur «l'innovation» ou «la connaissance», un discours prêchant la nécessité d'investir dans les ressources humaines se fait de plus en plus entendre, qui se traduit dans les demandes de voir le «niveau général d'éducation» s'élever» (Maroy, 2005).

Mais, selon les économistes, cette amélioration de la qualité de l'éducation ne doit pas se faire à n'importe quel prix et ce d'autant plus que la massification de l'éducation par un accès quasi-universel au secondaire supérieur dans les pays de l'OCDE pèse sur le budget des États. Face à ces contraintes budgétaires, les économistes mettent en avant leur capacité à apporter des solutions pour «produire» plus à moindre coût – c'est-à-dire éduquer mieux avec des budgets stables voire moindres –, en définissant les paramètres institutionnels sur lesquels les politiques doivent jouer pour optimiser le système. Pour ce faire, l'analyse économique utilise le paradigme de la fonction de production, importée du monde de l'entreprise, l'entreprise combinant des facteurs de production (les *inputs*) pour élaborer un produit, l'*output*. Appliqué à l'éducation, ce modèle induit la transposition suivante: l'école

assemble différents *inputs* (matériels pédagogiques, professeurs aux caractéristiques particulières (nombre d'années de formation, ancienneté, etc.)...) pour produire de l'éducation, qui peut être évaluée d'un point de vue quantitatif (nombre d'années de scolarisation, par exemple) ou qualitatif (mesure des acquis des élèves). Aussi, l'économiste a-t-il besoin de données relatives aux *inputs* et aux résultats des systèmes éducatifs pour faire «tourner» ses modèles statistiques et mettre en évidence les configurations institutionnelles efficaces. Les données relatives aux résultats qualitatifs des écoles (acquis des élèves) sont acquises à travers la mise en œuvre de tests standardisés. Les économistes prônent donc un nouveau regard sur le fonctionnement des systèmes éducatifs appuyés par une régulation par les résultats (Wöessmann, 2007).

**Les choix de stratégies et d'organisation au niveau de l'établissement sont également cruciaux quant aux résultats des élèves. C'est le fameux slogan «School matters» («L'école compte») du courant de la *school effectiveness*.** La qualité du leadership, de la discipline scolaire, des relations entre enseignants et élèves... sont des facteurs fondamentaux de la réussite scolaire. Empiriquement les études de ce courant montrent que des écoles accueillant des publics identiques en termes de catégories socioprofessionnelles des parents peuvent présenter des résultats scolaires fort différents. D'où l'idée associée aux nouvelles politiques d'évaluation standardisée de rendre les écoles redevables en interne (auprès des autorités qui en ont la charge) et en externe auprès de la société civile.

Les tenants de l'approche économiste appliquée à l'éducation en soulignent cependant eux-mêmes les limites. Les conclusions issues des études fondées sur des fonctions de production sont difficiles à mettre en œuvre. «Ce modèle du changement technique où la science définit la voie à suivre, à laquelle les «exécutants» n'auront qu'à se rallier, peine à s'appliquer à l'éducation. Certes, les décideurs politiques peut s'appuyer sur les recherches pour décider d'éradiquer telle ou telle pratique dont on aurait démontré les effets nocifs (le redoublement ou les classes de niveau par exemple). Mais pour le reste... On connaît la faible portée des directives insufflées d'en haut, vue l'autonomie importante dont jouissent les enseignants dans leur classe [...]. On sait aussi que l'efficacité d'une pratique pédagogique est souvent relative à un contexte (efficace avec tel type de public ou à tel niveau scolaire seulement), ce qui rend les évaluations inévitablement indexées aux conditions dans lesquelles elles ont été réalisées» (Duru-Bellat et Jarousse, 2001).

**Au total, il apparait, à travers la définition de ces cadres théoriques politiques, que l'évaluation standardisée y joue un double rôle essentiel: instrument de régulation, elle permet l'articulation entre différentes politiques éducatives emblématiques des réformes d'envergure entamées dans la très grande majorité des pays de l'OCDE (Mons, 2007); outil de mesure, elle sert à l'évaluation de ces mêmes réformes.**

En effet, au-delà de son rôle traditionnel dans la mesure des acquis des élèves, l'évaluation standardisée apparait désormais comme un outil présentant de multiples facettes. **Instrument d'information** générateur de données quantitatives aisément comparables, elle appuie les politiques de redevabilité interne à l'institution (contrôle des écoles par les autorités en charge de l'éducation) et de redevabilité externe en direction de la société civile (informations sur les performances des élèves, notamment en lien, pour les parents, avec les politiques de libre de choix des établissements). **Outil de mise en cohérence impérative des actions micro des acteurs locaux et macro des décideurs politiques**, l'évaluation standardisée permet également d'imposer, dans des systèmes historiquement ou nouvellement décentralisés, des lignes directrices cognitives plus efficaces que les standards de contenus sur lesquels elle est le plus souvent assise. Au travers du développement du *testing* se joue donc une nouvelle répartition du pouvoir entre les écoles, les autorités locales en charge de l'opérationnalisation du service d'éducation et les décideurs-concepteurs rattachés au niveau central

ou fédéral (Broadfoot, 2000). L'instrument joue donc aussi le double **rôle d'outil de direction cognitive** (définition claire et imposée des contenus d'enseignement jugés prioritaires) et **de contrôle social sur les agents opérationnels**. Enfin, les nouveaux dispositifs d'évaluation standardisée, fondés désormais le plus souvent sur une redevabilité des écoles (par opposition à la responsabilisation des autorités supérieures en charge de l'éducation) s'impose comme un **outil de transfert de responsabilité**: alors qu'hier l'élève était érigé au rang de premier acteur de la réussite scolaire et que le politique se devait de rendre des compte sur les réformes engagées, l'accent est désormais mis en priorité sur la redevabilité des écoles qui, en contrepartie, se voient octroyer davantage de marges de manœuvre. **On voit donc à travers ces rôles multiples que les effets attendus de l'évaluation sont nombreux, en lien avec les politiques de décentralisation, d'autonomie scolaire et de libre choix de l'école.**

Au-delà des effets politiques attendus de l'évaluation standardisée, **l'instrument est aussi censé jouer un rôle dans le domaine strictement pédagogique. Les enjeux se situent alors, non plus au niveau du pilotage des systèmes éducatifs, mais dans l'établissement et la classe.**

## **B Le cadre théorique pédagogique de l'évaluation standardisée: des modèles encore en construction**

Si d'autres politiques éducatives comme l'autonomie scolaire, la décentralisation ou encore le libre choix de l'école ont suscité des recherches théoriques d'envergure bien que toujours très convergées, le cadre théorique pédagogique de l'évaluation standardisée, centré sur les mécanismes en jeu dans l'établissement et la classe, reste à ce jour peu développé. Au niveau des apprentissages, l'objectif de l'introduction du *testing* est bien une amélioration des acquis des élèves. Pour autant de nombreux auteurs (Linn, 2001; Nichols, 2007...) pointent du doigt le fait que les théories expliquent rarement précisément par quels processus intermédiaires l'évaluation standardisée est censée conduire à une progression des acquis des élèves.

Une série d'arguments, assez vagues, sont avancés par les défenseurs de ces dispositifs. Suivant les types d'*accountability* développés – «*accountability* dure» fondée sur l'organisation de tests à forts enjeux (*high stakes testing*) pour les élèves et les écoles sur le modèle américain ou anglais ou modèle que l'on pourrait qualifier d'«*accountability* douce» (le modèle de l'école européenne continentale) –, les modes d'explication diffèrent, voire s'opposent sur certaines dimensions.

**Le modèle de l'«*accountability* dure»** (Goodwin, Eglert *et al*, 2002; Raymond et Hanushek, 2003; Mc Donnell, 2005; Haertel et Heman, 2005, Phelps, 2005; Wöessmann, 2007...) attache aux résultats des tests une série de sanctions et de récompenses qui peuvent avoir de lourdes conséquences pour les écoles, les enseignants ou les élèves (financement de l'école, diplôme ou redoublement des élèves...). **Ce modèle est fondé sur l'existence des mécanismes suivants qui, déclenchés par le *testing* et fonctionnant de concert dans une dynamique positive, permettraient de rendre le système éducatif plus efficace:**

- Des **élèves**, il est attendu un travail plus important et donc, au final, de meilleurs apprentissages du fait à la fois d'une définition plus claire des objectifs scolaires à atteindre, de la désignation de champs d'enseignement jugés prioritaires et de la pression psychologique exercée par le test lui-même puisqu'il conditionne leurs carrières scolaires. En effet, dans ce modèle d'«*accountability* dure», les mesures de redoublement ainsi que l'accès à la certification peuvent dépendre du résultat du test. Le *testing* agirait sur le comportement

des élèves en développant une motivation renforcée pour les études, motivation au moins extrinsèque <sup>(2)</sup>.

- Les **enseignants**, pour leur part, responsabilisés sur les résultats des élèves devant leurs supérieurs hiérarchiques et la société civile par la publication des performances de leur école, travailleraient davantage à la réussite des élèves et amélioreraient leur professionnalisme à la fois par des formations et des échanges entre collègues au sein des écoles, les résultats aux tests donnant lieu à des analyses réflexives dans les établissements. Des performances élevées aux tests dynamiseraient la motivation des équipes pédagogiques tandis que des résultats moyens ou faibles auraient pour conséquence de mettre au travail les moins motivées.
- Pour les **personnels administratifs et politiques encadrant localement les établissements scolaires**, l'évaluation standardisée permettrait une meilleure connaissance des difficultés locales et la mise en œuvre de mesures efficaces de récompenses et de sanctions quand l'établissement ne parvient pas à améliorer ses résultats dans les délais impartis (changement de leadership, restructuration de l'école ou fermeture...).

**Au-delà de l'amélioration de l'efficacité, certains auteurs mettent en avant les effets positifs que les dispositifs d'évaluation standardisée peuvent avoir en termes d'inégalités scolaires** (Grissmer *et al.*, 2000; Hong et Youngs, 2008). En imposant des standards communs à tous les élèves, ces mesures obligent les enseignants à développer des attentes uniformes à leur égard quelles que soient leurs caractéristiques individuelles (handicap, appartenance à une communauté ethnique minoritaire ou à un milieu social défavorisé...) et quels que soient les contextes scolaires dans lesquels ils évoluent (écoles accueillant ou non un public favorisé...). Ceci aurait un effet pédagogique homogénéisateur (même exposition aux contenus d'enseignement, même nombre d'heures utiles, attentes face aux élèves identiques...), qui tendrait à limiter les inégalités scolaires globales et d'origine sociale, en permettant d'améliorer les résultats des élèves appartenant aux groupes sociaux considérés comme défavorisés.

En outre, l'imposition d'un *reporting* aux écoles sous la forme de statistiques décomposées par groupes sociaux ou ethniques permet d'informer parents et décideurs des défaillances de certaines écoles à traiter les populations à risque. Confrontées à des résultats considérés comme faibles mais à des objectifs identiques, les équipes pédagogiques des écoles accueillant des publics difficiles se trouveraient dynamiser par ces nouveaux défis collectifs. Les responsables administratifs locaux apporteraient à ces établissements un soutien plus rigoureux sous la forme de ressources supplémentaires. L'application de sanctions, jusqu'au stade ultime de la fermeture de l'école, pour les établissements qui ne parviennent pas durablement à améliorer leurs performances, permet de faire disparaître des lieux de scolarisation qui semblent vouer à l'échec pour des raisons multiples (médiocrité collective de l'équipe pédagogique, école ghetto...).

C'est pour soutenir cette dynamique imposée aux écoles défavorisées qu'une partie des acteurs politiques réfute l'idée selon laquelle les performances statistiques des établissements scolaires rendues publiques devraient tenir compte, à travers un calcul de valeur-ajoutée contextuelle <sup>(3)</sup>, des

---

<sup>(2)</sup> Par opposition à la motivation intrinsèque qui trouve dans le sujet lui-même ses causes (par exemple l'élève travaille parce que la discipline lui plaît), la motivation extrinsèque est supportée par des causes externes (l'élève va travailler pour le test parce que ce dernier lui permet d'obtenir un diplôme)

<sup>(3)</sup> Statistiques permettant de calculer les performances des établissements en tenant compte des caractéristiques socio-économiques des publics scolaires attendus.

milieux sociaux et ethniques des élèves, les objectifs des écoles devant en effet restés identiques pour tous.

Aux côtés de ce modèle anglo-saxon du *testing* à forts enjeux, **le modèle continental européen, associé à une philosophie d'*accountability* que l'on peut qualifier de douce**, met en avant des arguments qui ne recourent pas, sur toutes les dimensions, ceux qui viennent d'être exposés. Nous présentons ici les deux modèles principaux, qui pour l'un a irrigué la réflexion sur les standards et leur évaluation en Allemagne, en Suisse et en Autriche, dans le cadre du rapport Klieme (2004) et, pour l'autre, fut développé en France par Thélot sous le nom de l'«effet miroir».

**Dans le cadre du rapport «Le développement de standards nationaux de formation: une expertise», Klieme, chercheur à l'Institut Allemand de Recherche Pédagogique Internationale (le DIPF), et son équipe préconisent la mise en place de «standards de résultat» (Klieme, 2004, p. 48). Ces «output standards» décrivent des objectifs d'enseignement dont l'atteinte doit être validée par des évaluations standardisées par opposition aux «standards de contenus» ou «input standards» qui sont axés sur les contenus à enseigner (proches mais en moins détaillés des notions de curricula, programmes scolaires ou plans d'études). Selon l'équipe dirigée par Klieme, ces dispositifs qui allient standards et *testing* auraient une mission centrale: «ils font ressortir sous une forme claire et concise ce qui importe dans le cadre de notre système scolaire. Cette fonction d'orientation est utile tant pour les élèves que pour leurs parents; mais elle sert aussi au professionnalisme des enseignant(e)s et au développement de la qualité au niveau institutionnel. Concrétisés en procédés de test, les standards sont utilisés dans le cadre du monitoring de la formation et de l'évaluation des établissements scolaires. [...] Leur objectif est d'analyser les effets (et les effets secondaires) des méthodes pédagogiques et de permettre ainsi un agir professionnel, rationnel. Les enquêtes sous forme de tests ne sont donc utiles que si elles contribuent à développer le professionnalisme du personnel enseignant ainsi que la qualité de l'école et de l'enseignement» (Klieme *et al.*, 2004, p. 46).**

**Le rapport détaille ensuite ce qu'apportent les standards axés sur les résultats à chacun des acteurs intervenant directement dans le processus pédagogique.** Pour les élèves et leurs parents, les standards de résultats couplés à des tests permettent une meilleure information sur les champs d'apprentissage prioritaires et constituent un outil de dialogue privilégié avec l'équipe enseignante. Par contre, contrairement à l'*accountability* dans sa version dure, qui allie responsabilisation simultanée de l'école et de l'élève à travers des tests à forts enjeux pour les carrières scolaires, le rapport Klieme s'oppose très explicitement à l'usage des standards de résultat dans le cadre des carrières scolaires individuelles. Les données produites dans le cadre de ces dispositifs ne doivent pas être exploitées pour les décisions de passage de classe ou de certification.

**Pour les enseignants**, il est attendu des standards de performances qu'ils donnent une orientation à leurs enseignements («Les standards fournissent aux enseignants un cadre de référence» p. 49) et lancent un signal de responsabilisation («Ils soulignent la responsabilité des enseignants et des écoles pour les résultats d'apprentissage» p. 48). Par contre, contrairement au modèle de l'«*accountability* dure» qui rend les équipes pédagogiques redevables devant leurs concitoyens grâce à la publication des performances des établissements auprès du grand public, le modèle Klieme préconise que «les *feed-backs* [soient] adressés au corps enseignant et aux organes d'un établissement scolaire et non au grand public. [...] Nous voyons dans la mesure des acquis scolaires (compétences des élèves) une opportunité pour les écoles de vérifier les résultats du travail réalisé et d'y réagir professionnellement» (p. 53).

**En France, le modèle de «l'effet miroir», théorisé par Thélot (1994, 1998 et 2003, cité par Pons, 2008), ancien directeur de la direction chargée de la prospective et des statistiques au ministère de l'Éducation, ancien président du HCEE <sup>(4)</sup>, s'oppose aussi à la vision d'une «*accountability* dure». Pour lui, l'évaluation standardisée doit réussir l'«effet miroir» (1998, cité par Pons, 2008). Cela signifie que l'évaluation ne doit pas nécessairement donner des schémas explicatifs de la réussite scolaire mais confronter les acteurs du système éducatif avec les résultats de leurs actions. Il faut renvoyer à ces derniers une image de leurs pratiques afin qu'ils puissent les améliorer si elles ne donnent pas lieu aux résultats escomptés. «Il faut réussir l'effet miroir, donner des résultats sans nécessairement avoir des schémas explicatifs, parce qu'on ne les a pas toujours» (Thélot, 1998). L'«effet miroir» est une modélisation qui s'appuie exclusivement sur une sanction symbolique.**

**Que l'on se positionne dans le cadre d'une responsabilisation dure ou douce, les deux modèles décrits précédemment reposent sur un ensemble de présupposés** qui, d'après certains auteurs (Linn, 2000; Nichols, 2007...) doivent, pour les uns, encore être démontrés et, pour d'autres, entrent en contradiction avec certaines études empiriques déjà menées. Ces présupposés sont les suivants:

- Les tests permettent une bonne mesure de la qualité des enseignements dispensés par les écoles et des compétences et connaissances réelles des élèves;
- Cette mesure, même si elle ne suit pas une logique de valeur-ajoutée contextuelle, n'est pas affectée par des différences entre élèves en termes de motivation, maîtrise de la langue, statuts sociaux et ethniques;
- Les enseignants et personnel des écoles, motivés par un dispositif de sanctions et de récompenses et le regard extérieur posé sur leur travail par les parents, en particulier, et l'opinion publique, en général, cherchent à améliorer leurs enseignements et disposent des ressources personnelles et collectives pour le faire;
- Les résultats des tests aident les équipes pédagogiques à améliorer leurs enseignements;
- Les résultats des tests permettent aux responsables administratifs en charge de l'éducation de faire progresser la gestion des établissements dont ils ont la charge;
- Les écoles peuvent être tenues pour responsable principal des performances des élèves;
- Les parents comprennent la signification des tests et peuvent interpréter les résultats de leur enfant et des écoles dans leur globalité. Dans le cadre d'un système de libre choix de l'établissement, ils utilisent ces indicateurs pour demander une inscription dans l'école qui se révèle la plus performante, stimulant ainsi une concurrence positive entre établissements qui tend globalement à améliorer les résultats du système éducatif.

Comme nous le verrons dans la suite de ce rapport, **un champ désormais large de recherches a montré qu'une partie de ces présupposés, notamment ceux concernant les comportements des enseignants et des parents, ne rejoignent pas la réalité de terrain.**

**Au total, tout comme le cadre politique décrit précédemment, les fondements théoriques pédagogiques de l'évaluation standardisée apparaissent à ce jour devoir être soit explorés**

---

<sup>(4)</sup> Haut Conseil à l'Évaluation de l'École. Cette instance pluraliste, récemment supprimée, avait pour fonction de veiller au bon développement des évaluations en France à tous les niveaux du système éducatif.

**plus avant, soit ré-analysés pour les dimensions qui entrent en contradiction avec les conclusions de certaines recherches empiriques.**

Après avoir analysé dans cette première partie les processus par lesquels l'évaluation standardisée est censée intervenir théoriquement dans la régulation des systèmes éducatifs, nous présentons dans une seconde partie les effets empiriques réels qui sont associés à ces politiques. En particulier, est-ce que les dispositifs de *testing* sont liés à des niveaux de performance moyenne des élèves plus élevés et à des disparités scolaires globales et sociales plus faibles? La revue de la littérature empirique que nous présentons ici cherche donc à présenter une synthèse des effets de l'évaluation standardisée en termes à la fois d'efficacité et d'égalités scolaires. Nous nous interrogeons aussi dans une troisième partie sur les processus mis en jeu par le *testing* en ce qui concerne les enseignants, les élèves, les cadres intermédiaires du système éducatif ainsi que les parents.

Pour réaliser ces deux parties fondées sur des matériaux empiriques, nous présenterons ici un ensemble large de recherches tant quantitatives que qualitatives, émanant de sources multiples (articles scientifiques mais aussi, pour l'analyse des comportements et points de vue des acteurs, des rapports d'inspection, des enquêtes parlementaires, des sondages ...). Nous ne pratiquons donc pas une méta-analyse qui permet de confronter, sur une base standardisée et des critères précis de sélection des études, les résultats de différentes enquêtes. Notre choix s'est davantage orienté vers la présentation d'une large gamme d'études fondées sur des méthodologies diverses. Enfin, dernière caractéristique, cette revue de la littérature est centrée sur les pays développés à la fois européens et nord-américains.





## II. LES EFFETS RÉELS DE L'ÉVALUATION STANDARDISÉE SUR LES PERFORMANCES DES SYSTÈMES ÉDUCATIFS

---

C'est principalement une littérature nord-américaine qui a traité de l'évaluation des dispositifs de *testing* en termes d'efficacité et d'égalité scolaires du fait, entre autres, du vif débat public qui a entouré ces réformes Outre-Atlantique. **Avant de détailler les études qui ont porté sur les effets des politiques d'évaluation standardisée sur les performances des systèmes éducatifs, il nous semble intéressant de commencer par présenter deux études de cas américaines – les réformes du Texas et du district de Chicago.** Ces expériences ont donné lieu à de multiples recherches dont les résultats, bien que portant sur les mêmes données, se sont avérés souvent contradictoires. Ces deux cas sont doublement exemplaires dans la recherche empirique sur les effets du *testing*: premièrement, ils mettent clairement en évidence l'absence de consensus empirique sur l'impact de ces dispositifs – selon les niveaux d'enseignement ou les disciplines observées, les conclusions des études varient fortement; deuxièmement, ces études révèlent un ensemble de chausse-trappes méthodologiques quand lesquelles l'analyse des effets du *testing* tombe encore souvent, enlevant toute validité scientifique aux dispositifs d'évaluation de ces réformes.

**Au début des années 1990, l'État américain du Texas décide d'imposer des tests d'évaluation des acquis aux élèves à la fin des 4<sup>e</sup>, 8<sup>e</sup> et 10<sup>e</sup> années (Haney, 2000). Ce dispositif du Texas Assessment of Academic Skills (TAAS) a pour but de responsabiliser à la fois les écoles et les élèves.** Sur la base des performances de leurs élèves à cette évaluation, les écoles sont, en effet, classées en plusieurs catégories («exemplaires», «bon niveau», «acceptables», «non acceptables»). Ces catégories sont elles-mêmes associées à des récompenses sous forme de financements supplémentaires ou à des sanctions pouvant aller jusqu'à la fermeture de l'école. Les résultats au test impactent aussi fortement la carrière scolaire des élèves puisqu'ils déterminent les redoublements et entrent même en ligne de compte dans l'obtention du diplôme de fin de secondaire supérieur (*high school diploma*).

**Les premières recherches qui visaient à évaluer les effets de ce dispositif sur les acquis des élèves mirent en évidence des résultats positifs.** En particulier, les études de Grissmer et Flanagan (1998, 2001) montrèrent que les résultats au test TAAS avaient fortement progressé durant les années 1990 à la fois en moyenne et pour les différents groupes ethniques (blancs mais aussi hispaniques et afro-américains). Par exemple, selon Hong et Youngs (2008), le pourcentage d'élèves atteignant le seuil minimal requis en dixième année avait bondi de 1994 à 2000, en particulier pour les jeunes noirs-américains. Pour cette population, le pourcentage de jeunes réussissant le test passa ainsi de 28 % en 1994 à 78 % en 2002.

**Ce premier enthousiasme fut rapidement refroidi par une série de recherches qui, à partir des mêmes données liées au test TAAS et de la prise en compte de tests nationaux, ré-analysaient les effets de ce dispositif.** En effet, si les progressions au test local du Texas – le TAAS – pouvaient paraître particulièrement fortes, la prise en compte des résultats des élèves texans au test standardisé américain administré au niveau fédéral – le NAEP <sup>(5)</sup> – ainsi qu'une analyse chronologique plus longue révélaient des progressions qui étaient soit très fortement diminuées soit non significatives selon les disciplines (Treisman et Fuller, 2001). **Ces premiers résultats mirent ainsi en évidence quelques règles basiques méthodologiques à respecter pour permettre une évaluation scientifiquement fondée des dispositifs d'évaluation standardisée. Leurs effets ne doivent tout d'abord pas être analysés à partir des résultats du test mis en œuvre localement: l'évaluation locale ne peut servir à la fois de moyen d'intervention et être son propre outil d'évaluation.**

---

<sup>(5)</sup> National Assessment of Educational Progress.

L'écart entre les résultats à des épreuves externes et ceux des tests locaux a été mis en évidence de façon consistante sur une pluralité d'études (Nichols, 2007): le test local, surtout lorsqu'il est associé à des forts enjeux, voit, avant d'atteindre un plafond, ses résultats fortement progresser lors des premières années de sa mise en œuvre, principalement sous l'effet d'un phénomène d'entraînement intensif au test (le phénomène du «teaching to the test» sur lequel nous reviendrons plus loin). **Une évaluation d'un dispositif de testing qui se limiterait aux résultats des élèves à l'outil d'évaluation locale ne serait en fait qu'une démonstration statistique du phénomène de «teaching to the test» et non pas une analyse du bien-fondé de cette politique éducative. Autre point fondamental de ces premières études: les effets de l'évaluation standardisée doivent être analysés sur le long terme, du fait d'effets probables artificiels lors des toutes premières années.**

Les études portant sur le cas texan ne devaient pas s'arrêter là. **Les effets exceptionnels du nouveau dispositif sur les résultats scolaires des minorités ethniques conduisirent certains auteurs à s'intéresser de près à l'administration du test lui-même. Ils découvrirent ainsi qu'une partie du «mythe texan» (Haney, 2000) s'expliquait par l'exclusion des élèves présentant des difficultés scolaires** (Haney, 2000; McNeil, 2005). En effet, le test fédéral américain NAEP autorise les États à éliminer de leur échantillon les élèves handicapés qui bénéficient de Plans d'Éducation Spécialisés (*Individualized Education plans* – IEPs) et ceux, d'origine étrangère qui présentent une faible maîtrise de l'anglais). De 1992 à 1996, le taux d'exclusion au Texas avait progressé de 8 % à 11 % en quatrième année scolaire et de 7 à 8% en huitième année alors qu'au niveau national les taux régresaient de 8% à 6% en quatrième année et de 7% à 5% en huitième année.

**Des analyses qualitatives, comme celle de Booher-Jennings (2005) montrèrent également que les comportements des enseignants avaient changé depuis l'introduction du test:** ils tendaient à catégoriser les élèves en trois groupes – les «cas sûrs», les «cas nécessitant un traitement», les «cas désespérés» avec une attention particulièrement portée au groupe du milieu dont les progrès permettaient d'améliorer à court terme les résultats de leur école tandis que les élèves les plus faibles recevaient moins d'attention. **Au total, l'expérience texane interrogea sur la consistance des effets positifs de l'évaluation standardisée, leur pérennité dans le temps, les effets pervers dont elle pouvait être assortis ainsi que sur les écueils méthodologiques de certaines recherches qui visaient à en analyser les conséquences.**

**Une seconde expérience – celle du district de Chicago – permet, quant à elle, de mettre en lumière à la fois l'inconsistance des résultats de certaines expériences et les problèmes posés par les contenus des tests.** Au printemps 1995, le district de Chicago – troisième plus grand district des États-Unis marqué par une population socialement et ethniquement défavorisée – décida de mettre fin à la promotion automatique. Il fut donc décidé que pour passer dans la classe supérieure tous les élèves devaient avoir atteint un seuil de compétences minimales à l'*Iowa Test of Basic Skills* (ITBS) lors de leur troisième, sixième et huitième années scolaires (Roderick, Jacob et Bryck, 2002). Ceux qui n'avaient pas le niveau bénéficiaient de soutiens spéciaux (camps d'été...) ou redoublaient s'ils ne parvenaient pas malgré cette aide à améliorer leurs résultats. A la fin de la 8<sup>ème</sup> année, ceux qui avaient échoué au test par deux fois étaient orientés vers des «centres de transition». **Cette politique se solda par des effets majeurs immédiats: les deux premières années, un tiers des élèves de troisième, sixième et huitième années redoublèrent. Après ce premier choc, les résultats s'améliorèrent là encore de façon spéculaire** (Roderick et Nagaoka, 2005). Le nombre de redoublants dans les années scolaires testées baissa très fortement. Par exemple, pour la sixième année, le taux de redoublement atteignait en 1995 37 %, il ne s'élevait plus qu'à 14 % en 1999. De

même, Jacob (2002) montra qu'entre 1990 et 2000, les résultats moyens au test ITBS s'améliorèrent de façon très significative pour les trois années testées.

**Des études plus approfondies montrèrent là encore que les gains n'étaient pas aussi consistants que le laissait envisager la première analyse. Roderick, Jacob et Bryck (2002) montrèrent qu'ils variaient fortement selon les années scolaires envisagées, les disciplines testées et les populations d'élèves analysées.** Ainsi, si en lecture l'introduction du test semblait avoir été positive pour les élèves en difficulté, en mathématiques, les élèves les plus avancés étaient avantagés. Jacob (2002) montra également que les résultats des élèves en mathématiques au test local, le ITBS, étaient très faiblement corrélés avec ceux de l'*Illinois Goals Assessment program*, un test davantage axé sur des exercices de réflexion. Les bons résultats semblaient donc s'expliquer principalement en mathématiques par la réussite aux exercices basiques d'arithmétique et de calcul mental auxquels il était plus facile d'entraîner les élèves. **De la même façon qu'au Texas, le rétrécissement des curricula et le peu de soutien apporté aux élèves en grande difficulté furent mis en évidence par des analyses qualitatives** (Lipman, 2004; Anagnostopoulos, 2006).

**La présentation de ces deux études de cas américaines, dont les données ont été parmi les plus analysées, montre ainsi les difficultés inhérentes à l'évaluation du *testing*: les enquêtes s'appuyant sur les tests locaux ne sont pas fiables; les dates retenues dans l'analyse sont fondamentales; l'étude des résultats des élèves en termes de valeur absolue ou de progression fait changer les conclusions des recherches. De façon générale, dans une même enquête, les effets de l'évaluation standardisée semblent varier suivant la discipline ou l'âge testés, sans que, nous le verrons à travers une méta-analyse présentée par la suite, il soit possible de trouver une logique à ces variations.**

**Pour élargir la démonstration au-delà de ces deux cas, nous présentons donc maintenant une revue d'études empiriques sur le sujet, qui portent tant sur des expériences nationales et régionales que sur des comparaisons internationales.**

**Comme pour les deux études de cas présentées précédemment, aucun consensus ne se dessine autour des recherches conduites aux niveaux national et régional** <sup>(6)</sup>. Il s'agit principalement d'une littérature nord-américaine. **Une partie de ces études s'est plus particulièrement intéressée aux tests fondés sur les compétences minimales.** Dans cette famille, certaines recherches ont mis en évidence des gains d'acquis des élèves importants en lien avec les dispositifs de *testing*. C'est par exemple le cas d'une étude pionnière américaine (Fredericksen, 1994) qui montra, en s'appuyant sur le test fédéral NAEP en mathématiques, que les évaluations fondées sur des compétences minimales étaient associées sur le long terme à une progression en mathématiques des résultats moyens des États. Cependant une réanalyse de ces dispositifs par Jacob (2001) montra, à travers les données d'un autre test national américain – le NELS <sup>(7)</sup> –, que la même politique n'était pas liée à des résultats plus élevés, pour la douzième année, ni en mathématiques ni en lecture. Bishop *et al* (2001) aboutirent à des résultats plus mitigés. La mise en place d'un test de compétences minimal conduit à des résultats faiblement liés aux acquis des élèves sauf si cet examen est fortement en lien avec le curriculum scolaire.

**D'autres analyses également nationales se sont intéressées à des dispositifs d'*accountability* plus larges que les compétences minimales. Les années 1990 furent en effet marquées dans**

---

<sup>(6)</sup> Nous ne retenons ici que les évaluations conduites à partir d'un test externe et non du test local inclus dans la politique d'*accountability* elle-même compte tenu de l'écueil méthodologique mis en évidence précédemment.

<sup>(7)</sup> *National Assessment of Educational Progress*

**de nombreux pays par le passage d'un *testing* fondé sur une analyse des compétences minimales à des évaluations portant sur un spectre plus large d'acquisitions (*proficiency testing*). Là encore, suivant les études et même suivant les réanalyses des données similaires, les résultats varient fortement.** Amrein et Berliner (2002) ont ainsi présenté une recherche étudiant chronologiquement les effets des nouvelles politiques d'*accountability* développées aux États-Unis par une multiplicité d'États depuis le début des années 1990. Ils cherchèrent ainsi à mettre en relation l'introduction de ces dispositifs avec de potentielles augmentations des acquis des élèves appréhendées à travers le test américain fédéral le NAEP. Ils montrèrent ainsi l'inconsistance des résultats à la fois pour les acquis en mathématiques et en lecture en 4<sup>e</sup> et 8<sup>e</sup> années: dans certains cas, ces réformes étaient liées à des gains en termes de performances, dans d'autres à des résultats négatifs. Rosenshine (2003) réanalysa les mêmes données en utilisant une nouvelle approche méthodologique et montra que les résultats moyens des États au NAEP avaient davantage crû dans les pays ayant mis en place un dispositif de *testing* à forts enjeux que dans les États qui n'avaient pas adopté de telles réformes. Amrein-Beardsley et Berliner (2003) reprirent les mêmes données et méthodologie que Rosenshine mais en incluant les taux d'exclusion du NAEP relatifs aux élèves en difficulté. Comme Rosenshine, ils montrèrent que les États ayant adopté un *testing* à forts enjeux voyaient leurs résultats progresser en quatrième année. Par contre, l'introduction, comme variable de contrôle, des taux d'exclusion au NAEP rendait cet effet non significatif.

Par la suite les études américaines se sophistiquèrent avec le développement des dispositifs d'*accountability* liés à la loi fédérale No Child Left Behind de 2002 <sup>(8)</sup>. Plutôt que d'envisager une vision dichotomique de ces réformes (le dispositif existe ou n'existe pas), plusieurs études créèrent des variables continues de façon à mettre en évidence le caractère plus ou moins contraignant des mesures d'*accountability* liées au *testing* (récompenses ou sanctions plus ou moins prégnantes). Dans cette perspective, Carnoy and Loeb (2002) aboutirent à des résultats mitigés. Ils montrèrent ainsi qu'une progression importante en mathématiques en huitième année au test fédéral américain NAEP sur la période 1996-2000 était associée à un renforcement de l'*accountability*, pour les élèves des différentes communautés ethniques dominantes et minoritaires (blancs, hispaniques et afro-américains). À l'opposé, les résultats du test en quatrième année étaient beaucoup plus faiblement liés à l'intensité de l'*accountability* pour les élèves des minorités ethniques. Pour les élèves issus de familles blanches, la relation disparaissait complètement. La recherche ne mit pas non plus en évidence de liens statistiques entre les dispositifs d'*accountability* et des changements à la fois dans les taux de redoublement en 9<sup>ème</sup> année et dans les taux de poursuite de scolarisation dans le secondaire supérieur. Appuyée également sur la construction d'un index d'*accountability*, l'étude de Hanushek et Raymond (2005) montra au contraire que les progressions des acquis entre les quatrième et huitième années, appréhendées à travers le test NAEP, étaient fortement liées à l'introduction de ces dispositifs et à leur pérennité dans le temps.

Nichols, Glass et Berliner (2006) reprirent l'idée d'une échelle d'intensité de l'*accountability* (le Taux de Pression de l'Évaluation – *Assessment Pressure Rating*) et fondèrent sa conception sur une documentation large institutionnelle des politiques conduites dans 25 États américains. Leurs résultats

---

<sup>(8)</sup> Adoptée en janvier 2002 grâce au soutien bipartisan des Démocrates et des Républicains, la loi *No Child Left Behind* (NCLB) fixe aux États américains en charge de l'éducation dans ce système éducatif fédéral une série d'objectifs de haut niveau en termes à la fois d'acquis des élèves, de niveau de diplomation en fin de secondaire supérieur (*high school*) et de qualité des enseignants recrutés. Les efforts et les progrès des États doivent être contrôlés à travers la mise en place de dispositifs de tests standardisés en mathématiques et lecture – en troisième et huitième années – et en sciences – trois épreuves au moins pendant la scolarité. Les écoles défaillantes en termes de progression annuelle des résultats de leurs élèves doivent recevoir des soutiens techniques spécifiques ou se voir définitivement fermées si les équipes pédagogiques ne parviennent pas à inverser la tendance malgré l'aide reçue.

là aussi sont mitigés. L'étude montre des corrélations importantes entre leur index de pression et les résultats des élèves en mathématiques en quatrième année au test américain fédéral, le NAEP. Plus la pression exercée par l'évaluation est forte, meilleurs sont les résultats. Par contre, parmi les multiples corrélations mises en évidence pour le test de 8<sup>e</sup> année en mathématiques, les résultats apparaissent inconsistants: les corrélations sont positives ou négatives suivant les cas. De même, les relations entre leur index de pression d'*accountability* et les performances au test en lecture aussi bien en 4<sup>e</sup> qu'en 8<sup>e</sup> années se sont avérées très faibles. Par contre, quand les relations sont apparues significatives, elles étaient négatives, suggérant ainsi que les dispositifs d'évaluation standardisée peuvent peser négativement sur les performances des élèves en lecture surtout en quatrième année. Au total, **les relations entre efficacité et dispositifs de *testing* n'apparaissent ni uniformément significatives ni automatiques.**

**Les recherches internationales qui peuvent aussi être mobilisées sur le sujet montrent la même inconsistance.** Une partie d'entre elles met en évidence les effets positifs apportés par le *testing*. C'est le cas de différentes études de Wöessmann (synthétisées dans Wöessmann, 2007). En s'appuyant sur un ensemble d'évaluations standardisées internationales conduites dans le secondaire (TIMSS <sup>(9)</sup> 1995, TIMSS-Repeat de 1999, PISA <sup>(10)</sup> 2000) et en utilisant un traitement statistique multi-niveau permettant de tenir compte des caractéristiques individuelles des élèves – comme le milieu social –, l'économiste allemand a montré que l'existence d'examen final externe dans le secondaire était associée à des performances plus élevées. Les données de l'enquête PIRLS <sup>(11)</sup> lui permirent de montrer que des effets positifs du *testing* peuvent aussi exister au primaire. L'impact bénéfique de l'*accountability* serait accru quand ces dispositifs sont développés dans le cadre d'une autonomie des établissements.

S'appuyant en partie sur les mêmes données (PISA 2000), mais les traitant au niveau national ce qui permet l'introduction de variables de contrôle relatives au contexte du pays (le niveau de développement économique, par exemple), Mons (2007) a au contraire montré que l'existence de dispositifs d'*accountability* n'était pas associée aux indicateurs d'efficacité. Les politiques d'évaluation standardisée dans l'enseignement obligatoire ont été appréhendées à travers plusieurs variables créées par l'auteur. Une première variable permet de rendre compte de l'existence dans les pays d'une évaluation centralisée nationale qui peut revêtir la forme à la fois d'examens nationaux donnant lieu à une certification (principalement à la fin du premier cycle du secondaire) et de tests standardisés (si des échantillons larges d'élèves ont été concernés). Cette évaluation centralisée standardisée s'oppose à des évaluations locales, c'est-à-dire pratiquées soit par les collectivités locales, soit par les établissements scolaires. La recherche a mis en évidence que, à niveau de développement économique égal, l'existence d'examens ou de tests centralisés n'était en relation ni avec les performances moyennes des élèves, ni avec le pourcentage des élèves dont le niveau scolaire est le plus élevé (niveau 5 de l'étude PISA), ni avec le pourcentage des élèves présentant des difficultés scolaires (niveau 1 de l'étude PISA). Par contre, en absence d'introduction dans les modèles de l'indicateur PIB/habitant comme variable de contrôle, les résultats mis en évidence par Wöessmann (2007) sur une liaison entre *accountability* et performances scolaires réapparaissent. Il semblerait donc que les conclusions issues des recherches internationales précédentes peuvent s'expliquer en partie par le manque de variables de contrôle relatives au niveau de développement des pays. En effet, les évaluations standardisées se sont surtout développées dans les pays les plus riches de l'OCDE. Les bons résultats en termes d'efficacité pourraient donc s'expliquer davantage par

---

<sup>(9)</sup> *Third International Mathematics and Science Study.*

<sup>(10)</sup> *Programme for International Student Assessment.*

<sup>(11)</sup> *Progress in International Reading Literacy Study.*

le niveau de développement économique élevé de certains pays qui peut, entre autres, autoriser l'organisation d'épreuves nationales coûteuses plutôt que par l'existence même de ces examens.

**De même, le dernier rapport PISA de l'OCDE (2007) n'a pas réussi à mettre en évidence des associations fortes entre des dispositifs de *testing* et l'efficacité.** «Quels liens les politiques et les pratiques d'*accountability* entretiennent-elle avec les performances des pays? La réponse est difficile, notamment parce que ces politiques sont associées à d'autres politiques et pratiques des établissements)» (OCDE, 2007, p. 243). Pour répondre à cette question, l'OCDE a développé des modèles multi-niveaux intégrant les caractéristiques socio-économiques des élèves ainsi que différentes caractéristiques des dispositifs d'*accountability* (liens avec des standards, caractère public ou non des résultats, usage ou non des données pour évaluer les enseignants...). Les résultats sont mitigés. Si les dispositifs d'évaluation fondés sur des standards sont associés positivement aux scores nationaux en science à l'épreuve de PISA 2006, cette relation disparaît dès lors que l'on tient compte des facteurs contextuels démographiques et socio-économiques. Seules les organisations qui intègrent la publication grand public des performances des écoles seraient liées à des résultats supérieurs. «Pour les autres caractéristiques des dispositifs d'*accountability* examinés dans PISA, les relations avec les performances sont faibles et statistiquement non significatives» (OCDE, 2007, p. 243).

**Au total, que l'on examine la littérature qui s'est concentrée sur les cas nationaux ou celle qui s'est appuyée sur les comparaisons internationales, les relations entre efficacité et *testing* apparaissent peu consistantes: elles ne sont ni univoques, ni automatiques.**

**L'évaluation standardisée est-elle davantage associée à une réduction des inégalités scolaires** entre groupes sociaux et ethniques comme l'ont affirmé certains de ses défenseurs? **Là aussi, la littérature abondante, principalement nord-américaine et rarement fondée sur des comparaisons internationales, n'aboutit pas à un consensus empirique.** Comme nous l'avons exposé précédemment les multiples recherches développées autour des expériences du Texas et de Chicago ont mis en évidence les résultats variables du *testing* pour les élèves issus de milieux socialement défavorisés. L'étude déjà citée de Carnoy et Loeb (2002), quant à elle, montre des effets positifs du *testing* sur les jeunes des minorités ethniques américaines pour certaines épreuves. Les résultats de l'étude de Hanushek et Raymond (2005) sont mitigés. Alors que la recherche met en évidence des résultats moyens positifs, elle montre que l'*accountability* ne tend pas à réduire les écarts de performances entre les publics scolaires blancs et afro-américains. Ces politiques seraient davantage bénéfiques pour le groupe hispanique. Les recherches de Lee et Wong (2004) et Nichols *et al* (2006) mettent en évidence, quant à elles, que les effets de ces dispositifs ne sont pas significatifs pour les élèves issus des minorités ethniques. De même, l'étude de l'OCDE (2007), fondée sur PISA 2006, n'a pas établi de liens entre les différentes formes d'*accountability* analysées dans le cadre de cette enquête et les inégalités scolaires d'origine sociale.

**Seul un type d'*accountability* particulier – l'organisation d'examens nationaux externes dans le secondaire – semble être en lien avec des inégalités plus faibles.** On retrouve ce résultat de façon consistante dans une multiplicité d'études nationales (Harris et Herrington, 2006). À l'international une seule étude s'est intéressée aux relations entre cette forme d'*accountability* et les inégalités sociales à l'école (Mons, 2007). Alors que la variable créée par l'auteur autour des examens nationaux externes n'était pas en lien avec les indicateurs d'efficacité comme nous l'avons précisé plus haut, la mise en œuvre d'évaluations centralisées est apparue associée à des inégalités scolaires d'origine sociale faibles. Autrement dit, l'existence d'épreuves externes nationales est liée à une relation d'intensité faible entre les performances des élèves et leurs caractéristiques socio-économiques. Les évaluations standardisées permettraient ainsi de limiter les phénomènes de

reproduction sociale au sein de l'école peut-être, entre autres, parce qu'elles homogénéisent les exigences académiques qui sont imposées aux enseignants et limitent ainsi les dérives curriculaires dans les établissements qui accueillent les élèves de milieux défavorisés. La pratique de l'évaluation standardisée postule implicitement que les attentes doivent être identiques pour tous les élèves qui sont reconnus capables d'un même niveau d'éducabilité, puisqu'ils sont soumis au même test.

Bishop (2006) a défini plus avant les critères qui doivent, selon lui, permettre de rendre plus efficaces les examens externes nationaux. On peut faire l'hypothèse que certains de ces critères participent également d'une réduction des inégalités scolaires. **Selon Bishop (2006), ces épreuves de certification nationales peuvent avoir des effets positifs si elles sont directement mises en lien avec des curricula et des standards externes, si elles mesurent un ensemble étendu de niveaux de performances et non un niveau d'acquis exclusif et, enfin, si elles couvrent de façon très large chaque cohorte de jeunes. Harris et Herrington (2006) insistent aussi sur le fait que ces examens externes auraient des effets positifs non pas parce qu'ils exercent des pressions psychologiques sur les élèves et les enseignants mais parce qu'ils constituent l'occasion d'approfondir l'exposition des élèves aux contenus d'enseignement en termes à la fois de temps alloué à l'instruction et d'exigence sur les contenus.**

Du fait de la multiplicité des analyses et des conclusions contradictoires auxquelles les études sur les effets de l'évaluation standardisée aboutissent, Lee (2008) en a réalisé **une méta-analyse visant à comparer et synthétiser leurs résultats de façon à essayer de mettre en lumière des grandes lignes directrices**. Pour ce faire, il y a sélectionné 14 études qui répondent à un ensemble de critères précis et qui recouvrent en grande partie celles que nous avons citées auparavant. En particulier, pour limiter les effets d'entraînement aux tests, ces recherches doivent avoir mobilisé des résultats à un test en mathématiques et/ou en lecture indépendant du test local, comparable nationalement et à faible enjeu (comme le NAEP). Les résultats auxquels aboutit le chercheur américain sont décevants. L'analyse synthétique de l'effet moyen des coefficients estimés dans les modèles construits dans ces études conduit certes dans un premier temps à un résultat positif. Autrement dit, la méta-analyse semble tout d'abord montrer que, au vu de l'ensemble de ces études, le *testing* aurait des effets positifs sur les performances des élèves. Cependant, outre le fait que cette moyenne cache en fait des disparités très importantes entre les résultats des études, cet effet moyen disparaît dès lors que l'on traite correctement le problème des liens de dépendance entre certaines recherches (de nombreuses études ayant ré-analysé les mêmes données). De la même façon, Lee cherche également à mettre en évidence des effets du *testing* qui pourraient être particulièrement liés à une discipline, un niveau scolaire ou une durée de mise en œuvre de ces politiques – certaines études ayant par exemple montré que l'*accountability* a des résultats plus probants en mathématiques qu'en lecture, pour les années scolaires du primaire (vs celles du secondaire) ou sur des durées longues. Là encore, la méta-analyse ne révèle pas d'effets significatifs. Enfin, dernière interrogation de Lee: l'évaluation standardisée conduit-elle à une réduction des inégalités scolaires? Une fois de plus, les résultats synthétiques de sa recherche apparaissent neutres.

Au-delà de l'efficacité et de l'égalité scolaires, il est également possible de s'interroger sur **l'efficience des politiques d'évaluation standardisée**, compte tenu de la rhétorique économiste à partir de laquelle ces réformes se sont développées. La promesse était bien celle d'une amélioration des performances des élèves à un moindre coût (Linn, 2000). Dès lors, il est étonnant de constater que peu de recherches empiriques se sont penchées sur le sujet (Behrens, 2006). Seule l'économiste américaine Hoxby (2004) a cherché à évaluer le coût financier de certains dispositifs outre-Atlantique pour montrer le faible investissement attaché à ces politiques. Ailleurs, au mieux dispose-t-on de quels indices très partiels. Par exemple, le coût de développement, d'administration, de notation et de

*reporting* du programme de tests de Floride est évalué à 42 millions de dollars par an (*Florida Department of Education*, 2003, cité par Jones, 2007). De même, le dernier rapport du parlement anglais sur le *testing* (House of Commons, 2007) renseigne sur certains éléments de coûts liés aux dispositifs d'*accountability* du pays: les 70 tests que passe en moyenne chaque élève durant sa carrière scolaire mobilisent annuellement 54 000 examinateurs et modérateurs. 68 % des écoles primaires emploient du personnel supplémentaire dédié aux tests.

**Au total, que l'on se retourne vers les études nationales ou internationales, vers les recherches qui portent sur les relations entre le *testing* et l'efficacité ou vers celles qui étudient les liens avec les inégalités scolaires, aucun consensus empirique ne se dégage actuellement autour des bienfaits de l'évaluation standardisée.** Comme le suggèrent Nichols (2007) et Lee (2008), la divergence des résultats des différentes enquêtes doit conduire à des nouvelles recherches qui analyseraient plus finement les dispositifs en jeu. On peut en effet faire l'hypothèse que la diversité des politiques menées conduit à ces conclusions variables. Ainsi, dans certaines configurations institutionnelles, le *testing* pourrait entraîner des effets bénéfiques alors que dans d'autres les effets pervers seraient les plus notables. Tout dépend en effet des processus activés durant la mise en œuvre, qui eux-mêmes sont en lien avec les comportements des différents acteurs impliqués dans la réforme. C'est ce que nous détaillons dans la troisième partie de ce rapport.



### III. ÉVALUATION STANDARDISÉE ET PROCESSUS ÉDUCATIFS: LES COMPORTEMENTS DES ENSEIGNANTS, CADRES INTERMÉDIAIRES, ÉLÈVES ET PARENTS FACE AU TESTING

---

La littérature, tant européenne que nord-américaine, sur les effets du *testing* sur le comportement des acteurs des systèmes éducatifs est désormais féconde. Comme le souligne Jones (2007), si des effets bénéfiques de l'évaluation standardisée ont bien été identifiés, la littérature empirique est également riche maintenant d'études ayant montré que des effets dits pervers peuvent également se développer dans certains contextes d'*accountability* externe, en particulier ceux des tests à forts enjeux. Nous présentons donc maintenant une revue de ces recherches sur les processus en jeu dans l'évaluation standardisée. Pour ce faire, un peu artificiellement nous analysons les réactions que les enseignants, les cadres administratifs des systèmes éducatifs, les élèves, puis les parents développent face à ces instruments de régulation.

#### A. Enseignants: une tendance à la résistance à la culture de l'évaluation standardisée quantitative

Si la littérature empirique relève des effets positifs du *testing* pour les enseignants, elle met aussi en évidence des évolutions notables des pratiques pédagogiques (centration sur l'entraînement aux tests, rétrécissement des curricula...) qui questionnent l'intérêt de ces dispositifs et peuvent expliquer la résistance de ce corps professionnel à la culture d'évaluation standardisée que l'on retrouve dans de nombreux pays.

**Les dispositifs d'évaluation standardisée peuvent avoir des effets bénéfiques sur l'activité d'enseignement. Des études menées dans plusieurs pays montrent tout d'abord que les enseignants adhèrent à l'idée de standards de performances en éducation. Plusieurs organisations syndicales se sont ainsi prononcées en faveur de ces réformes.** Par exemple, aux États-Unis l'*American Federation of Teachers* (AFT, 2001, cité par Behrens, 2006) affirme que le syndicat «a été un des premiers avocats des réformes fondées sur les standards. Dès 1992, du fait d'un souci croissant au sujet des acquis des élèves dans le cadre d'une économie devenue globale et de l'existence d'un écart intolérable entre les performances des Afro-américains et des Blancs [...], l'ancien président de l'AFT a demandé aux États de prendre pour modèle les pays réussissant le mieux et de mettre en place des standards scolaires rigoureux et clairs qui s'appliqueraient à tous les élèves. Pour l'AFT, une réforme fondée sur les standards doit s'articuler autour des composantes suivantes logiquement ordonnées: des standards de qualité qui sont supportés par des curricula clairs permettant leur mise en œuvre, la formation continue pour les enseignants, des batteries de tests alignés sur les standards et, enfin, des incitants justes et des ressources financières suffisantes pour aider les élèves à passer dans la classe supérieure» (p. 9).

**Des sondages conduits auprès des enseignants montrent également leur adhésion aux dispositifs de standards de performances, au moins dans certains pays.** Selon Johnson et Duffet (2003), malgré des réserves sur les modalités d'applications (que nous développerons par la suite), pour 80 % des enseignants américains les standards fournissent de bonnes lignes directrices pour mieux comprendre ce qui doit être enseigné et pour améliorer les performances des élèves. 87 % d'entre eux pensent également que les élèves doivent passer des tests pour être promus dans la classe supérieure et, qu'en cas d'échec, ils doivent être envoyés dans des camps de remédiation l'été ou redoubler la classe qu'ils viennent de suivre si leur situation ne s'est pas améliorée grâce à ce soutien. Selon Jones (2007), des enquêtes conduites dans plusieurs États américains (Floride, Ohio...) ont également montré que les enseignants envisageaient positivement les tests: en effet selon eux, ces dispositifs permettent de mieux organiser les contenus d'enseignement en fonction des années scolaires, donnent une réalité aux standards qui, en leur absence, pourraient ne pas être

appliqués, permettent d'établir un diagnostic des faiblesses des élèves et développent chez les enseignants une culture du résultat.

**En Suède, un sondage administré par l'Agence Nationale pour l'Éducation (2004, *National Agency for Education*), montre également que les enseignants adhèrent dans leur très grande majorité aux tests nationaux développés dans leur pays.** Une grande majorité d'entre eux déclarent en effet que l'évaluation standardisée donne des lignes directrices claires sur les contenus à enseigner, aide à mettre en évidence les faiblesses et les atouts des élèves et qu'elle ne constitue pas un cadre contraignant pour leurs enseignements. L'outil est également apprécié parce qu'il donne un cadre national aux contenus d'enseignement dans un système aujourd'hui fortement décentralisé et dans lequel les collectivités locales jouent un rôle non négligeable dans les activités pédagogiques, ce qui peut laisser craindre l'apparition d'inégalités entre les régions et entre les établissements. Une étude des chercheurs norvégiens Helgoy et Homme (2007), comparant les situations norvégiennes et suédoises, a mis en évidence des résultats très proches pour la Suède. Par contre, leur enquête montre une réticence plus marquée de la part des enseignants norvégiens, ce qui laisse à penser que le positionnement de ce corps de professionnel face au *testing* dépend bien d'éléments contextuels.

**Au Royaume-Uni (Angleterre), tout comme aux États-Unis ou en Suède, malgré les fortes contestations dont font l'objet les tests SATs actuellement, le principe même de l'évaluation standardisée n'est pas remis en cause par les représentants des enseignants (House of Commons, 2007). Le *testing* qui met l'accent sur une culture de l'évaluation et le développement d'objectifs cognitifs semble, de plus, bien influencer les pratiques éducatives des enseignants.** Par exemple, dans une étude comparative sur l'école primaire française et anglaise, Broadfoot, Osborn, Sharpe et Planel (2001) ont montré, à travers l'administration d'un questionnaire, que les enseignants anglais mettent au premier rang de leurs priorités éducatives les «compétences en évaluation», les «connaissances disciplinaires» et des «objectifs clairs»<sup>(12)</sup>. Pour les chercheurs, cette évolution dans les priorités des enseignants est à mettre en lien avec l'introduction du *National Curriculum* en 1988 et le développement du dispositif d'évaluations standardisées mis en place au début des années 1990. Pour Hargreaves (2002), les enseignants anglais grâce aux tests SATs ont désormais une perception très claire de ce qui est attendu aux différentes étapes de la carrière scolaire des élèves qui sont explicitement représentées par niveau scolaire, les *Key stages*.

**En France, le rôle positif d'orientation joué par les évaluations standardisées a aussi été analysé:** «Une voie permettant d'agir sur les programmes enseignés consiste à agir sur les épreuves nationales et sur les examens. Par exemple, il a suffi que les évaluations diagnostiques à l'entrée en sixième testent des compétences en géométrie durant quelques années pour que cet aspect des mathématiques, un temps négligé, se réactive» (IGEN-IGAENR, 2005). La force normative qu'imposent les tests standardisés apparaît d'autant plus forte qu'en France, aucun enjeu n'était lié à ces évaluations diagnostiques. Plus généralement, «le pilotage des disciplines a fortement recouru durant les dernières années au levier que constitue l'introduction de nouvelles formes d'épreuves. (...) Bien entendu, ces épreuves demandent un temps d'explicitation (...). Mais l'effet en retour s'observe très vite» (IGEN-IGAENR, 2005).

Demilly (2001) propose également une vision positive des apports de l'évaluation standardisée pour le travail enseignant: «les effets formatifs de l'évaluation ne sont pas négligeables, ainsi que le décloisonnement des cultures professionnelles et le développement de l'attitude à coopérer dans un cadre d'action interprofessionnel. Réflexion sur le «réfèrent», identification d'indicateurs pertinents

---

(12) En bas de leurs priorités apparaissent désormais: les «relations avec les enfants» et le «maintien de la discipline».

permettant de décrire le fonctionnement des dispositifs, mises à plat parfois douloureuses des pratiques, telles sont les occasions d'une explicitation et d'un enrichissement d'un certain nombre de savoir-faire professionnels».

**Au total, la perception générale par les enseignants des standards et des dispositifs de tests associés ainsi que dans, certaines configurations institutionnelles, la réalité des effets de ces mesures sur l'activité pédagogique semblent positives: ces réformes permettent en effet de donner des guides clairs pour mettre en œuvre les curricula, d'empêcher l'apparition de fortes inégalités dans le développement des syllabi locaux, de mettre l'accent sur les résultats réels des élèves, en particulier pour ceux issus des milieux défavorisés, et de favoriser un travail en équipe autour de l'analyse des résultats des évaluations.**

Cette ouverture au *principe* de l'évaluation standardisée, dans certains pays, n'empêche pas, par ailleurs, la dénonciation par le corps enseignant de certains dispositifs concrets perçus comme négatifs parce que trop rustiques dans les compétences analysées, ne tenant pas compte des caractéristiques sociales des publics accueillis ou faisant le lien entre les performances des élèves et leurs rémunérations. Cette remise en cause est particulièrement virulente aux États-Unis comme le montre un sondage récent – 70 % des enseignants affirment que les tests sont trop nombreux <sup>(13)</sup>, au Royaume-Uni (Angleterre) où les syndicats d'enseignants ont récemment appelé au boycott des tests nationaux ou encore en France où les syndicats enseignants du primaire se sont violemment prononcés contre les nouveaux tests mis en place en 2009. Selon un récent rapport d'inspection, bien que s'agissant d'évaluations obligatoires, 70 % des écoles seulement ont accepté de transmettre à l'administration centrale du Ministère leurs résultats pour le test de 5<sup>e</sup> année et 85 % pour celui de seconde année.

**Pour que l'évaluation soit perçue positivement, Demailly (2001) observe que trois conditions doivent être réunies dans la définition et la mise en œuvre de l'évaluation:** 1) l'évaluation se développe de manière participative avec une forte association des personnels enseignants, 2) les objectifs de l'évaluation sont démocratiques (par opposition à l'évaluation autoritaire), 3) les promoteurs du projet présentent une capacité de conviction et manifestent un volontarisme certain.

Ces conditions n'étant pas réunies dans un certain nombre de pays et de dispositifs, une littérature, désormais très riche et développée dans des contextes nationaux multiples, a mis l'accent sur **les effets pervers de l'évaluation standardisée pour les activités d'enseignement. Ces recherches montrent clairement que les tests, en particulier quand ils sont associés à de forts enjeux, peuvent conduire à la fois à une évolution défavorable des pratiques pédagogiques ainsi que, dans certaines circonstances, à un sentiment de déprofessionnalisation, qui se traduit par des phénomènes de démotivation des enseignants.**

En effet, comme nous l'avons déjà signalé pour les cas du Texas et du district de Chicago, **certaines dispositifs d'évaluation standardisée peuvent tout d'abord conduire au phénomène désormais bien analysé du «teaching to the test» ou focalisation sur l'entraînement intensif aux tests** (Gordon et Reese, 1997; Jones et Egley, 2004; Belair, 2005; Jones, 2007...). Face aux impératifs de résultats auxquels ils sont soumis, dans certains cas, les enseignants consacrent désormais une grande partie du temps d'instruction à l'entraînement à des exercices proches de ceux qui seront administrés dans le test. Ces nouveaux comportements ont particulièrement été analysés aux États-

---

<sup>(13)</sup> Sondage d'opinion sur l'éducation réalisé par l'agence américaine Public Agenda et intégré dans la série "Reality Check 2006". La série de 2006 est disponible à l'adresse suivante:  
<http://www.publicagenda.com/files/pdf/rc0603.pdf>

Unis (Jones, 2007) et au Royaume-Uni (Angleterre), deux contextes marqués par la mise en œuvre de tests à forts enjeux. Ainsi, selon le rapport parlementaire anglais «Testing and assessment» (House of Commons, 2007), une étude de la *Royal Society* de 2003 a montré qu'il existe, au Royaume-Uni, une différence très importante dans le temps consacré aux évaluations entre l'Angleterre – région dans laquelle l'évaluation standardisée est très soutenue – et l'Écosse qui a développé une politique plus souple en la matière. Ainsi, dans le secondaire, les professeurs anglais passeraient deux fois plus de temps que leurs collègues écossais sur des activités d'évaluation. Le même rapport évalue qu'au trimestre du printemps, 70 % des écoles du primaire consacrent trois heures par semaine au seul entraînement au test correspondant au *Key stage 2*, administré en sixième année.

Au-delà du phénomène d'entraînement intensif aux tests, **les dispositifs d'évaluation standardisée peuvent aussi entraîner ce qui a été synthétisé sous l'appellation de «rétrécissement des curricula»** (Behrens, 2006; Jones, 2007; House of Commons, 2007...). Cette famille de pratiques pédagogiques revêt plusieurs formes. Elle peut tout d'abord consister en une contraction du spectre des disciplines enseignées. Les tests se concentrant en général sur un nombre limité de matières, les enseignants, surtout dans les classes de primaire, tendent à accorder à la fois moins d'importance et moins d'heures d'enseignement aux disciplines qui ne sont pas évaluées. Ainsi, certaines études empiriques anglaises et américaines ont-elles montré que les sciences sociales, les disciplines artistiques ou sportives voyaient leurs volumes horaires nettement réduits du fait des épreuves standardisées (Jones, Jones et Hargrove, 2003; House of Commons 2007...). Le rétrécissement du curriculum se traduit également par **une focalisation des enseignants sur les compétences testées qui le plus souvent s'avèrent également basiques par opposition à des compétences complexes – par exemple la résolution de problèmes – plus rarement évaluées. Au-delà des contenus d'enseignement, les épreuves standardisées tendent également à focaliser les enseignants sur des objectifs strictement cognitifs (Osborn, 2006; Jones, 2007 ...) au détriment des autres missions de l'école (socialisation, développement de la créativité, autonomie, participation à la vie citoyenne...).**

Au-delà des contenus et des objectifs d'enseignement, dans certaines circonstances, **les évaluations standardisées peuvent faire évoluer les pédagogies elles-mêmes. Couvrant un large spectre de connaissances qu'il faut désormais faire assimiler aux élèves dans un temps limité, le testing peut conduire les enseignants à se concentrer sur des méthodes pédagogiques fondées sur la mémorisation rapide plutôt que sur une exploration active davantage consommatrice de temps** (Gordon et Reese, 1997). Plus largement, c'est la perception même de l'essence du métier qui évolue. Par exemple, comme l'a montré une analyse comparative menée au Royaume-Uni (Angleterre), au Danemark et en France (Osborn, 2006), avec le nouvel accent mis sur l'évaluation standardisée, les enseignants anglais considèrent désormais comme prioritaire leur rôle de transmetteur de connaissances et de compétences au détriment de leurs activités d'accompagnement individualisé et affectif des élèves (*pastorale care*).

**Les tests à forts enjeux, et plus particulièrement les indicateurs de résultats médiatisés qui leur sont associés, ont également des conséquences sur la perception des élèves développée par les enseignants, sur la centration de leur attention sur certains d'entre eux ainsi que sur la nature des publics recrutés par les écoles.** Comme nous l'avons vu précédemment dans les expériences au Texas ou dans le district de Chicago, les enseignants peuvent être amenés, dans certains cas, à catégoriser leurs élèves (élèves brillants, élèves qui peuvent réussir le test avec un soutien, élèves en échec durable). Ce classement peut alors les conduire à isoler les élèves en très grande difficulté qui quel que soit le soutien apporté ne réussiraient pas l'évaluation à court terme et

donc ne permettraient pas à leur école d'améliorer ses performances. Au Royaume-Uni (Angleterre), Levacic (2001) a ainsi montré que l'indicateur de performance des écoles le plus médiatisé – le GCSE-1<sup>(14)</sup> – tend à orienter la conduite des enseignants dans un contexte de forte concurrence entre écoles. À travers une comparaison des cas français et anglais, van Zanten (1999, p.145) aboutit également au même constat et fait le lien entre le choix de l'école, les évaluations standardisées et la mécanique de recrutement des élèves: «si les chefs d'établissement et les enseignants ont tendance en toute circonstance à rechercher des «bons clients», cette tendance est nettement accentuée par une logique de concurrence. Dans ce contexte, les établissements qui le peuvent sont conduits à devenir encore plus sélectifs de façon à recruter les élèves qui améliorent encore leur image en apportant de la valeur ajoutée: des élèves dont les résultats scolaires contribuent à l'image d'un établissement performant, bien classé dans les évaluations nationales (...) mais aussi des élèves dont la tenue, le langage et le comportement jouent le rôle de marqueurs de la «qualité sociale» de l'établissement». Une fois recrutés, ces bons élèves bénéficieraient d'un traitement spécial. Dans certains établissements anglais, les moyens financiers et le travail des enseignants seraient déplacés, souvent contre le gré de ces derniers, vers des activités programmées pour les élèves les plus brillants (van Zanten, 1999). La chercheuse souligne également que les élèves de niveau moyen dont les progrès scolaires permettent d'améliorer les performances des établissements sont également ciblés par des actions pédagogiques spécifiques.

Au-delà d'une nouvelle attention portée à certaines populations d'élèves, **les dispositifs d'accountability externe dure peuvent également conduire certaines équipes pédagogiques à pratiquer directement des tricheries qui visent à augmenter les résultats de leur école.** Prenant appui sur le cas de l'État canadien de l'Ontario, Bélair (2005) affirme que «dans certains cas cités par des enseignants, des écoles sont même allées jusqu'à identifier des élèves en programmes spécialisés, afin qu'un nombre moindre d'élèves faibles passent ces tests pour favoriser un score élevé de réussite». Les mêmes phénomènes ont été observés également au Texas comme nous l'avons vu précédemment avec l'exclusion de certains élèves en grande difficulté du test fédéral NAEP. En 2006, aux Pays-Bas – pays marqué par un dispositif d'évaluation standardisée à forts enjeux, l'inspection, suite à des rumeurs, a également mené une enquête dans certaines écoles sur des pratiques d'exclusion du test de fin du primaire. Il s'est avéré que dans certains cas, ne participaient pas à l'épreuve les élèves qui très certainement seraient orientés vers la voie scolaire la moins prisée: le *Leerwegondersteunend onderwijs* (LWOO; *learning support*)<sup>(15)</sup>. Là aussi, ces agissements visaient à présenter des performances d'établissement supérieures à la réalité.

Ces nouvelles pratiques, qui entrent en contradiction avec le référentiel professionnel des enseignants – et en particulier avec ce qu'ils perçoivent comme devant être leurs missions pédagogiques – entraînent des **phénomènes de démotivation dans ce corps de professionnels** (Debard et Kubow, 2002; *Center on Education Policy*, 2006; Jones, 2007...). En particulier, dans certains cas, **la perception du métier devient négative, la satisfaction dans l'activité professionnelle diminue**

---

<sup>(14)</sup> Le *General Certificate in Secondary Education* (GCSE) est un examen national, externe, qui est passé par les élèves anglais de 15-16 ans, à la fin de leur 11<sup>e</sup> année. Il clôt la scolarité obligatoire. La plupart des élèves choisissent d'être évalués dans 8 à 10 disciplines. Les résultats nationaux sont ventilés par école. Ils sont appréhendés à travers deux indicateurs. Le GCSE-1 est l'indicateur le plus médiatisé (à travers les *League Tables*): il correspond à la proportion des élèves obtenant des notes de A à C (bon niveau) dans au moins 5 matières. Le GCSE-2 reflète plus largement les performances des établissements mais fait l'objet de moins d'attentions puisqu'il ne renseigne pas sur leur capacité à produire des «bons élèves»: il correspond à la proportion des élèves obtenant des notes de A à G dans au moins 5 matières. «G» est la note minimale pour décrocher son certificat dans une discipline.

<sup>(15)</sup> Le rapport en néerlandais est disponible à l'adresse suivante:  
[http://www.owinsp.nl/nl/home/naslag/Alle\\_publicaties/Eindtoets\\_po](http://www.owinsp.nl/nl/home/naslag/Alle_publicaties/Eindtoets_po)

**tandis que la perception d'un stress nouveau apparaît.** Par exemple, une enquête conduite en Caroline du Nord a montré que 84 % des enseignants considèrent que leur métier est devenu plus stressant depuis l'introduction des tests à forts enjeux (cité dans Hargrove *et al.*, 2004). Les évaluations standardisées sont aussi considérées comme une des causes du départ du métier des enseignants. Par exemple, dans l'étude américaine de Hoffman, Assaf et Paris (2001), 85 % des enseignants affirment que les meilleurs d'entre eux quittent la profession du fait des tests à forts enjeux. D'autres études américaines montrent qu'une proportion non négligeable des enseignants qui veulent demeurer dans cette activité a demandé à ne plus enseigner dans les années scolaires concernées par les évaluations (Tobin et Ave, 2006 cité par Jones, 2007).

**Ces phénomènes de démotivation semblent particulièrement importants dans les écoles accueillant des publics défavorisés et dont les résultats bruts apparaissent faibles, du fait de la non prise en compte des caractéristiques sociales de leurs élèves** (Jones, 2007). La définition d'indicateurs sous forme de valeur-ajoutée permettant d'évaluer les apports pédagogiques réels des établissements s'avère être une demande récurrente de la part des syndicats notamment aux États-Unis et au Royaume-Uni (Angleterre) (Behrens, 2006). Ces écoles rencontreraient également, toujours selon les enseignants, des difficultés à recruter des professionnels de qualité. Jones (2007) souligne cependant que, au-delà de ces études qui portent sur la perception des enseignants, des enquêtes mesurant la réalité des difficultés de recrutement et l'importance du *turn over* (*attrition rates*) doivent être conduites.

**Ce phénomène de démotivation des enseignants s'expliquerait, selon certains chercheurs, par un profond sentiment de déprofessionnalisation de leur métier.** C'est la thèse des chercheurs belges Maroy et Cattonar (2002) qui soulignent que la seule explication corporatiste ne peut expliquer l'opposition des enseignants à certains dispositifs de *testing*. Osborn (2006) pour le Royaume-Uni (Angleterre) et Dupriez (2005) qui a comparé les cas anglais et belges rejoignent également cette thèse. En effet, jusque-là les enseignants s'intégraient dans ce que les sociologues des organisations ont appelé des bureaucraties professionnelles (Bidwell, 1965, cité par Maroy et Cattonar, 2002): c'est un modèle d'organisation hybride alliant, d'un côté, les règles rigides, dépersonnalisantes mais rationnelles de la bureaucratie (dans le sens wéberien (1922)) et, de l'autre, une latitude d'action qui résulte de la reconnaissance forte d'une expertise professionnelle de haut niveau. De par cette dernière caractéristique, les enseignants sont assimilés par les sociologues à une *profession*, dans l'acception anglo-saxonne du terme, ce qui pourrait se rapprocher dans l'univers francophone de l'expression «profession libérale». En effet, si l'enseignant est bien soumis à des règles très contraignantes (organisation scolaire fixe, programmes scolaires à suivre dans de nombreux pays, existence d'un supérieur encadrant et notant son travail), par ailleurs, il dispose d'une liberté importante dans la conduite pédagogique et l'activité au quotidien de sa classe du fait de ses compétences reconnues. C'est cette latitude d'action et donc son statut de «semi-professionnel» que le développement de tests standardisés viendrait remettre en cause. Il ne serait plus considéré comme le seul acteur capable de poser un jugement souvent définitif et lourd de conséquences sur les compétences acquises par «ses» élèves.

Pour Maroy et Cattonar (2002), «les récentes réformes touchant le curriculum (davantage centralisé et redéfini sur base d'une approche par compétences) et les pratiques d'évaluation (avec la diffusion des batteries d'épreuves étalonnées), qui sont davantage prescrits et échappent désormais en partie aux enseignants en exercice pour devenir le produit d'acteurs externes, limitent la sphère d'activité traditionnelle reconnue au groupe. On pourrait parler d'une certaine «déqualification» (Lessard, 1999)».

**Au total, la position des enseignants apparaît bien mitigée par rapport à l'évaluation standardisée: acceptation du principe général mais dénonciation des mécanismes à forts enjeux qui orientent trop les démarches pédagogiques.**

## **B. Les cadres intermédiaires du système éducatif: une prise en main de l'outil**

Même si la littérature sur le positionnement face aux évaluations standardisées des responsables opérationnels encadrant le service éducatif est à ce jour encore peu développée, des enquêtes qualitatives menées dans plusieurs pays ainsi que des sondages montrent que **ce corps professionnel remet moins en cause que les enseignants à la fois le principe mais aussi les dispositifs concrets de *testing* qui leur sont proposés, même si les interrogations ne sont pas totalement absentes de la réflexion conduite sur le sujet.**

**Ainsi, aux États-Unis, Farkas, Johnson et al (2003) ont montré que la très grande majorité des personnels d'encadrement, surtout les chefs d'établissement et les *superintendents* <sup>(16)</sup>, étaient acquis à l'évaluation standardisée:** «seule une poignée d'entre eux pensent que c'est une erreur, et beaucoup indiquent qu'ils se concentrent depuis déjà longtemps sur les performances des élèves, la qualité des enseignants et l'*accountability*. La très grande majorité affirme qu'ils travaillent à la réduction de l'écart entre les performances des élèves des minorités ethniques et des élèves blancs et à l'amélioration des capacités d'expression des élèves non anglophones [...]. Les *superintendents* des districts urbains semblent spécialement acquis aux standards» (p. 48). De même, un sondage récent du Groupe new-yorkais de sondages d'opinion Public Agenda (2006) <sup>(17)</sup> montre que 90 % des *superintendents* et 85% des chefs d'établissement pensent que les données d'évaluation quantitative des élèves peuvent être utiles pour améliorer l'enseignement. Ceci n'exclut pas pour autant une vision négative du dispositif concret en cours d'application aux États-Unis, la loi *No Child Left Behind* de 2002: moins de la moitié des mêmes personnels d'encadrement pensent que cette législation permettra de faire progresser les standards scolaires.

**Au Royaume-Uni (Angleterre), les auditions parlementaires menées par la Commission *Children, School and Family* pour son rapport «Testing and Assessment» (House of Commons, 2007) ont montré une adhésion des personnels d'encadrement au *testing*.** Auditionné, le Secrétaire général de l'Association Nationale des Chefs d'établissement (*National Association of Head Teachers*) affirme: «personne dans notre association ne veut retourner à la situation des années 1970 quand nous ne savions pas ce que l'école juste à côté faisait» (p. 12). Également auditionné, un représentant du comté de l'Hampshire affirme aussi: «les écoles reconnaissent tout à fait la nécessité d'un monitoring des progrès des élèves, de la fourniture d'information régulière aux parents et de l'utilisation des évaluations pour améliorer les résultats des écoles» (p. 12).

**En France, la perception et l'usage des évaluations standardisées semble dépendre des corps d'encadrement considérés** (Mons et Pons, 2006). Ainsi si l'inspection générale (IGEN-IGAENR, 2005) note que les chefs d'établissement et les administrations académiques (administrations

---

<sup>(16)</sup> Le plus souvent placé à la tête du *district* (l'autorité locale en charge de l'opérationnalisation du service éducatif en liaison avec l'administration de l'État dans le système fédéral américain), le *superintendent*, dans la majorité des cas, est responsable pour les établissements de son district des politiques de sélection et de recrutement des enseignants, de la définition des budgets opérationnels et plus largement de la définition et du contrôle des politiques d'établissement au sens large.

<sup>(17)</sup> Public Agenda réalise des sondages d'opinion sur l'éducation notamment la série "Reality Check". Le sondage, présenté ici, appartient à la série de 2006 et est disponible à l'adresse suivante: <http://www.publicagenda.com/files/pdf/rc0603.pdf>

déconcentrées) ont peu recours aux résultats produits par les épreuves standardisées, les cadres intermédiaires plus directement en relation avec les enseignants (les inspecteurs d'académie et les IEN<sup>(18)</sup> dans le primaire) les utilisent de façon plus intensive, en particulier les évaluations diagnostiques. Dans certains cas, ils sont promoteurs d'épreuves complémentaires. Ces usages sous-tendent parfois des messages implicites adressés aux enseignants. «Certains fractions de l'encadrement de première ligne du travail enseignant (...) trouvent dans le maniement de l'obligation de résultats et la culpabilisation afférente un instrument de pouvoir commode sur les personnels qu'ils dirigent et un instrument de valorisation de leur propre rôle professionnel, même si pour leur propre compte ils ne sont pas disposés à subir la même évaluation» (Demailly, 2001). L'usage des instruments reste cependant circonscrit. «On reste surpris par la faible utilisation par le premier degré des évaluations d'entrée en 6<sup>e</sup> (19), plus utilisées en collège pour mettre en place des correctifs qu'en primaire dans un souci d'anticipation des difficultés à venir», note l'inspection (IGEN-IGAENR, 2005).

**En Belgique, Maroy et Cattonar (2002) notent que le développement de l'évaluation standardisée, qui s'insère dans le cadre plus global de référentiels de compétences centralisés, a entraîné dans ce pays historiquement décentralisé, la création d'une nouvelle techno-structure pédagogique.** Experts en sciences de l'éducation ou anciens enseignants, ils sont chargés, entre autres, de développer les socles de compétences et les batteries d'épreuves standardisées. «On pourrait [...] parler d'une *knowledge elite*, soit une élite intellectuelle ou techno-pédagogique au sein de la profession, qui n'est sans doute pas neuve mais qui tend au cours de la dernière décennie à s'étoffer considérablement. (...) Les enseignants les perçoivent de manière ambivalente, simultanément comme des soutiens potentiels dans l'exercice de leurs tâches mais aussi comme des vecteurs d'une standardisation et d'une formalisation de leurs pratiques pédagogiques» (Maroy et Cattonar, 2002). Les réformes actuelles belges conduisent également aux côtés de cette élite pédagogique à l'accroissement des pouvoirs de l'élite administrative plus traditionnelle. Au total, ces phénomènes induiraient une redéfinition du contexte et des relations dans lesquels s'insère le travail enseignant. «Notre hypothèse est que la division du travail entre ces catégories s'accroît et que de plus en plus les enseignants se trouvent dans une situation de dépendance soit de nature «technique et professionnelle», vis-à-vis de l'élite technico-pédagogique, soit de nature administrative et gestionnaire vis-à-vis des administrateurs des établissements ou du système scolaire» (Maroy et Cattonar, 2002).

**Au total, paradoxalement, alors que les politiques d'évaluation standardisée s'inscrivent en rupture avec les régulations bureaucratiques antérieures, il semble qu'elles puissent conduire à un développement ou une régénération des encadrements intermédiaires et donc à un contrôle effectif accru des pratiques et méthodes pédagogiques et non pas seulement des résultats académiques.**

Si les professionnels de l'éducation présentent face aux tests des positions variées, comment les élèves, principaux intéressés de ces réformes éducatives, perçoivent-ils ces dispositifs?

### **C. Les élèves et le fardeau des tests**

Comme nous l'avons vu dans la première partie de cette section sur les effets théoriques attendus des évaluations standardisées, **ces dispositifs sont censés conduire à une plus grande**

---

(18) Inspecteur de l'Éducation Nationale en charge de l'évaluation et de l'animation pédagogique des enseignants travaillant dans sa circonscription.

(19) 1<sup>ère</sup> année du secondaire français ou 6<sup>e</sup> année de l'enseignement obligatoire.



**motivation des élèves pour les études au moins en termes de motivation extrinsèque<sup>(20)</sup>. Le peu d'études ayant analysé ce point présentent des résultats contradictoires.** Ainsi, lors d'une enquête qualitative menée dans un district de l'Ohio, 83 % des élèves en primaire et 45 % dans le secondaire ont affirmé que les tests les avaient conduits à travailler davantage (DeBard et Kubow, 2002). A l'opposé, d'après les enseignants, leur motivation intrinsèque serait remise en cause par ces dispositifs ou du moins peu dynamisée. Ainsi, plusieurs recherches ont montré que le *testing* soit réduirait le fait d'«aimer apprendre» («love of learning»), qui est une des dimensions de la motivation intrinsèque, soit serait neutre (Jones *et al*, 1999; Rapp, 2002; Yarbrough, 1999, cités par Jones, 2007). Dans certaines situations, les évaluations standardisées parce qu'elles conduisent à des pratiques pédagogiques moins stimulantes pour les élèves (mémorisation rapide plutôt qu'apprentissage actif par exemple) semblent conduire à un désengagement des élèves.

**Si les études sur la motivation des élèves demeurent peu nombreuses, de nombreuses recherches au contraire ont porté sur le nouveau stress ressenti du fait du développement de ces dispositifs.** A la fois les élèves et les enseignants ont témoigné, dans certains contextes marqués par des épreuves à forts enjeux, de phénomènes d'anxiété, d'irritation, de pleurs ou de douleurs liés au *testing* (Jones *et al*, 1999; Hoffman, Assaf et Paris, 2001; DeBard et Kubow, 2002; Gregory et Clarke, 2003; House of Commons, 2007; Jones, 2007...). En Suède, alors que les enseignants adhèrent majoritairement au principe de l'évaluation standardisée et qu'ils déclarent ne pas être contraints dans leurs enseignements par cet instrument, un cinquième d'entre eux rapportent cependant que leurs élèves souffrent de stress à cause des tests nationaux (Agence Nationale pour l'Éducation, 2004).

Au-delà du stress, **le *testing* peut également jouer négativement sur les carrières des élèves.** Obligeant certains à redoubler, stigmatisant les difficultés scolaires des autres, ces dispositifs peuvent entraîner à terme une augmentation des abandons scolaires (Haney, 2000; Jacob, 2001; Amrein et Berliner, 2003 ...). Comme nous l'avons vu auparavant, le peu d'attention, dans certains cas, porté aux élèves en grande difficulté ne pouvant pas réussir aux tests finit par créer une démotivation qui conduit à des sorties plus précoces de la carrière scolaire.

**Ces phénomènes seraient aggravés dans le cas des élèves issus des milieux défavorisés (Lipman, 2004; Jones, 2007; Hong et Youngs, 2008).** Disposant de peu de ressources pour préparer les tests, labellisés comme en échec dans des écoles qui publient des performances médiocres, ces élèves subissent de plus un entraînement intensif aux tests et un rétrécissement du curriculum qui seraient encore plus marqués que dans les écoles accueillant des publics favorisés. Cette centration sur les évaluations standardisées se ferait au détriment d'activités d'éveil et de culture générale dont le manque s'avère particulièrement préjudiciable pour des élèves dont les familles ne peuvent prendre le relais sur ces terrains.

**Au total, si les tests peuvent avoir dans certains contextes des effets d'entraînement positifs sur le travail scolaire, dans les cas de dispositifs liés à de forts enjeux, la pression exercée sur les élèves ainsi que le nombre d'épreuves à subir – 70 évaluations par exemple dans la carrière scolaire d'un jeune anglais – ont des conséquences négatives sur leur comportement vis-à-vis de l'école.**

---

<sup>(20)</sup> Voir note de bas de page précédente n° 3 sur l'opposition entre motivation intrinsèque qui trouve dans le sujet ses causes et la motivation extrinsèque supportée par des causes externes.

Dernier acteur que nous analysons ici, les parents semblent supporter l'idée de *testing* même si les sondages effectués montrent clairement que les objectifs cognitifs visés par les tests ne constituent pas pour eux la totalité des missions assignées aux écoles.

#### **D. Les parents: une réception positive du *testing* mais des attentes pour l'école qui vont au-delà**

La littérature sur la position des parents face à l'évaluation standardisée s'est à ce jour développée dans peu de pays.

Aux États-Unis, les vifs débats développés autour de la loi *No Child Left Behind* (NCLB), notamment au moment de sa récente révision, ont donné lieu à de multiples sondages, dont les résultats sont relativement convergents. Il apparaît tout d'abord que **les parents adhèrent massivement à l'idée de standards et de *testing***. Ainsi, d'après un sondage effectué aux États-Unis (Johnson et Duffet, 2003), 82 % des parents pensent que des lignes directrices claires au sujet des contenus d'enseignement permettent de faire progresser les acquis des élèves. Selon le rapport américain du *MrRel Research Institute* (Goodwin, 2003), fondé sur une soixantaine d'interviews qualitatives de groupes (focus groups), les parents se prononcent pour des standards accompagnés de *testing*, l'absence d'évaluation rendant les standards inutiles à leurs yeux car non appliqués. C'est principalement parce que l'outil leur fournit des informations sur les acquis de leurs enfants dans un univers scolaire qui leur semblent encore peu ouvert à la communication externe qu'ils le soutiennent, même si certains d'entre eux font état de conséquences psychologiques négatives pour leur progéniture. Contrairement aux enseignants, ils sont d'ailleurs nombreux – 80 % – à penser que leurs enfants atteindront les standards fixés par la loi NCLB en mathématiques et en lecture d'ici 2013-2014 contre environ un cinquième des enseignants (un sondage AP-AOL Learning Services, avril 2006) <sup>(21)</sup>.

D'après l'enquête qualitative du *MrRel Institute* (Goodwin, 2003), **ce soutien massif à l'évaluation standardisée n'exclut pas pour autant une prise de recul par rapport à l'instrument. Ainsi, les parents affirment que les écoles ne doivent pas être évaluées sur la seule base des résultats aux tests**. Ce point revient dans de nombreux sondages effectués aux États-Unis, qui montrent également que les parents remettent en cause le fait que les écoles puissent être tenues pour seules responsables des résultats des élèves, le milieu social des familles étant jugé primordial dans la réussite scolaire.

**Autre résultat: les parents perçoivent le *testing* comme inséré dans des dispositifs d'*accountability* envers les autorités administratives plutôt qu'envers les parents et s'inquiètent du fait que ces mesures puissent de fait limiter les communications avec la communauté civile plutôt que les conforter**. Enfin, plusieurs sondages (par exemple, Public Agenda, 2006) montrent également que, **pour les parents, les objectifs cognitifs qu'évaluent les tests ne sont qu'une des missions des écoles. Interrogés sur leurs principales inquiétudes face à l'école, les familles ne citent le bas niveau scolaire qu'en toute fin de classement loin derrière la sécurité, la discipline, le respect pour les enseignants et l'enseignement des valeurs**.

---

<sup>(21)</sup> Le sondage AP-AOL Learning Services, qui a interrogé 810 enseignants et 1085 parents d'élèves depuis la première année de maternelle jusqu'à la douzième année, a été administré en ligne du 13 au 23 janvier par la compagnie *Knowledge Networks*, après que les répondants aient été initialement contactés par téléphone.

**Au Canada, dans la province de Colombie Britannique, les sondages menés auprès des parents ont suscité davantage de polémique**, suite à la vive controverse opposant le Ministère de l'Éducation provincial et les syndicats d'enseignants autour des tests standardisés et en particulier le *Foundation Skills Assessment (FSA)* et le classement du *Fraser Institute*. L'Institut Fraser, *think tank* néo-libéral soutenant le libre choix de l'école et l'évaluation standardisée à forts enjeux, a montré lors d'un sondage mené en avril 2008 que 83 % des parents supportaient largement les tests et que 66 % d'entre eux ne s'opposaient pas à leur utilisation pour classer les écoles. Le syndicat d'enseignant a cependant remis en cause la formulation des questions jugées trop générales et ne portant pas sur le dispositif concrètement mis en œuvre (Vancouver Sun, 17/04/2008) <sup>(22)</sup>.

Les sondages menés au Royaume-Uni (Angleterre) auprès des parents mettent en évidence des traits communs avec les autres pays (House of Commons, 2007). Si les parents consultent les résultats des palmarès, le choix de l'école de leurs enfants ne se fonde que très partiellement sur ces statistiques. Ces indicateurs apparaissent de plus en plus clairs à leurs yeux.

\*

\* \*

**Au total, les réactions des différents acteurs impliqués dans le système éducatif – enseignants, cadres de l'éducation, élèves et parents – apparaissent fortement conditionnées par le contexte dans lequel se développent les dispositifs d'évaluation standardisée. Contrairement à d'autres politiques qui peuvent susciter des phénomènes de rejet ou d'adhésion de principe (la création de filières scolaires dans l'école unique, la décentralisation, l'autonomie scolaire...), il semble que cet instrument certainement parce qu'il est associé à des pratiques traditionnelles de l'école – l'évaluation et la notation – n'entraîne pas un positionnement – positif ou négatif – *a priori*. L'outil apparaît, dans un premier abord, neutre et semble le demeurer quand les conséquences qui y sont attachées demeurent faibles. Il semblerait que ce soit principalement le *testing* à forts enjeux qui suscite des réactions négatives des acteurs et conduise au développement d'effets pervers qui jouent négativement sur les processus d'apprentissage.**

En conclusion, cette revue de la littérature féconde qui s'est développée autour du sujet polémique de l'évaluation standardisée met en évidence à ce jour plusieurs constats:

- Le cadre théorique des effets de l'évaluation standardisée sur les performances des élèves à la fois en termes d'efficacité et d'inégalités scolaires demeure aujourd'hui faible. Les processus qui sont en jeu, les mécanismes intermédiaires par lesquels ces instruments sont censés pouvoir avoir un effet sur les acquis des élèves doivent davantage être détaillés et évalués empiriquement;
- Les recherches empiriques menées sur les effets des dispositifs de *testing* sur les performances des systèmes scolaires ne permettent pas de dessiner, à ce jour, un consensus, car elles présentent des résultats le plus souvent contradictoires. Les réformes fondées sur l'évaluation standardisée, soutenues par une rhétorique politique forte, se sont développées alors que les effets de ces instruments apparaissent encore largement aléatoires. De même, l'évaluation des coûts de ces politiques est encore peu explorée, alors que la filiation théorique de ces réformes les inscrit de façon privilégiée dans la veine économiste, qui se fait fort, à juste titre, de maximiser l'utilisation des ressources publiques, budgétaires en particulier.

---

<sup>(22)</sup> <http://communities.canada.com/vancouver/blogs/reportcard/archive/2008/04/17/parents-support-standardized-tests-poll-shows.aspx>

- Les résultats variables des effets du *testing* peuvent certainement en partie s'expliquer par le fait que cet outil peut s'intégrer dans des modèles d'*accountability* fort différents. De futures recherches sont encore nécessaires dans ce domaine. Cependant, les résultats empiriques obtenus à ce jour permettent déjà de mettre en évidence certains des paramètres cruciaux sur lesquels les décideurs politiques doivent s'interroger lors de la construction ou la rénovation des dispositifs **d'évaluation standardisée (pour une synthèse plus complète sur le sujet, Mons et Pons, 2006)**:
  - De façon générale, un questionnement de fond doit porter, tout d'abord, sur la liaison entre l'évaluation standardisée et d'autres réformes qui si elles ne sont pas mises en lien risquent de limiter les bénéfices potentiels du *testing*. En particulier, une articulation doit être pensée entre, d'un côté, les dispositifs d'évaluation standardisée et, de l'autre, les standards de contenus ou les programmes scolaires, les plans de formation continue des enseignants, le financement des écoles en difficulté et, de façon plus générale, le développement de plans de support pour soutenir les établissements suite à l'analyse qu'ils font de leurs résultats aux tests.
  - Une interrogation est également nécessaire sur l'intensité et la nature des enjeux attachés aux tests à la fois pour les écoles et les enseignants. Quelles sanctions et quelles récompenses peuvent être associées aux résultats des évaluations? Combien de tests sont nécessaires pour donner une image sinon fidèle – ce qui ne sera jamais possible à travers des outils strictement quantitatifs – du moins la moins biaisée possible de la réalité du travail pédagogique réalisé par les écoles? Ne prévoir qu'un seul test comportant de forts enjeux à la fois pour les élèves et les écoles semble porteur de forts effets pervers. Construire une épreuve unique pour atteindre des objectifs multiples (direction des carrières scolaires (promotion, diplomation des élèves...), pilotage des établissements, évaluation générale du système scolaire) conduit à des dysfonctionnements majeurs.
  - Les formes de participation des enseignants aux dispositifs d'*accountability* doivent aussi être pensées. Plus ceux-ci participent à la fois à la conception, l'administration et l'analyse des résultats, plus leur implication dans le processus s'en trouve développée et la culture d'évaluation assimilée. Une réflexion est donc nécessaire sur leur intervention dans le *testing* lui-même (conception, administration...) et, plus largement, dans le développement de modèle d'évaluation interne des établissements qui doivent être mis en lien avec l'*accountability* externe quantitative.
  - **Une réflexion doit enfin être conduite sur la publicité des résultats des tests**: les écoles doivent-elles être classées? Quels indicateurs faut-il présenter aux parents sur quelles bases: le résultat général du pays ou de la région, ceux des écoles – anonymes ou pas – ou seulement ceux de leurs enfants? Quels indicateurs doivent être mis en avant (taux bruts ou indicateurs de valeur ajoutée liés aux contextes qui permettent de tenir compte des publics scolaires accueillis dans les établissements)?

On perçoit à travers la liste longue des questions soulevées ici qu'il n'existe pas de recette toute faite qui permettrait de construire un dispositif d'évaluation standardisée associé uniquement à des effets vertueux. **Les interrogations ouvertes présentées ci-dessus dessinent en creux des modèles multiples d'évaluation standardisée déjà mis en évidence dans les travaux sur les politiques publiques** (Duran et Monnier, 1992): à une **évaluation gestionnaire** (dispositif technique conçu majoritairement par les autorités supérieures administratives pour encadrer les agents dont elles ont la charge) peut très schématiquement s'opposer une **évaluation démocratique** dans laquelle la construction du dispositif et la lecture des résultats des épreuves sont majoritairement le fait du

politique (par opposition à l'administration) et de la société civile. Un **nouveau modèle d'«évaluation professionnelle» (dispositif d'évaluation formative donnant une large place à l'intervention des professionnels de l'éducation qui en sont les premiers utilisateurs) peut certainement compléter ce tableau traditionnel.** Si les modèles peuvent être mixés dans le cadre de la construction concrète des dispositifs d'évaluation standardisée, l'un d'entre eux doit cependant prendre le pas sur les autres pour conduire la philosophie politique générale du projet dont découleront les réponses données aux questions posées précédemment (quels enjeux pour les tests?, quelle publication des résultats des épreuves?, quelle implication des enseignants?...).



## RÉFÉRENCES

---

- Amrein A. L., Berliner D. C. (2002), 'High-Stakes Testing, Uncertainty and Student Learning', *Education Policy Analysis Archives*, vol 10-18
- Amrein-Beardsley A. A., Berliner D.C. (2003), Re-analysis of NAEP math and reading scores in states with and without high-stakes tests: Responses to Rosenshine  
*Education Policy Analysis Archives*, vol 11-25
- Anagnostopoulos D. (2006), "'Real students' and 'true demotes': Ending social promotion and the moral ordering of urban high school", *American Educational Research Journal*, vol 43-1, p. 5-42
- Behrens M. (Ed.) (2006), *Analyse de la littérature critique sur le développement, l'usage et l'implémentation de standards dans un système éducatif*, Neuchâtel: IRDP
- Belair L.M. (2005), Les dérives de l'obligation de résultats ou l'art de surfer sans planche, in C. LESSARD et P. MEIRIEU (Eds), *L'obligation de résultats en éducation*, Bruxelles: De Boeck Université
- Berliner D.C., Biddle, B.J. (1995), *The Manufactured Crisis: Myths, Fraud, and the Attack on America's Public Schools*, Reading, MA: Addison-Wesley
- Bishop J. H., Mane F., Bishop M., Moriarty J. (2001), The role of end-of-course exams and minimum competency exams in standards-based reforms, in D. Ravitch D. (Eds) *Brookings Papers on education Policy 2001*, p. 267-330, Washington D.C.: Brookings Institution
- Bishop J. H. (2006), Drinking from the fountain of knowledge: Student incentive to study and learn. In A. Hanushek et F. Welsh (Eds), *Handbook of the economics of education*, p. 909-944, Amsterdam: North-Holland
- Booher-Jennings J. (2005), "Below the bubble: 'Educational triage' and the Texas accountability System", *American Educational Research Journal*, vol 42-2, p. 231-268.
- Broadfoot P. (2000), «Un nouveau mode de régulation dans un système décentralisé: l'État évaluateur», *Revue Française de Pédagogie*, n° 130, p. 43-55
- Broadfoot P., Osborn M., Sharpe K., Planel C. (2001), Pupil assessment and classroom culture: a comparative study of the language of assessment in England and France, in D. Scott (Ed) *International Perspectives in Curriculum Series Vol. 1: Assessment and the Curriculum*, Greenwood Publishing Group
- Carnoy M., Loeb S. (2002) 'Does external accountability affect student outcomes? A cross-State Analysis', *Educational Evaluation and Policy Analysis*, vol 24-4, p. 305-331
- Center on Education Policy (2006), *From the capital to the classroom: Year 4 of the No Child Left Behind Act*. Mars 2006, disponible://www.cep-dc.org/nclb/Year4/Press
- DeBard R., Kubow P. K. (2002), From compliance to commitment: The need for constituent discourse in implementing testing policy. *Educational Policy*, vol 16-3, p. 387-405
- Demilly L. (2001), Enjeux de l'évaluation et régulation des systèmes scolaires, in L. Demilly (Ed), *Evaluer les politiques éducatives*, Bruxelles: Editions De Boeck Université

- Duran P., Monnier E. (1992), «Le développement de l'évaluation en France. Nécessité techniques et exigences politiques», *Revue Française de Science Politique*, vol. 42, 2, p.235-262
- Duru-Bellat M., Jarousse J.-P. (2001), «Portée et limites d'une évaluation des politiques et des pratiques éducatives par les résultats», *Éducation et Société*, n°8, p. 97-134
- Farkas S., Johnson J., Duffett, A. (2003), *Rolling Up Their Sleeves. Superintendents and Principals Talk about What is Needed to Fix Public Schools*, New York, NY: Public Agenda
- Fredericksen N. (1994), *The influence of minimum competency tests on teaching and learning*. Princeton, NJ: Educational Testing Service
- Goodwin B., Englert K., Cicchinelli L. F. (2002), *Comprehensive Accountability Systems. A Framework for Evaluation*, Mid-continent Research for Education and Learning
- Gordon S. P., Reese M. (1997), 'High-stakes testing: Worth the price?' *Journal of School Leadership*, vol 7, p. 345-368
- Gregory K, Clarke M. (2003), 'High-Stakes Assessment in England and Singapore', *Theory into Practice*, vol 42-1, p. 66-74
- Grissmer, D., Flanagan, A. (1998), *Exploring rapid achievement gains in North Carolina and Texas*. Washington, DC: National Education Goals Panel. (ERIC Document Reproduction No. 425204.)
- Grissmer, D., Flanagan, A. (2001), *Searching for indirect evidence for the effects of statewide reforms*. In D. Ravitch (Ed.), *Brookings papers on education policy*, p. 181-207, Washington, DC: The Brookings Institution
- Grissmer, D.W., Flanagan, A., Kawata, J., Williamson, S. (2000), *Improving student achievement: What state NAEP scores tell us*. Santa Monica, CA: The Rand Corporation
- Haertel E. H., Herman J. L. (2005), A historical perspective on validity arguments for accountability testing. In E. H Haertel., J. L. Herman (Eds) *Use and misuses of data for educational accountability and improvement. The 104<sup>th</sup> Yearbook of the National Society for the Study of Education, Part II*, p. 1-34, Malden MA: Blackwell
- Haney W. (2000) 'The myth of the Texas miracle in education', *Education Policy Analysis Archives*, vol 24-1, Disponible sur le site internet: <http://epaa.asu.edu/epaa/v8n41/>
- Hanushek E., Raymond M. (2004) 'Does School Accountability Lead to Improved Performance? and Student Performance', *Journal of Policy Analysis and Management*, vol 24-2, p. 297-327
- Hanushek E., Raymond M. (2003), 'High-Stakes Research: Accountability works after all', *Education Next*, Version électronique, disponible à l'adresse Internet suivante: <http://www.hoover.org/publications/ednext/3347781.html>
- Hanushek E., Raymond M. (2006), 'School Accountability and Student Performance', *Federal reserve bank of St Louis Regional economic development*, vol 2-1, p. 51-61
- Hargreaves D. H. (2002), *Assessing assessment*, Communication présentée au RSA Lecture programme, 13 février 2002, ([www.rsa.org.uk](http://www.rsa.org.uk))



- Hargrove T. W., Bradford L. H., Richard A., Corrigan S. Z.; Moore C. (2004), 'No Teacher Left Behind: Supporting Teachers as They Implement Standards-Based Reform in a Test-Based Education Environment', *Education*, vol 124-3, p 567-581
- Harris D. N., Herrington C. D. (2006), 'Accountability, Standards and the Growing Achievement Gap: Lessons for the Past Half-century', *American Journal of Education*, 112, p. 209-239
- Helgoy I., Homme A. (2007) 'Towards a New Professionalism in School? A Comparative Study of Teacher Autonomy in Norway and Sweden', *European Educational Research Journal*, vol 6-3, p. 232-249
- Hoffman J., Assaf L., Paris, S. (2001), 'High-stakes testing in reading: Today in Texas, tomorrow?' *The Reading Teacher*, vol 54, p. 482-492
- Hong W-P., Youngs P. (2008) 'Does high-stakes testing increase cultural capital among low-income and racial minority students?', *Education Policy Analysis Archive*, vol 16-6, p. 2-18
- House of Commons (2007), *Testing and assessment: Third report of Session 2007-08*, Children, Schools and Families Committee, London
- Hoxby C. M. (2002), 'The Cost of Accountability,' *NBER Working Papers* 8855, National Bureau of Economic Research, Inc
- IGEN-IGAENR (2005), *Les acquis des élèves, pierre de touche de la valeur de l'école?*, Paris: IGEN-IGAENR
- Jacob B. A. (2001), 'Getting tough? The impact of high school graduation exams', *Educational Evaluation and Policy Analysis*, vol 23-2, p. 99-121
- Jacob B. A. (2002), *Accountability, incentives, and behavior: The impact of high-stakes testing in the Chicago Public Schools*, National Bureau of Economic Research Working Paper 8968
- Jones B. (2007), 'The unintended outcomes of High-Stakes Testing', *Journal of Applied School psychology*, vol 23-2, p. 65-86
- Jones B. D., Egley R. J. (2004), 'Voices from the frontlines: Teachers' perceptions of high-stakes testing'. *Education Policy Analysis Archives*, vol 12-39
- Jones M. G., Jones B. D., Hargrove T. Y. (2003), *The unintended consequences of high-stakes testing*. Lanham, MD: Rowman & Littlefield Publishers, Inc.
- Johnson J., Duffett A. (2003), *Where We Are Now. 12 Things you need to Know about Public Opinion and Public Schools*. New York, NY: Public Agenda
- Klieme E. et al. (2004), *Le développement de standards nationaux de formation: Une expertise*. Bonn: Ministère fédéral de l'Education et de la Recherche (BMBF)
- Lee J. (2008), 'Is Test Driven External Accountability Effective? Synthesizing the Evidence From Cross-State Causal-Comparative and Correlational Studies', *Review of Educational Research*, 78-3, p. 608-644
- Lee J., Wong K. K. (2004), 'The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes', *American Educational Research Journal*, vol 41-4, p. 797-832

- Levacic R. (2001), *An Analysis of Competition and its Impact on Secondary School Examination Performance in England*, National Center for the Study of Privatization in Education, Teachers College, Columbia University Occasional Paper n°34
- Linn R. L. (2000) 'Assessment and accountability', *Educational Researcher*, vol 29-2, p. 4-16
- Linn, R. L. (2001). *The design and evaluation of educational assessment and accountability systems*. Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
- Lipman P. (2004), *High stakes education: Inequality, globalization, and urban school reform*, New York: Routledge Palmer.
- McDonnell L. M. (2005), Assessment and accountability from the policy-maker's perspective, In E. H. Haertel, J. L. Herman (Eds) *Use and misuses of data for educational accountability and improvement*. The 104<sup>th</sup> Yearbook of the National Society for the Study of Education, Part II, p. 1-34, Malden MA: Blackwell
- Maroy C. (2008), 'Vers une régulation post-bureaucratique des systèmes d'enseignement en Europe?', *Sociologie et Société*, vol 40-1, p. 31-55
- Maroy C. (2005). – «Vers une régulation post-bureaucratique des systèmes d'enseignement en Europe?», *Cahier de recherche en Éducation et Formation n° 49*, Louvain-la-Neuve: Université de Louvain
- Maroy C., Cattonar B. (2002). – «Professionnalisation ou déprofessionnalisation des enseignants? Le cas de la Communauté française de Belgique», *Cahier de recherche en 'ducation et Formation n° 18*, Louvain-la-Neuve: Université de Louvain
- Manin B. (1996), *Principes du gouvernement représentatif*, Paris: Flammarion
- McNess E., Broadfoot P., Osborn M. (2003), 'Is the effective compromising the affective?', *British Educational Research Journal*, vol 29-2, p. 243-257
- Goodwin B. (2003), 'Digging deeper: where does the public stand on standards-based education?', *Issues Brief*, Aurora, Colorado: McRel
- Mons N. (2007), *Les nouvelles politiques éducatives*, Paris: PUF
- Mons, N., Pons, X. (2006). *Les standards en éducation dans le monde francophone. Une analyse comparative*, Neuchâtel: IRDP
- Mons, N., Pons, X. (2009), Pourquoi le pilotage par les résultats? Une mise en perspective théorique et historique de ce nouveau mode de gouvernance, In Mons N, Emin J.-C., P. Santana, *Piloter par les résultats*, Paris: CNDP
- Nichols S. L. (2007), 'High-Stakes Testing: Does It Increase Achievement?', *Journal of Applied School Psychology*, vol 23-2, p. 47-64
- Nichols S. L., Glass G.V., Berliner D.C. (2006), 'High-Stakes Testing and Student Achievement: Does Accountability Pressure Increase Student Learning?', *Education Policy Analysis Archives*, vol 14-1

- McNeil L. M. (2005), Faking equity: High-stakes testing and the education of Latino youth. In A. Valenzuela (Ed), *Leaving children behind: How 'Texas-style' accountability fails Latino youth* (p.57-111). Albany, NY: State University of New York Press
- OCDE (2007), *PISA 2006 Science Competencies for Tomorrow's World, Volume 1*, Paris: OCDE
- Osborn M. (2006), 'Changing the context of teachers' work and professional development: a European perspective', *International Journal of Educational Research*, vol 45, p. 242-253
- Public Agenda (2006), *Is support for standards and testing fading?* Reality Check 2006, Issue n°3, N.Y.: Public Agenda
- Phelps R. P. (2005), *Defending standardized testing*, Mahwah, NJ: Erlbaum
- Pons, X. (2008), *L'évaluation des politiques éducatives et ses professionnels. Les discours et les méthodes (1958-2008)*, thèse de doctorat de science politique, IEP Paris,
- Treisman P. U., Fuller E. J. (2001), *Comment by Philip Uri Treisman and Edward J. Fuller*. In D. Ravitch (Ed.), *Brookings papers on education policy*, p. 208-218, Washington DC: The Brookings Institution
- Roderick M., Jacob B. A., Bryk A. S. (2002), 'The impact of high stakes-testing in Chicago on student achievement in promotional gate grades', *Educational Evaluation and Policy Analysis*, vol 24-4, p.333-357
- Roderick M., Nagaoka J. (2005), 'Retention under Chicago's high-stakes testing program: Helpful, harmful, or harmless?' *Educational Evaluation and Policy Analysis*, vol 27-4, p. 309-340
- Rosenshine B. (2003), High-stakes testing: Another analysis. *Education Policy Analysis Archives*, vol 11-24
- Woessmann L. (2007), 'International Evidence on School, Competition, Autonomy and Accountability: a Review', *Peabody Journal of Education*, vol 82-2-3, p. 473-497
- Zanten Van A. (1999), Les chefs d'établissements et la justice des systèmes d'enseignement en Angleterre et en France: les pratiques et les éthiques professionnelles à l'épreuve de la concurrence entre établissements. In D. MEURET (Ed). *La justice du système éducatif*. Bruxelles: Département De Broeck Université

