REVUE FRANÇAISE DE PÉDAGOGIE

Revue française de pédagogie
Recherches en éducation

189 | octobre-novembre-décembre 2014
L'internat et ses usages, d'hier à aujourd'hui

Les effets de l'évaluation externe sur les pratiques enseignantes : une revue de la littérature

The effects of external assessment on teaching methods: A literature review

Esteban Rozenwajn et Xavier Dumay

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/rfp/4636
DOI : 10.4000/rfp.4636
ISSN : 2105-2913

Éditeur
ENS Éditions

Édition imprimée
Date de publication : 31 décembre 2014
Pagination : 105-138
ISBN : 978-2-84788-678-8
ISSN : 0556-7807

Référence électronique
Esteban Rozenwajn et Xavier Dumay, « Les effets de l'évaluation externe sur les pratiques enseignantes : une revue de la littérature », *Revue française de pédagogie* [En ligne], 189 | octobre-novembre-décembre 2014, mis en ligne le 31 décembre 2017, consulté le 19 avril 2019. URL : http://journals.openedition.org/rfp/4636 ; DOI : 10.4000/rfp.4636

Les effets de l'évaluation externe sur les pratiques enseignantes : une revue de la littérature

Esteban Rozenwajn
Xavier Dumay

L'utilisation de l'évaluation externe des élèves comme outil de régulation des systèmes scolaires et des pratiques enseignantes suscite un nouvel engouement politique depuis les années 1980 et 1990. Cet engouement s'inscrit en partie dans la croyance en la capacité d'un tel outil à contribuer à l'amélioration des pratiques enseignantes et par là des « performances » des systèmes scolaires. Cet article cherche à faire un état de la littérature ayant considéré les effets des évaluations externes sur les pratiques mises en place par les enseignants.

Mots-clés (TESE) : évaluation externe, évaluation des enseignants, pratique pédagogique, autonomie des enseignants, programme d'études, profession d'enseignant.

Introduction

Le recours à des évaluations standardisées élaborées par des structures indépendantes des établissements scolaires et imposées à l'ensemble des élèves d'un niveau et d'une région déterminée n'est pas une pratique nouvelle dans les systèmes scolaires modernes. Aux États-Unis, l'*Iowa Test of Basic Skills* (ITBS) est présenté aux élèves du niveau 8 (13-14 ans) depuis 1935 dans l'Iowa (Carnoy & Loeb, 2002). Du côté européen, l'Islande instaure un test à la fin de l'enseignement primaire dès 1946 ; l'Irlande du Nord et le Portugal appliqueront des épreuves similaires en 1947 (Eurydice, 2009). Néanmoins, le recours à l'évaluation externe comme outil de régulation des systèmes scolaires et des pratiques enseignantes semble susciter un engouement politique particulièrement prononcé depuis les années 1980.

Un tel engouement aura impliqué le déplacement d'une logique de régulation du système éducatif par l'« *input* » à une logique de pilotage par l'« *output* » qui trouve sa justification dans la volonté d'améliorer la qualité de l'enseignement. C'est ainsi qu'Elmore et ses collaborateurs précisent qu'« en principe, la focalisation sur la performance des élèves devrait éloigner les États [nord-américains] d'une régulation par l'*input* – évaluer les établissements sur la base du nombre de livres dans la bibliothèque et la proportion de personnel qualifié, par exemple – vers un modèle de pilotage par les résultats – utilisant des récompenses, des sanctions, et une assistance pour faire avancer les établissements vers des niveaux de performance plus élevés. En d'autres termes, la reddition de comptes en éducation devrait orienter l'attention des établissements sur l'amélioration de l'apprentissage pour les élèves plutôt que sur la conformité envers les règles » (Elmore, Abelmann & Fuhrams, 1996, p. 65). Dans une perspective similaire, du côté européen, le rapport Eurydice dédié aux évaluations standardisées (2009) considère ce changement de perspective comme une conséquence du processus plus général de décentralisation des systèmes éducatifs entamé en Europe dans les années 1980 :

> À travers l'Europe, un mouvement en faveur de la décentralisation et de l'autonomie des établissements scolaires a commencé à voir le jour dans les années 1980, avec une tendance générale au sein des systèmes éducatifs à passer de cadres réglementaires normatifs à des cadres politiques permettant une plus grande participation démocratique et accordant davantage de libertés aux établissements, mais créant aussi de nouvelles mesures d'évaluation des résultats pédagogiques (Eurydice, 2009, p. 19-20).

D'un point de vue théorique, cet intérêt pour les résultats fait notamment écho à divers courants influents du domaine des politiques publiques tels que le *New Public Management* et le mouvement de la *Policy Evaluation* ; ainsi qu'à divers courants associés plus directement à l'éducation comme l'analyse de la fonction de production en économie de l'éducation et la *School Effectiveness Research* (Mons, 2009). Les évaluations standardisées serviraient ainsi à orienter le comportement des acteurs dans le contexte d'un désengagement de l'État prôné par le *New Public Management* et d'outil d'évaluation des politiques dans le cadre séquentiel proposé par les tenants de la *Policy Evaluation*. Les résultats d'épreuves de compétences scolaires, pas seulement externes, apparaissent également comme le critère par lequel peuvent être distinguées les « bonnes » des « mauvaises » pratiques dans le cadre pragmatique de la *School Effectiveness Research*.

Toutefois, si les dispositifs d'évaluation externe partagent une même matrice et certains objectifs, leur diversité est malgré tout encore très importante à l'heure actuelle. Les critères sur la base desquels ils peuvent se distinguer les uns des autres sont nombreux. Ils peuvent évidemment varier sur la base de leur contenu : le type de questions et les matières évaluées. Les questions peuvent prendre la forme de questions à choix multiples, de réponses courtes ou de réponses ouvertes, voire d'une combinaison de ces différentes possibilités dans la plupart des cas. Il est courant que tous les élèves soient soumis à des versions identiques d'une même épreuve, mais d'autres configurations peuvent être envisagées. Les épreuves informatisées présentées tout au long de la scolarité obligatoire au Danemark présentent ainsi la particularité de s'adapter constamment au niveau de l'élève en fonction des bonnes ou mauvaises réponses qu'il fournit. Les matières évaluées comprennent le plus souvent la lecture/écriture de la langue du pays et les mathématiques, bien que d'autres disciplines soient également soumises à l'évaluation dans de nombreux cas.

Les dispositifs d'évaluation externe peuvent également se distinguer par les modalités et les acteurs chargés de la conception, de la passation et de la correction des épreuves. L'élaboration des épreuves peut ainsi être prise en charge par un organisme public, avec ou sans interventions externes ; par un organisme privé ; ou par une université. L'administration et la correction des épreuves peuvent à leur tour être prises en charge par les enseignants titulaires des classes concernées, par d'autres enseignants ou par des personnels externes à l'établissement.

Les épreuves externes peuvent aussi se distinguer sur la base de leur fréquence, du niveau auquel elles sont imposées ou des enjeux qu'elles impliquent pour les élèves. À cet égard, les épreuves dites diagnostiques (ou formatives), contrairement aux épreuves dites certificatives (ou sommatives), ne sont pas déterminantes quant au parcours scolaire d'un élève et servent principalement à informer les enseignants du niveau et des difficultés de leurs élèves.

Cependant, en dépit de toutes ces variations possibles, il est commun dans la littérature scientifique de distinguer les épreuves externes en fonction des conséquences qui sont associées aux résultats pour les enseignants et les établissements scolaires. D'une part, sont considérés comme des « dispositifs d'évaluation à enjeux élevés » (*high-stakes tests*) les dispositifs où les résultats de l'évaluation externe sont associés à des conséquences relativement importantes pour les enseignants ou les établissements. Concrètement, ces conséquences peuvent impliquer un risque de licenciement en cas de résultats insatisfaisants récurrents, ou des sanctions et/ou des récompenses financières dépendantes des résultats obtenus. En outre, ce type de dispositifs s'intègre généralement dans un ensemble plus large de mesures de reddition de comptes régulièrement désignées par la notion d'« *high-stakes accountability* », parmi lesquelles figure souvent la diffusion médiatique des résultats aux épreuves.

D'autre part, sont considérés comme des « dispositifs à faibles enjeux » (*low-stakes tests*) les dispositifs dont les résultats de l'évaluation externe n'entraînent pas de conséquences formelles sur le statut professionnel de l'enseignant ou sur l'établissement. En raison du contraste avec les dispositifs à enjeux élevés, ce genre de dispositifs essentiellement développés en Europe continentale est parfois désigné par la notion de « *low-stakes accountability* » ou « *soft accountability* ».

Dans tous les cas néanmoins, les dispositifs d'évaluation externe présentent la volonté d'influencer les pratiques pédagogiques dans le sens d'une amélioration de la qualité de l'enseignement. L'argument commun à tous les auteurs prônant le recours à l'évaluation externe suppose ainsi que les résultats devraient fournir un *feedback* permettant aux acteurs de terrain d'améliorer leurs compétences. Dans le cadre des dispositifs d'évaluation à faibles enjeux, cette logique réflexive serait suffisante à produire un tel processus (Klieme, 2004 ; Thélot, 2002). Ainsi Claude Thélot s'exprime-t-il en ces termes :

> […] l'évaluation est capitale : comme levier, comme outil de régulation interne, comme soutien, comme seul moyen pour que les professeurs, mais aussi les cadres de l'école, améliorent leurs façons de faire. Pourquoi ? Parce que l'on crée un « effet miroir ». Dès lors que vous allez donner à ces professeurs et à ces cadres le résultat de leur action, si ce résultat n'est pas conforme à ce qu'ils souhaitent ou à ce qu'on souhaite qu'ils fassent, ils changeront leur façon de faire (Thélot, 2002, p. 325-326).

Pour les auteurs prônant les modèles aux enjeux élevés par contre (Phelps, 2005 ; Raymond & Hanushek, 2003 ; Woessman, 2007), le moteur du changement pédagogique serait la motivation de l'individu à éviter les sanctions où à obtenir des récompenses. Le *feedback* constitué par les résultats ne serait, par lui seul, pas suffisant pour produire les transformations susceptibles d'aboutir à l'amélioration de la qualité de l'enseignement. Dans les termes d'Hanushek et Raymond donc :

> Ces systèmes se basent sur l'idée que la préoccupation pour les résultats des élèves mènera à des changements comportementaux de la part des élèves, enseignants, et établissements afin d'atteindre les objectifs du système. Une partie de cet effet est supposé plus ou moins automatique. Mais une partie vient aussi du développement d'incitants explicites qui mèneront à l'innovation, à l'efficience et à la résolution de n'importe quel problème de performance (Hanushek & Raymond, 2002, p. 81).

Dans le cadre des dispositifs d'évaluation externe donc, les attentes formulées à l'égard des enseignants changent : ils ne sont plus simplement appelés à enseigner un programme et à se conformer aux divers prescrits institutionnels, mais aussi à orienter et à modifier leurs

pratiques pédagogiques en fonction des informations qui leurs sont transmises par les résultats. Aussi, leur position vis-à-vis de l'évaluation change. Si l'évaluation supposait traditionnellement une distinction claire entre un enseignant évaluateur et un élève évalué sur lequel reposait la responsabilité de la réussite, la possibilité de comparer les classes, établissements, régions ou pays au travers des résultats des épreuves standardisées place potentiellement les enseignants dans une position d'évalué où ils auraient également une part de responsabilité quant à la performance de leurs élèves.

Mais qu'en est-il de l'impact réel des dispositifs d'évaluation externe sur le travail enseignant ? Telle est la question autour de laquelle s'articule cet article et à laquelle nous tenterons de répondre à travers une revue de la littérature empirique sur le sujet. Notre intérêt se porte donc spécifiquement sur les études empiriques dont l'objet fait référence à la façon dont les enseignants réagissent aux dispositifs d'évaluation externe, que nous avons tenté de rassembler à travers des recherches sur des bases de données diverses telles que *google scholar*, ERIC, *scopus*, *Education Policy Analysis Archives*, *Francis*, *Erudit* ou CAIRN. Les mots-clés ayant permis la localisation de ces études se sont essentiellement basés sur une combinaison du terme « enseignant » (ou « *teacher* ») avec les différents termes susceptibles de désigner les épreuves standardisées sur lesquelles portait notre intérêt : *évaluation externe*, *évaluation standardisée*, *reddition de comptes*, *accountability*, *external testing*, *high/low-stakes testing*, *national assessments*, *school accountability*, *standardized achievement tests*, *school performance feedback systems*. Après un premier survol des articles repérés, nous avons également élargi les mots-clés à des termes faisant référence à des pratiques régulièrement associées et observées dans le cadre des dispositifs d'évaluation externe, à savoir les mots *bachotage*, *cheating* (fraude), *teaching to the test*.

La sélection des références à travers cette procédure s'est limitée aux publications dans les revues scientifiques, aux rapports institutionnels[1] et aux chapitres d'ouvrages. Une lecture attentive des résumés des articles susceptibles de correspondre à notre question de recherche nous a finalement permis d'aboutir à une sélection finale de 77 publications : 62 faisant référence à des études menées dans le contexte de dispositifs aux enjeux élevés nord-américains et anglais ; et 15 faisant référence à des études menées dans le contexte de dispositifs à faibles enjeux du continent européen. L'ensemble de ces études sont référencées en annexe.

La suite de cette note de synthèse vise à rendre compte des principaux résultats mis au jour par ces études empiriques concernant les pratiques enseignantes, tant dans le cadre des dispositifs aux enjeux élevés que dans le cadre de ceux à faibles enjeux. Mais avant cela, afin de rendre compte des contextes dans lesquels sont menées les études sélectionnées, nous poursuivons notre introduction en décrivant plus précisément les dispositifs d'évaluation externe dans les pays anglo-saxons d'une part et sur le continent européen d'autre part.

Les dispositifs à enjeux élevés

Malgré la popularité que les dispositifs à enjeux élevés ont eue au cours des dernières décennies, l'idée d'associer des conséquences importantes pour les enseignants aux performances de leurs élèves n'est pas récente. Déjà en 1862, sous l'impulsion du vice-président du ministère de l'Éducation de l'époque, Robert Lowe, l'Angleterre et le Pays de Galles instaurent un système de financement des établissements primaires reposant sur les notes attribuées aux élèves par les inspections annuelles qui se maintiendra jusqu'en 1897 (Rapple, 1994). L'évaluation des élèves en écriture, lecture et arithmétique (les dits « *three Rs* ») était alors prise en charge par

1 Dans certains cas, les études sont menées par des chercheurs dans le cadre d'une structure privée ou publique. Les résultats de leurs travaux sont alors généralement diffusés sous forme de rapports plutôt que dans des revues scientifiques.

les inspecteurs eux-mêmes selon une procédure standardisée et déterminait la paie de l'enseignant sur la base du nombre de réussites. Dans le contexte nord-américain par ailleurs, des exemples de variations salariales liées aux résultats semblent traverser le XXe siècle (Linn, 2000). Ce qui est davantage caractéristique des 25 dernières années cependant, c'est probablement la croissance exceptionnelle, dans les pays anglophones, de dispositifs à enjeux élevés en une période relativement courte.

Les États-Unis

Aux États-Unis, la décennie des années 2000 aura été marquée par l'imposition de mesures « *high-stakes* » au niveau fédéral. En effet, si les États sont les acteurs pionniers du développement de dispositifs aux enjeux élevés dans les années 1990, le *No Child Left Behind Act* (NCLB) instauré par l'administration Bush en 2001 impose de fait la mise en place d'épreuves standardisées auxquelles sont associées des conséquences importantes pour les enseignants et les établissements sur l'ensemble du territoire nord-américain. Sous cette réforme, le versement des fonds fédéraux aux États est conditionné par la mise en place d'épreuves standardisées permettant l'évaluation des élèves des niveaux 3 (8-9 ans) et 8 (13-14 ans) en lecture et mathématiques, pour lesquelles il est demandé aux États de définir un rythme de progression annuel (*Adequate Yearly Progress* [AYP]) au terme duquel devait être atteint un taux de réussite de 100 % pour l'année académique 2013-2014. La définition du contenu de l'épreuve et des standards de performances à atteindre demeurait toutefois entre les mains des États.

Les objectifs de progression s'appliquaient à toutes les écoles publiques et les États devaient maintenir un registre permettant d'identifier les établissements ne parvenant pas à les atteindre. De plus, les établissements défavorisés bénéficiant des financements fédéraux *Title I*[2] ne parvenant pas à atteindre leurs objectifs de progression pendant plus de deux années consécutives étaient soumis à un régime de sanctions particulier. Après deux années en dessous des standards de performances, ces établissements étaient sommés de proposer un plan d'amélioration (*school improvement plan*) et de dédier au moins 10 % des fonds fédéraux à la formation continue des enseignants. Dans le cas où les objectifs de progression n'étaient toujours pas atteints par la suite, une série de « mesures correctives » était progressivement imposée à l'établissement. À partir d'une troisième année consécutive en dessous des objectifs de progression s'ouvrait ainsi la possibilité pour les parents de changer leurs enfants d'établissement. À partir de la quatrième année dans la même situation, l'établissement devait mettre en place une série de « services éducatifs supplémentaires » (par exemple des séances de tutorat extra-scolaire). Au terme de 5 années consécutives sans parvenir à atteindre les standards prédéterminés par l'État, l'établissement était soumis à un plan de restructuration déterminé par le district qui devait devenir effectif à la septième année de sous-performance. Ce dernier pouvait alors choisir une des options de restructuration parmi les suivantes : remplacement de l'équipe éducative, prise en charge de la gestion de l'établissement par un organisme privé, réouverture de l'école en tant que « *Charter School* »[3] ou prise en charge par l'État (*state takeover*). Dans le cas où aucune de ces solutions ne convenait, une dernière option permettait aux districts de déterminer eux-mêmes d'autres formes de restructuration.

La sévérité des sanctions prévues par le NCLB paraît donc importante, mais, dans les faits, leur application s'est avérée limitée par l'autonomie dont faisaient preuve les États et les districts. En premier lieu, l'autonomie des États en ce qui concerne la définition du contenu de

2 Pour être considéré comme relevant du *Title I* et bénéficier des fonds fédéraux qui sont associés à cette catégorie, un établissement doit comprendre un minimum de 40 % d'élèves vivant en dessous du seuil de pauvreté.

3 Les *Charter Schools* sont des établissements scolaires bénéficiant d'une autonomie plus importante que les écoles publiques. En contrepartie néanmoins, les financements publics attribués aux *Charter Schools* sont généralement moins importants que ceux attribués aux *Public Schools*.

l'épreuve et des standards de performances à atteindre leur donnait une certaine marge de manœuvre pour éviter que les établissements ne se retrouvent en dessous des objectifs de performance, au moins à moyen terme. En jouant sur la difficulté des épreuves et sur le rythme de progression imposé, les établissements se retrouvaient dans une situation plus ou moins contraignante en fonction de l'État où ils se trouvaient. De telles variations permettaient ainsi de comprendre des situations telles qu'observées dans l'année académique 2010-2011, où seuls 11 % des établissements du Wisconsin se trouvaient en dessous des objectifs de progression contre 89 % en Floride (Gamoran, 2012) (en partant du principe que ces différences n'étaient pas dues à des variations exceptionnelles dans les compétences des enseignants ou des élèves). Toutefois, l'imposition de l'objectif des 100 % de réussite pour 2014 obligeait les États à déterminer une courbe de progression qui placerait inévitablement un nombre croissant d'établissements scolaires en sous-performance au fur et à mesure de l'approche de l'échéance. Les données officielles du gouvernement fédéral indiquent ainsi que pour l'année académique 2005-2006, 9 903 établissements scolaires étaient « en besoin d'amélioration » (situés en dessous des standards de performance pendant au moins 2 années consécutives) sur l'ensemble du territoire ; tandis qu'ils étaient 19 270 pour l'année académique 2012-2013. Parallèlement, le nombre d'établissements soumis à des plans de restructuration (situés donc en dessous des objectifs de performance pendant au moins 6 années consécutives) était également en nette augmentation : pour 925 établissements dans cette situation au cours de l'année académique 2005-2006, ils étaient 6 105 pour l'année académique 2011-2012.

Néanmoins, une fois l'établissement soumis à un plan de restructuration, le district scolaire détenait à son tour une marge de manœuvre assez large en raison de la possibilité de définir lui-même les mesures à prendre dans le cas où l'option « autre mesure de restructuration majeure » était retenue. Le remplacement de l'équipe éducative ou de la direction pouvait ainsi être évité. Malheureusement, les données concernant la proportion de choix effectués pour chacune des options de restructuration possibles ne semblent pas disponibles au niveau fédéral. Elles le sont cependant au niveau des rapports annuels étatiques (*Consolidated State Reports*), qui précisent aussi en quoi consistent les mesures tombant sous cette catégorie « autres ». Pour l'État de Californie par exemple, qui semble constamment en tête du peloton des États dans lesquels le nombre d'établissements n'atteignant pas les objectifs de progression est le plus important, 80 % des 340 établissements en restructuration pour l'année académique 2009-2010 étaient soumis à cette catégorie « autres ». Parmi les mesures entrant dans cette catégorie se trouvaient notamment la participation des enseignants à des équipes de coordination (les dites *Professional Learning Communities*) ou la collaboration avec des équipes de soutien liées au district.

Par ailleurs, l'approche de l'échéance de 2014 semble avoir rendu évidente l'impossibilité d'atteindre le seuil des 100 % de réussite. L'administration Obama offre alors en 2011 la possibilité pour les États de renoncer à cet objectif ainsi qu'à l'obligation de mettre en place les services d'éducation supplémentaires. Les conditions pour l'obtention d'une telle dérogation étaient de 1) maintenir des standards et des évaluations permettant de mesurer la performance et la progression des élèves, 2) d'élaborer un système de reddition de comptes permettant de reconnaître les établissements aux performances et progression élevées, et 3) d'élaborer des systèmes d'évaluation des enseignants et des directions. Pour autant qu'ils soient acceptés par le comité d'experts chargé de se prononcer sur les nouveaux programmes d'évaluation, cette procédure de dérogation diminuait l'importance qui était attribuée aux résultats des épreuves externes sous le NCLB puisqu'elle offrait la possibilité aux États de baser l'évaluation des enseignants et des établissements sur une diversité d'indicateurs de performance.

Cette possibilité d'obtenir des dérogations aux exigences établies par le NCLB semble d'ailleurs attractive puisque 43 des 50 États en bénéficient depuis 2012, et deux autres sont actuellement en procédure pour en obtenir. Ainsi, les standards de performance étant

désormais beaucoup plus flexibles, le nombre d'établissements en sous-performance sur l'ensemble du territoire est passé de 19 270 pour l'année académique 2011-2012 à 10 647 pour l'année académique 2012-2013.

En tous les cas, l'échéance prévue par le NCLB étant désormais passée sans qu'aucun des États ne soit parvenu à atteindre les 100 % de réussite, le gouvernement fédéral doit sous peu se prononcer sur une nouvelle politique éducative fédérale. Rien n'est encore certain donc, mais il est fort probable que le régime de sanctions imposé par le NCLB soit remis en question. Encore récemment, dans un discours prononcé le 12 janvier 2015, le secrétaire d'État à l'éducation, Mr Arne Duncan, déclarait vouloir « […] travailler ensemble – Républicains et Démocrates – pour aller au-delà de la dépassée, exténuée et prescriptive législation du *No Child Left Behind* ».

Le plus probable sera donc d'assister à la disparition formelle des AYP, ce qui est déjà le cas pour les États qui se sont vu accorder les dérogations. Cependant, les interventions de l'administration Obama dans le domaine de l'enseignement laissent supposer que le recours à l'évaluation standardisée des élèves pour évaluer, et éventuellement sanctionner/récompenser le travail enseignant, n'est pas près de disparaître pour autant. Outre le fait que le maintien d'épreuves standardisées et de systèmes d'évaluation des enseignants et directions était dès le départ des conditions nécessaires à l'octroi d'une dérogation aux règles définies par le NCLB, le concours initié par l'administration Obama « *Race to The Top* » auquel ont participé la plupart des États incitait à la mise en place d'un système d'évaluation des enseignants en partie basé sur l'évolution des résultats des élèves. Les États qui mettaient en place ce genre de dispositifs se voyaient en effet attribuer des points supplémentaires permettant d'améliorer leur position dans cette compétition où les meilleures places permettaient l'obtention de financements supplémentaires pour l'enseignement. L'enjeu était ainsi de récompenser les enseignants « hautement efficaces » tout en permettant de « soustraire les enseignants titulaires ou non titulaires inefficaces ayant eu d'amples opportunités d'amélioration […] » (US Department of Education, 2009, p. 9).

De plus, à travers ce concours interétatique, l'administration Obama est parvenue à s'attaquer au problème de l'hétérogénéité des standards de performance en octroyant des points supplémentaires aux États adoptant l'ensemble des standards proposés par le *Common Core State Standards Initiative* (CCSSI). Ils sont ainsi désormais 43 États à avoir adopté ces standards qui, sans pour autant l'éliminer totalement, réduisent la diversité potentielle des épreuves conçues.

En fin de compte, malgré un assouplissement des mesures définies par le NCLB pendant les mandats d'Obama, les résultats des épreuves standardisées paraissent toujours associés à des conséquences importantes pour les enseignants et établissements nord-américains pouvant prendre la forme de sanctions et/ou de récompenses. De plus, l'autonomie dont les États font toujours preuve en matière de définition des politiques éducatives entraîne une hétérogénéité des dispositifs d'*accountability* susceptible de placer les enseignants dans des situations plus ou moins contraignantes en fonction des États ou districts au sein desquels ils exercent leur métier.

L'Angleterre

Du côté européen, l'Angleterre est souvent considérée comme le seul pays européen disposant d'un système d'*accountability* aux enjeux élevés pour les établissements et les enseignants. Dès 1988, l'*Education Reform Act* restreint l'autonomie dont faisaient preuve les établissements et les autorités locales en matière de définition des contenus enseignés par l'instauration d'un curriculum national où sont définis des objectifs d'apprentissage pour quatre « niveaux-clés » (*Key Stages* [KS]), concernant les élèves de 5-7 ans (KS1), 8-11 ans (KS2), 11-14 ans (KS3) et 14-16 ans (KS4). Dans la foulée sont également prévues des épreuves standardisées obligatoires, les *National Curriculum Assessments* (aussi connues sous la dénomination de « SATs »), pour les KS1 (1991), KS2 (1995) et KS3 (1998), auxquelles sont associés des objectifs de réussite. Dans une logique de

reddition de comptes visant à informer les parents de l'offre éducative, l'*Education Reform Act* prévoit aussi la diffusion publique des résultats aux SATs (par établissement), largement médiatisés par la presse sous forme de classements : les *league tables*. Les évaluations externes anglaises s'inscrivent donc, bien plus qu'aux États-Unis, dans une reddition de comptes de type marchand (Harris & Herrington, 2006), qui a donné lieu à des phénomènes de stigmatisation des établissements peu performants aussi désignés par le « *naming and shaming* ».

De plus, toujours en mobilisant l'argument de renforcer la qualité de l'enseignement, le système d'inspection subit une réforme en 1991 au travers de laquelle un nouveau nom lui est attribué : *Office for Standards in Education, Children's Services and Skills* (OFSTED). Derrière ce changement de nom change également la mission du service, qui devient une structure indépendante et qui n'est plus seulement amené à fournir un descriptif du système éducatif, mais aussi à contribuer à son amélioration. Les inspections, désormais plus fréquentes, aboutissent alors à des rapports qui sont également publiés et dans lesquels l'OFSTED attribue une note globale à l'établissement en se basant notamment sur les scores de réussite aux SATs. Cette note varie entre 1 et 4 pour les qualifications suivantes : exceptionnel (1), bon (2), requière amélioration (3) et inadéquat (4). Les établissements tombant dans la catégorie « inadéquat » s'exposent alors au risque de se voir soumis à des « mesures spéciales » qui peuvent voir l'inspection requérir le changement de l'équipe éducative ou la fermeture de l'établissement.

Dans les faits toutefois, les mesures les plus dramatiques demeurent rares. Entre 2009 et 2014, seuls 2 % des établissements se retrouvent dans la catégorie 4, et tous ne sont pas soumis à des changements de personnel ou à des fermetures. De plus, l'impact de la diffusion médiatique des résultats sur le choix des parents semble limité. Certaines enquêtes indiquent ainsi que les classements de performance ne constituent pas une source d'information privilégiée par les parents pour déterminer leur établissement de préférence (House of Commons, 2008).

Le recours aux épreuves standardisées et, plus généralement, le système de reddition de comptes dans son ensemble seront toutefois remis en question dans le courant des années 2000. Sous la pression des parents réticents à ce que leurs enfants de 7 ans soient confrontés à un cadre aussi formel que celui des SATs, les épreuves de KS1 seront supprimées au profit d'évaluations « internes » menées par les enseignants dès 2004. En 2009 seront également retirées les épreuves de KS3, de telle sorte que les SATs de KS2 sont à l'heure actuelle les seules épreuves standardisées obligatoires dans la scolarité des élèves anglais et portent sur l'anglais et les mathématiques. D'autres épreuves standardisées telles que les *General Certificate of Secondary Education* (GCSE) sont toutefois présentées à la grande majorité des élèves dans le secondaire et constituent également des indicateurs de performance centraux dans la reddition de comptes et pour les interventions de l'OFSTED.

En parallèle, la remise en cause du dispositif d'*accountability* est discutée au Parlement au cours de l'année 2009, à partir d'un rapport présenté par le *Children, Schools and Families Committee* (House of Commons, 2009). Celui-ci dénonce les effets délétères de la publication des résultats sous forme de *league tables* et plaide pour une nouvelle planification de la reddition de comptes basée sur des fiches d'établissement individuelles (*School Report Cards*) tentant d'intégrer une diversité d'indicateurs. Aussi, le même rapport questionne l'importance donnée aux scores des élèves lors des inspections et invite également l'OFSTED à diversifier les indicateurs utilisés pour l'élaboration de ses rapports. Tout comme dans les dernières réformes observées aux États-Unis donc, l'enjeu est également de réduire le poids accordé aux scores des élèves dans l'évaluation des établissements et des enseignants.

Pour autant, même si les interventions de l'OFSTED se sont adoucies dans le sens où elles sont moins fréquentes et censées se baser davantage sur des données qualitatives telles que les observations en classe et les interactions avec les enseignants, les scores demeurent une préoccupation importante. Des classements de performance des établissements en fonction des scores aux épreuves standardisées sont ainsi toujours accessibles sur le site web du

Département de l'éducation. En outre, celui-ci continue d'établir des objectifs de performances progressifs basés sur la réussite des élèves aux épreuves standardisées. Ainsi par exemple, un établissement du secondaire était considéré en sous-performance s'il n'atteignait pas les 60 % d'élèves atteignant un niveau 4 ou supérieur aux épreuves de KS2 en lecture, écriture et mathématiques pour l'année 2013. Ce seuil passe ensuite à 65 % pour l'année 2014 et progresse régulièrement pour atteindre les 85 % en 2016.

Il convient finalement de préciser que le système de reddition de comptes anglais ne repose pas exclusivement sur des sanctions potentielles. Des récompenses pour les enseignants apparaissent en effet sous la forme de récompenses salariales (les dites « *Upper Pay Scales* ») basées sur des indicateurs de performance mesurables parmi lesquels peuvent se retrouver les résultats aux épreuves standardisées.

Les dispositifs à faibles enjeux

Contrairement aux pays anglophones, l'ensemble des pays du continent européen se sont orientés vers la mise en place de dispositifs de reddition de comptes à faibles enjeux. Au sein de ceux-ci, les évaluations externes sont censées assumer au moins une des 3 fonctions suivantes : la prise de décisions sur le parcours scolaire des élèves ; l'identification des besoins des élèves par les évaluations dites formatives ; et le pilotage du système éducatif dans son ensemble, c'est-à-dire la collecte d'informations permettant de contrôler les performances et d'orienter la politique éducative.

Quel que soit leur objectif néanmoins, la mise en place de dispositifs d'évaluation externe en Europe continentale apparaît à nouveau comme une tendance prononcée à partir des années 1990 qui semble s'accentuer dans la décennie 2000 avec l'impact médiatique des enquêtes internationales PISA. Ainsi, ne prenant en compte que les dispositifs d'évaluation externe destinés aux élèves de 16 ans ou moins, le rapport Eurydice de 2009 dénombre 17 pays ou régions ayant développé des dispositifs d'évaluation externe entre 1946 et 1990 (donc pour une période de 44 ans) ; tandis qu'ils sont respectivement 11 et 19 à en avoir développé pour les décennies 1990 et 2000 (donc 30 dispositifs en tout, pour une période de 20 ans). En fin de compte, seuls 5 pays ou régions (Communauté germanophone de Belgique, République tchèque, Grèce, Pays de Galles et Liechtenstein) ne disposent d'aucune évaluation standardisée des élèves au terme de la décennie précédente.

Le cas de la Communauté française de Belgique est assez illustratif à cet égard. Suite aux rapports de l'OCDE des années 1990 et aux études PISA en 2000 qui pointent du doigt la faiblesse du système scolaire belge francophone, la mise en place d'un dispositif d'évaluation externe suscite l'intérêt du gouvernement. Un décret instaurant des épreuves non certificatives obligatoires dans le primaire et le secondaire paraît alors en 2006. En 2009, l'épreuve certificative permettant l'obtention du Certificat d'études de base (CEB) à la fin de l'enseignement primaire devient obligatoire. En 2013, c'est l'épreuve certificative de la fin du premier degré de l'enseignement secondaire, le CE1D, qui le devient. En 2015 sont prévues des épreuves certificatives pour la fin de l'enseignement secondaire en français et en histoire. En moins d'une décennie donc, dans une région précédemment caractérisée par l'absence d'épreuves standardisées communes à tous les établissements, les élèves sont, à peu de choses près, soumis à des épreuves standardisées tous les deux ans (certificatives ou non certificatives).

Dans la plupart des pays européens où sont imposées des épreuves standardisées, les résultats agrégés par établissements ne sont généralement pas diffusés publiquement. Seuls les Pays-Bas, la Suède, la Pologne et l'Islande procèdent à la diffusion de ces données. L'Italie constitue un exemple un peu particulier étant donné que les établissements sont libres d'organiser la diffusion des résultats. Dans le reste des pays ou des régions, comme en Communauté française de Belgique, la confidentialité des résultats est habituellement prévue par des textes

légaux. Les seuls résultats généralement diffusés dans ce genre de cas concernent les moyennes nationales et ne permettent en aucun cas l'identification des classes ou des établissements.

Dans l'ensemble du continent européen, les résultats aux épreuves standardisées n'impliquent pas de conséquences formelles aussi contraignantes que dans les cas anglais et nord-américain pour les enseignants et les établissements. Certes, les services d'inspection peuvent dans certains cas prendre en compte les résultats des épreuves lors de leurs interventions, mais ils ne disposent pas de l'autorité leur permettant de recommander le changement de l'équipe éducative ou la fermeture de l'établissement sur la base d'indicateurs de performance. Aussi, contrairement au cas anglais, le salaire des enseignants est défini sur base statutaire et ne varie en fonction d'aucun indicateur de performances qui soit. Finalement, si les enseignants peuvent être incités à l'amélioration des résultats, il n'y a généralement pas d'objectifs de performances à atteindre qui soient définis sur le continent européen.

Pratiques enseignantes et dispositifs *high-stakes*

La plus grande partie des études prises en compte dans cette revue de la littérature prennent place dans le cadre de dispositifs d'évaluation à enjeux élevés parmi lesquels le contexte nord-américain se trouve largement surreprésenté : sur 62 études, seules 4 prennent place dans le contexte anglais. Il apparaît également que la préoccupation pour les pratiques enseignantes apparaît aux États-Unis dans le courant des années 1980 et semble particulièrement marquée par un questionnement relatif à la validité des résultats de l'évaluation externe.

En effet, dans le courant des années 1980, comme nous l'avons vu plus haut, la plupart des États nord-américains s'orientent vers la mise en place de dispositifs d'évaluation reposant sur des tests conçus et commercialisés par des agences privées. Ceux-ci avaient alors pour caractéristique d'être étalonnés autour d'une moyenne nationale (c'est-à-dire interétatique) obtenue sur un échantillon d'élèves visant la représentativité, mais qui n'était réactualisée qu'au terme d'un certain nombre d'années. Ainsi par exemple, tout élève obtenant un score supérieur à 50 au *Comprehensive Test of Basic Skills* (CTBS) se situait au-dessus de la moyenne nationale. Il était donc attendu des scores de l'ensemble des élèves soumis à l'épreuve qu'ils soient répartis équitablement autour du seuil des 50 points.

En ce qui concerne les enjeux associés aux résultats pour les enseignants et les établissements, si l'association de sanctions financières apparaît dans certains cas, c'est surtout la publication des résultats qui se généralise comme pratique courante à ce moment. Il arrive alors régulièrement que certains districts et États mettent en avant des résultats supérieurs à l'étalon national afin de vanter l'efficacité de leurs politiques éducatives. Certains chercheurs utiliseront d'ailleurs ces mêmes résultats pour prôner l'efficacité des dispositifs d'*accountability* (Popham, Cruse, Rankin *et al.*, 1985 ; Popham, 1987).

C'est dans ce contexte que John J. Cannell, un physicien travaillant à l'époque pour le compte d'une association dont il est lui-même le fondateur, *Friends for Education*, s'intéresse à la proportion d'élèves situés au-dessus de l'étalon national dans chaque État. Il publie ainsi un rapport indiquant que tous les États, quelle que soit l'épreuve utilisée, revendiquent des scores moyens supérieurs à la norme nationale déterminée par les agences distribuant les épreuves. Ainsi, dans l'ensemble des États-Unis, 70 % des élèves et 90 % des districts scolaires recensés présentaient des scores moyens supérieurs à l'étalon national, ce qu'il considère comme une improbabilité statistique étant donné que les épreuves sont conçues de façon à ce que les scores des élèves se répartissent normalement autour du standard national. De plus, l'auteur remet en question l'utilisation des scores aux épreuves standardisées comme preuve de l'amélioration de la qualité de l'enseignement étant donné qu'ils sont en décalage avec d'autres indicateurs tels que le taux de diplômés, la fréquentation de l'enseignement supérieur

ou le revenu *per capita*. En fin de compte, comme pour mettre en avant le manque de crédibilité des scores rapportés par les États, le rapport sera largement médiatisé sous la dénomination de « rapport Wobegon »[4], en référence à la ville mythique aux habitants idylliques mise en scène dans un ouvrage de Garrisson Keillor (1985).

Le rapport Wobegon pose ainsi la question de l'efficacité des dispositifs d'*accountability* et constitue le point de départ d'une polémique relative à la validité des résultats aux épreuves standardisées qui semble encore faire débat à l'heure actuelle. Toutefois, si les insinuations de Cannell (c'est-à-dire que les résultats de l'évaluation externe ne représentent pas une amélioration de la qualité de l'enseignement) sont dans certains cas remises en cause, Koretz (1991) va constater que les scores des élèves à une épreuve standardisée aux enjeux moins importants, mais couvrant des matières similaires, sont nettement plus faibles que ceux des épreuves dont les résultats sont rapportés dans la presse. En mathématiques par exemple, la différence atteint jusqu'à 16 points percentile. Le gain observé lors d'une épreuve standardisée aux enjeux élevés mobiliserait donc de la part des élèves des compétences spécifiques qui ne sont pas nécessairement généralisables à d'autres contextes que celui de l'épreuve à laquelle ils ont été préparés. Ce constat, bien qu'ayant été nuancé de nombreuses fois[5], instaure néanmoins définitivement le principe méthodologique selon lequel les résultats aux épreuves standardisées ne peuvent pas à eux seuls être utilisés comme mesure de leur propre efficacité.

Par ailleurs, et cela s'avère d'une pertinence majeure pour cette synthèse, le rapport Wobegon aura également eu comme conséquence d'attirer l'attention sur les pratiques enseignantes comme facteur explicatif de l'inflation des résultats aux épreuves standardisées. À cet égard, deux catégories majeures de comportements semblent attirer l'attention : d'une part, les pratiques stratégiques « non pédagogiques » spécifiquement orientées vers l'amélioration des scores aux épreuves ; et d'autre part, la modification des contenus enseignés, parfois désignée par la notion du *teaching to the test*.

Stratégies non pédagogiques : l'amélioration des scores par tous les moyens ?

Une première interprétation de l'inflation improbable des résultats est avancée par Cannell lui-même dans un second rapport (1989) au titre évocateur : *How educators cheat on standardized achievement tests*. Dans ce document, l'auteur se base sur une série de lettres d'aveux, articles de journaux et études (par exemple Frary & Olson, 1985 ; Ligon, 1985 ; Perlman, 1985) pour tenter de mettre en évidence l'étendue des pratiques frauduleuses de la part des enseignants pour l'obtention de meilleurs résultats, avec la complicité des directions et autres acteurs dans certains cas. Diffusion du contenu des épreuves avant passation, lecture des réponses pendant la passation, mise à l'écart des élèves moins performants, attribution de temps supplémentaire au temps réglementaire ou modification des réponses des élèves pendant la correction apparaissent comme autant de pratiques susceptibles d'expliquer l'inflation des résultats.

Depuis, la préoccupation pour les stratégies mises en place par les enseignants pour gonfler les scores de leurs élèves n'a pas disparu. L'arrivée du NCLB aura d'ailleurs probablement contribué à stimuler la discussion sur le sujet. Ainsi, parmi les études que nous avons recensées, sur 17 ans, 10 se sont penchées sur la question des pratiques non pédagogiques des enseignants entre 1984 et 2001, année de la ratification du NCLB (Aiken, 1991 ; Carnoy, Loeb & Smith,

4 Parmi les centaines de journaux ayant relayé le rapport se trouvent notamment le *New York Times* et le *Wall Street Journal*. L'impact médiatique sera tel que le rapport est même mentionné dans une réunion spéciale organisée par le secrétaire de l'éducation de l'époque, William Bennett.

5 Pour approfondir cette question, se référer aux articles qui posent la question de l'efficacité des évaluations standardisées à partir des scores : Amrein & Berliner, 2002 ; Amrein-Beardsley & Berliner, 2003 ; Braun, 2004 ; Carnoy & Loeb, 2002 ; Haney, 2000 ; Hanushek & Raymond, 2005 ; Lee, 2008 ; Nichols & Berliner, 2007 ; Rosenshine, 2003.

2000 ; Frary & Olson, 1985 ; Gay, 1990 ; Haladyna, Nolan & Haas, 1991 ; Haney, 2000 ; Ligon, 1985 ; McGill-Franzen & Allington, 1993 ; Perlman, 1985 ; Shepard & Dougherty, 1991). Leur nombre augmente à 12 pour la période entre 2001 et 2010 (sur 9 ans donc) (Amrein & Berliner, 2002 ; Amrein-Beardsley, Berliner & Rideau, 2010 ; Amrein-Beardsley & Berliner, 2003 ; Booher-Jennings, 2005 ; Braun, 2004 ; Carnoy & Loeb, 2002 ; Figlio & Getzler, 2002 ; Hanushek & Raymond, 2005 ; Jacob & Levitt, 2004 ; Nichols, Glass & Berliner, 2006 ; Toenjes & Dworkin, 2002). Cette littérature fait en outre partie intégrante du débat relatif à l'efficacité des mesures d'*accountability*, au sein duquel il ne convient pas seulement de démontrer les gains générés par les épreuves *high-stakes* sur la base des scores, mais aussi de s'assurer que de tels gains ne sont pas dus à des « tricheries » de la part des enseignants.

Cependant, l'identification de pratiques frauduleuses flagrantes telles que la modification des réponses des élèves ou les interventions pendant la passation destinées à leur fournir une aide est malaisée. L'observation de terrain étant peu envisageable, certains auteurs se sont alors orientés vers l'analyse des irrégularités dans les scores pour tenter d'illustrer l'étendue de ce genre de phénomènes (Frary & Olson, 1985 ; Jacob & Levitt, 2004 ; Perlman, 1985). Ainsi, par exemple, sur la base des résultats à l'ITBS, Perlman (1985) sélectionne 23 établissements présentant des anomalies suspectes dans les scores et 17 établissements servant de groupe contrôle qu'elle soumet à un re-test. Elle constate alors que sur les 80 classes ayant participé au re-test, 19 présentent une diminution significative des résultats. Parmi ces dernières, 17 appartiennent à des établissements suspectés d'avoir eu recours à des pratiques frauduleuses.

Par ailleurs, Jacob et Levitt (2004) développent pour leur part un algorithme visant à estimer la proportion de fraudes commises par les enseignants à partir de deux indicateurs composites : les fluctuations de scores inattendues et les configurations inhabituelles des réponses d'élèves dans une classe. Leur base de données comprend les réponses à l'ITBS de tous les élèves de Chicago des niveaux 3 à 7 entre 1993 et 1999, ce qui totalise environ 700 000 observations pour chaque année scolaire. Ils estiment alors, en se concentrant sur les classes présentant des valeurs extrêmes pour les deux indicateurs, que des pratiques frauduleuses seraient repérables dans 4 à 5 % des classes du primaire. En outre, la mise en place d'un dispositif d'*accountability* à enjeux élevés en 1996 semble accentuer l'incidence du phénomène pour les classes les moins performantes : les classes dont la performance se situe à un écart-type en dessous de la moyenne augmenteraient ainsi leur probabilité de frauder de près de 50 %.

En ce qui concerne les études reposant sur des données qui ne soient pas des indicateurs officiels, nous avons pu identifier certaines études basées sur des comptes rendus de la part des enseignants (Aiken, 1991 ; Amrein-Beardsley, Berliner & Rideau, 2010 ; Gay, 2010) ou sur des études de cas (Booher-Jennings, 2005 ; Ligon, 1985). En guise d'exemple, reposant sur une taxonomie des fraudes inspirée du droit pénal, Amrein-Beardsley, Berliner et Rideau (2010) s'intéressent à une large gamme de comportements qu'ils qualifient de fraudes au premier (modification des réponses des élèves pendant la correction), deuxième (encourager un élève à revenir sur une réponse pendant la passation) et troisième degré (utilisation de versions anciennes de l'épreuve pour préparer les élèves) en fonction de leur gravité relative. Afin de mettre en évidence l'étendue de ces diverses pratiques frauduleuses, ils combineront une enquête par questionnaire en ligne, entretiens individuels et entretiens en *focus group* auxquels seront conviés une bonne partie des enseignants de l'Arizona. Les résultats indiqueront en fin de compte que sur un total de 3 085 réponses au questionnaire, 8 entretiens individuels et 8 enseignants ayant participé au *focus group*, environ la moitié des enseignants admettent commettre des fraudes ou connaissent des collègues qui fraudent, essentiellement aux deuxième et troisième degrés.

Par ailleurs, les fraudes flagrantes ne constituent pas les seules possibilités pour les enseignants de gonfler artificiellement les scores de leurs élèves aux épreuves externes. Ainsi, l'usage abusif du redoublement ou du placement en éducation spécialisée apparaît comme des « vices de procédure » au travers desquels les enseignants, sans pour autant porter atteinte

au règlement de l'évaluation standardisée, peuvent espérer augmenter la moyenne de leur classe ou établissement. Tout d'abord, le redoublement retarde la passation des élèves les moins performants, évitant ainsi que ceux-ci ne risquent de réduire la moyenne des scores. Il s'agit pourtant là d'une solution à court terme puisque les élèves sont censés arriver au niveau concerné par l'épreuve à un moment où à un autre, excepté dans le cas des épreuves présentées en dehors de la scolarité obligatoire (après le niveau 9) où le redoublement servirait d'incitant à l'abandon scolaire, éliminant donc définitivement ces élèves moins performants. En effet, les élèves en retard scolaire auraient quatre fois plus de probabilités d'abandonner l'enseignement secondaire avant son terme (Rumberger, 1995). Ensuite, une autre possibilité pour les enseignants de restreindre la quantité d'élèves en difficulté ayant à présenter l'épreuve externe est de les placer en « éducation spécialisée ». Ces élèves sont alors censés bénéficier de services spécialisés adaptés à leurs difficultés individuelles (*Individualized Education Plans* [IEP]), et surtout, ne sont pas pris en compte dans les statistiques déterminant la performance de la classe ou de l'établissement. Or, les conditions permettant de catégoriser un élève en éducation spécialisée étant relativement ambiguës (notamment en ce qui concerne le « retard de développement » ou les « troubles de l'apprentissage spécifique »), les enseignants disposent d'une certaine marge de manœuvre relative au nombre et au type d'élèves exclus des échantillons utilisés pour la reddition de comptes.

Dans un cas comme dans l'autre, le redoublement et les placements en éducation spécialisée semblent augmenter suite à l'introduction d'un dispositif *high-stakes*. Haney (2000), dans un rapport qui déclenchera une vive polémique, s'intéresse ainsi aux raisons qui permettent d'expliquer l'amélioration extraordinaire des scores au *Texas Assessment of Academic Skills* (TAAS), associé depuis 1994 à un dispositif d'*accountability* à forts enjeux. Il rapporte que les élèves placés en éducation spécialisée au niveau 10 (15-16 ans, niveau de passation de l'épreuve) passent de 3,9 à 6,3 % entre 1994 et 1998, contrastant de la sorte avec les tendances nationales où, pour la même période, ils passent de 8 à 6 % au niveau 4 (9-10 ans) et de 7 à 5 % au niveau 8 (13-14 ans)[6].

Concernant les redoublements, en l'absence apparente de recensements précis sur le phénomène, Haney (2000) se base sur le rapport entre le nombre d'élèves inscrits au niveau 9 (14-15 ans) et le nombre d'élèves inscrits l'année précédente au niveau 8. Il part ainsi du principe que les élèves peu performants sont retenus au niveau 9 avant la passation de l'épreuve au niveau 10 et qu'il y aurait donc un nombre plus important d'élèves au niveau 9 en raison des redoublements. Et effectivement, le constat est que la proportion d'élèves doubleurs aurait augmenté de 10 % dans le courant des années 1990, et concernerait essentiellement les élèves des minorités afro-américaine et hispanique.

Cependant, si l'augmentation de la proportion d'élèves placés en éducation spécialisée et l'utilisation accrue du redoublement semblent persister, dans quelle mesure contribuent-elles à l'amélioration des scores aux épreuves standardisées ? En ce qui concerne l'usage abusif du redoublement, il semblerait qu'il ne puisse pas rendre compte de l'inflation des scores en contexte d'*high-stakes accountability* en raison du fait que la relation entre la sévérité du dispositif et les variations dans l'usage du redoublement ne paraît pas significative (Carnoy & Loeb, 2002). En ce qui concerne les placements en éducation spécialisée par contre, la question reste controversée. La décennie 2000 a ainsi été le théâtre d'un échange d'articles, mobilisant des bases de données comprenant scores au *National Assessment of Educational Progress* (NAEP) et pourcentages d'« exclusions », opposant les auteurs déclarant que l'augmentation de la proportion d'élèves placés en éducation spécialisée est suffisamment importante pour expliquer l'amélioration des scores constatée dans les États ayant instauré des dispositifs *high-stakes* (Amrein & Berliner, 2002 ; Amrein-Beardsley & Berliner, 2003 ; Nichols, Glass & Berliner,

6 Les données font référence à des niveaux différents car il s'agit des épreuves *National Assessment of Educational Progress* (NAEP) pour le niveau national et du TAAS pour le Texas.

2006) et ceux considérant à l'inverse qu'elle n'a pas d'impact significatif sur les scores (Braun, 2004 ; Carnoy & Loeb, 2002 ; Hanushek & Raymond, 2005). Ainsi, par exemple, sur la base d'une analyse des résultats au NAEP entre 1992 et 2002, Hanushek et Raymond (2005) constatent que les États ayant instauré des dispositifs d'*accountability* associés à des systèmes de sanctions ou récompenses présentent une évolution significative de leur performance qui n'apparaît pas pour d'autres États, et ce en contrôlant les exemptions de passation ; tandis qu'à partir d'une analyse des résultats au NAEP entre 1996 et 2000, Nichols, Glass et Berliner (2006) montrent que la corrélation positive entre la proportion d'élèves ayant réussi l'épreuve de mathématiques au niveau 8 et un indice de pression des dispositifs (*Accountability Pressure Rating* [APR]) disparaît lorsque les exemptions de passation sont prises en compte.

Finalement, il semble également qu'une autre stratégie utilisée par les enseignants dans le but précis de gonfler les scores de leurs élèves aux épreuves standardisées ait trait à la gestion stratégique de l'hétérogénéité de la classe. Contrairement à la plupart des études sur le sujet en effet, Booher-Jennings (2005) tente d'identifier la façon dont les enseignants se comportent dans le contexte des dispositifs d'*acountability* à travers une étude de cas combinant observations de terrain, observations participantes, entretiens individuels et analyse documentaire dans une école primaire défavorisée du Texas. Elle identifiera ainsi certaines pratiques destinées à gonfler les scores par lesquelles les enseignants « se jouent du système » (*game the system*), parmi lesquelles figure notamment l'utilisation abusive des placements en éducation spécialisée. Cependant, l'auteure insiste surtout sur les stratégies de catégorisation des élèves en fonction de leur niveau ayant déjà été identifiées dans le contexte anglais (Gillborn & Youdell, 2000), au travers desquelles les enseignants orientent davantage de ressources vers les élèves en difficulté les plus susceptibles de réussir l'épreuve externe : les « *bubble kids* ». Dans ce sens, les enseignants se basaient sur les résultats issus des simulations d'épreuves externes (en l'occurrence du *Texas Assessment of Knowldege and Skills* [TAKS], présenté aux élèves des niveaux 3 à 8) pour catégoriser les élèves en cas « sûrs », cas « nécessitant un traitement » et cas « désespérés ». La plupart des ressources étaient alors orientées vers ces élèves intermédiaires (attention accrue en classe, tutorats pendant le week-end, écoles d'été, etc.), et ce dans le but explicite d'obtenir de meilleurs scores lors de l'épreuve finale diffusée dans les médias. En conséquence, les élèves performants et les « cas désespérés » étaient pour leur part nettement moins soutenus tout au long de l'année. Certes, de telles pratiques de catégorisation auraient pu préexister à la mise en place d'un dispositif d'évaluation à forts enjeux. Difficile pourtant d'apporter une réponse à cette préoccupation en l'absence de données contrastées concernant les pratiques enseignantes précédant l'instauration du dispositif *high-stakes*. Néanmoins, les évaluations à forts enjeux semblent avoir contribué à rationaliser et à institutionnaliser ce phénomène de catégorisation cantonné au préalable à l'appréciation personnelle de l'enseignant (Booher-Jennings, 2005).

L'influence sur les pratiques pédagogiques : contenus et méthodes

En partant de la polémique concernant la validité des résultats aux épreuves *high-stakes*, bon nombre d'études ont également orienté leur attention sur la façon dont l'enseignement se trouvait affecté par les épreuves externes. À cet égard, deux constats principaux sont récurrents : la réduction du temps attribué aux matières non testées (Booher-Jennings, 2005 ; Boyle & Bragg, 2005 ; Diamond, 2007, 2012 ; Smith, 1991b ; Taylor, Shepard, Kinner *et al.*, 2003) et l'alignement des contenus enseignés sur le contenu de l'épreuve dans les matières testées (Barksdale-Ladd & Thomas, 2000 ; Boyle & Bragg, 2005 ; Bracey, 1987 ; Collins, Reiss & Stobart, 2010 ; Diamond, 2007 ; Firestone, Mayrowetz & Fairman, 1998 ; Gerwin & Visone, 2006 ; Herman, Abedi & Golan, 1994 ; Landman, 2000 ; Shepard & Dougherty, 1991 ; Shepard, 1990 ; Smith, 1991a, 1991b ; Valli & Buese, 2007). En ce qui concerne la réduction du temps attribué aux matières

non testées, à partir d'observations de terrain et d'une cinquantaine d'entretiens avec les enseignants dans une série d'établissements de Chicago, Diamond constate par exemple que les enseignants des niveaux 2 et 5 délaissent l'enseignement des sciences et sciences sociales au profit des matières testées, à savoir les mathématiques et l'anglais. Un des enseignants interviewés énonce ainsi assez clairement que « si vous devez laisser tomber tout le reste, c'est bon tant que les maths et la lecture sont couverts. Celles-là sont les deux matières où [les élèves] sont testés » (Diamond, 2007, p. 295). De telles pratiques apparaissent également dans l'étude de Booher-Jennings (2005) où les enseignants de gymnastique, de musique et de littérature (*library teachers*) substituent l'enseignement de leurs matières respectives à la préparation aux épreuves des *bubble kids*.

D'autre part, en ce qui concerne l'alignement de l'enseignement sur le contenu de l'épreuve, l'étude de Diamond permet aussi de constater que les enseignants de langue et de mathématiques déclaraient accorder une attention prononcée aux sujets ayant été particulièrement mal réussis lors des épreuves standardisées : « J'inonde mes élèves avec du vocabulaire parce que leur vocabulaire est tellement faible… Et s'ils veulent réussir l'*Iowa Test*, ils doivent se familiariser avec ces termes » ; ou encore : « Nous étions en train de faire… des soustractions de fractions… parce que je sais qu'ils en auront plein à l'épreuve… et les fractions posent toujours problème avec les enfants » (Diamond, 2007, p. 295).

Notons aussi que les dispositifs d'*accountability* paraissent avoir un impact récurrent sur le matériel pédagogique utilisé par les enseignants. Ces derniers auraient ainsi tendance à mobiliser du matériel pédagogique similaire ou identique au contenu de l'épreuve, notamment lors des exercices en classe, de la conception des épreuves internes ou des simulations d'épreuves (Amrein-Beardsley, Berliner & Rideau, 2010 ; Anagnostopoulos, 2003 ; Bol, 2004 ; Diamond, 2007 ; Firestone, Mayrowetz & Fairman, 1998 ; Grant, Gradwell, Lauricella *et al.*, 2002 ; Kirkup, Sizmur, Sturman *et al.*, 2005 ; Shepard & Dougherty, 1991 ; Shepard, 1990 ; Smith, 1991b, 1991a ; Taylor, Shepard, Kinner *et al.*, 2003 ; Valli & Buese, 2007). Ainsi, par exemple, se fondant sur des études de cas dans deux écoles primaires de l'Arizona, Smith (1991a) constate qu'environ 40 % des enseignants utilisent des manuels scolaires dont les contenus se rapprocheraient tellement de l'ITBS qu'ils équivaudraient à présenter une version parallèle du test. Même à contrecœur, une des enseignantes mentionnées pour illustrer leur propos aurait passé 80 % de son temps d'enseignement à travailler sur des exercices issus de tels manuels au cours des trois semaines précédant la présentation de l'ITBS. La situation ne semble pas avoir nettement changé près de 15 ans plus tard étant donné que Valli et Buese (2007) constatent dans leur étude longitudinale que l'achat de manuels scolaires dont le format et la présentation sont similaires aux épreuves standardisées est encore une pratique courante dans les établissements du Maryland.

Toujours en ce qui concerne l'influence des épreuves standardisées sur les contenus enseignés, une série d'études pointe aussi le séquençage des contenus et une accélération du rythme d'enseignement dans le cadre des dispositifs d'*accountability*. Ainsi, dans sa « métasynthèse qualitative » reposant sur 49 études nord-américaines, Au (2007) souligne que bon nombre d'entre elles (49 %) mettent en avant une fragmentation accrue des contenus enseignés dans le cadre des dispositifs d'évaluation externe, dans le sens où les savoirs seraient davantage transmis par fragments courts, individualisés et isolés. Concernant le rythme d'enseignement par ailleurs, les entretiens réalisés par Valli et Buese (2007) rapportent par exemple que les enseignants sont davantage contraints d'accélérer leurs cours afin de présenter l'ensemble des unités d'apprentissage susceptibles d'être couvertes par l'épreuve à temps. Un constat similaire est établi par Diamond (2007), qui précise que certains enseignants déclaraient avancer dans la matière en dépit du manque de maîtrise de la part des élèves.

Mais quels sont les facteurs susceptibles de favoriser l'apparition de tels effets sur les contenus enseignés ? Des variables telles que le niveau socio-économique du public accueilli dans l'établissement, l'intensité des enjeux associés aux résultats de l'épreuve externe ou le

niveau ciblé par ces épreuves ont été au centre de la préoccupation de certaines études visant à répondre à cette question. À partir d'un questionnaire complété par 341 enseignants, Herman, Abedi et Golan (1994) constatent par exemple que les enseignants confrontés à des publics défavorisés sont davantage susceptibles de focaliser leur enseignement sur le contenu de l'épreuve en raison d'une pression accrue à l'amélioration des scores causée par les difficultés de leurs élèves. Bol (2004) pour sa part montre à partir d'une étude par questionnaire transmise à 168 enseignants que ceux du secondaire supérieur confrontés aux dispositifs d'*accountability* sont davantage enclins à élaborer des évaluations internes imitant les épreuves standardisées que leurs collègues du primaire.

Concernant l'impact des enjeux associés aux résultats de l'épreuve externe, Firestone, Mayrowetz et Fairman (1998) ont quant à eux mené des études de cas dans deux États où les enjeux associés aux résultats de l'épreuve externe différaient : l'État du Maine, où les résultats sont publiés dans les médias ; et l'État du Maryland où, en plus d'être publiés, de mauvais résultats sont susceptibles d'aboutir à une reconstitution de l'équipe éducative. Les entretiens réalisés auprès des enseignants de mathématiques du premier cycle du secondaire (*middle school*) mènent alors les auteurs à la conclusion que les enseignants confrontés à des enjeux plus élevés ont davantage tendance à orienter la planification de leurs cours en fonction du contenu de l'épreuve. Mais si l'alignement de l'enseignement sur le contenu de l'épreuve varie en fonction des enjeux pour les enseignants, le rapport élaboré par Clarke, Shore, Rhoades *et alii* (2003) laisse supposer que la variation des enjeux de l'épreuve pour les élèves y contribue également. Sur la base d'entretiens réalisés auprès de 360 enseignants dans les États du Kansas, du Michigan et du Massachusetts, où les enjeux des épreuves varient en importance pour les élèves mais restent constants pour les enseignants, les auteurs constatent malgré tout une focalisation plus prononcée sur les matières présentées dans l'épreuve dépendante de l'importance de l'épreuve dans le parcours scolaire de l'élève.

Les dispositifs d'évaluation externe semblent donc exercer une influence significative sur les contenus enseignés. Qu'en est-il toutefois des méthodes pédagogiques, censées être portées par une dynamique d'innovation liée à la reddition de comptes ? À cet égard, si l'objectif des dispositifs d'*accountability* était d'inciter à l'innovation pédagogique, les données soulevées par les recherches sur le sujet ne permettent pas de le confirmer (Barksdale-Ladd & Thomas, 2000 ; Diamond, 2007 ; Firestone, Mayrowetz & Fairman, 1998). Les entretiens réalisés par Diamond (2007) mettent ainsi en évidence que l'épreuve externe ne constitue pas un élément déterminant dans les prises de décisions pédagogiques (c'est-à-dire quant aux méthodes) effectuées par les enseignants, qui continuent pour cela de se fonder sur des sources telles que les réflexions personnelles, les conseils des collègues, les manuels scolaires ou sur l'appréciation des compétences perçues des élèves. Plus encore, les entretiens réalisés par Barksdale-Ladd et Thomas (2000) indiquent qu'au lieu d'inciter au développement de nouvelles méthodes pédagogiques, les dispositifs d'évaluation externe inciteraient à la disparition de toute une série de pratiques pédagogiques : lectures collectives, expérimentation scientifique et activités créatives en général. En fin de compte, au-delà de ces exemples spécifiques, Au (2007) rapporte que 65 % des études de son échantillon signalent un développement accru de la pédagogie « centrée sur l'enseignant » sous les dispositifs d'évaluation externe, voulant dire par là que l'enseignement se baserait davantage sur la transmission des savoirs de l'enseignant vers l'élève plutôt que sur des interactions entre et avec les élèves.

Pratiques enseignantes et dispositifs à faibles enjeux

Nous avons précisé dans l'introduction de cette note que l'engouement politique pour les dispositifs d'évaluation externe n'a pas épargné le continent européen. Cela fait maintenant

quelques décennies qu'y sont développés des dispositifs d'évaluation externe qui, contrairement aux dispositifs développés dans les pays anglophones, se caractérisent par l'absence de systèmes de sanctions/récompenses associées aux résultats. Une telle tendance politique n'aura pas non plus manqué de générer un certain intérêt scientifique pour les dispositifs d'évaluation externe. Rien qu'en prenant en compte la littérature francophone sur le sujet, les publications ne sont pas particulièrement rares (voir par exemple Cattonar, Dumay & Mangez, 2010 ; Dupriez & Malet, 2013 ; Lopez & Crahay, 2009 ; Maroy & Voisin, 2013 ; Mons & Pons, 2006). Toutefois, les études empiriques centrées sur les pratiques enseignantes dans les contextes de dispositifs d'évaluation à faibles enjeux restent peu nombreuses, surtout en comparaison avec la quantité de recherches nord-américaines (cf. annexes). Tous les États membres ne sont pas non plus couverts par ces études. Les articles recensés dans notre revue de la littérature se limitent à l'étude des effets des dispositifs d'évaluation externe dans les contextes allemand, hollandais, belge (néerlandophone et francophone), français, suisse et luxembourgeois.

Contrairement au cas nord-américain, la question de l'amélioration des scores ne semble pas être une préoccupation centrale des études menées dans le cadre des dispositifs à faibles enjeux du continent européen. Une bonne partie d'entre elles assument un objectif prescriptif au travers duquel elles tentent de fournir des indications permettant d'améliorer les dispositifs d'évaluation existants. En particulier, les études non francophones (Maier, 2009, 2010 ; Schildkamp & Kuiper, 2010 ; Vanhoof, Verhaeghe & Van Petegem, 2012), s'inscrivant assez nettement dans le courant de la *School Effectiveness* et de la *School Improvement,* tentent d'identifier les « bonnes pratiques » développées par les enseignants dans le cadre des dispositifs d'évaluation externe et les facteurs sur lesquels il convient d'agir pour les stimuler. Shildkamp et Kuiper, par exemple, cherchent dans leur étude à déterminer les facteurs incitant à l'utilisation des données quantitatives et fondent leur recherche sur un présupposé favorable à l'utilisation des résultats de l'évaluation par les enseignants :

> Par exemple, un enseignant n'étant pas satisfait des résultats de l'évaluation peut décider d'analyser les résultats d'une façon plus critique. Sur la base de ces données, il peut arriver à la conclusion qu'il doit focaliser son enseignement sur certains sujets et qu'il devrait faire des changements dans sa manière d'enseigner. En conséquence, il peut commencer à utiliser différentes stratégies d'enseignement (amélioration de l'enseignant), et peut centrer l'enseignement sur certains sujets (amélioration du curriculum). Les données issues des résultats à une épreuve suivante peuvent lui dire si les changements opérés étaient satisfaisants dans le sens où ils ont amené à élever le niveau de réussite des élèves (amélioration de l'école) (Shildkamp & Kuiper, 2010, p. 483).

D'autre part, les études francophones, pour la plupart descriptives, sont essentiellement constituées de rapports institutionnels destinés à identifier les effets des dispositifs d'évaluation mis en place, voire simplement la perception que les enseignants ont des fonctions et de la qualité de l'évaluation externe (Dierendonck & Fagnant, 2010 ; Dierendonck, 2008 ; Longchamp & Gilliéron, 2009 ; Normand & Derouet, 2003).

Perception et acceptation des épreuves standardisées

Quels que soient toutefois les objectifs d'étude, la méthodologie ou le contexte étudié, un constat majeur émerge dans le cadre des dispositifs à faibles enjeux : celui de la faible utilisation des résultats par les enseignants pour prendre des décisions quant aux méthodes pédagogiques. À titre indicatif, dans leur méta-analyse de 52 articles publiés entre 1991 et 2010 et relatifs à la compréhension et à l'utilisation des *feedbacks* à l'école, Hellrung et Hartig soulèvent qu'« entre 50 et 98 % des enseignants d'Allemagne, de Belgique et des Pays-Bas refusent de changer ou d'innover en matière d'enseignement, de méthodes d'enseignement ou de curriculum en réaction au *feedback* fourni par les épreuves à faibles enjeux en langues et

mathématiques »[7] (2013, p. 7-8). Dans le même ordre d'idées, Verdière indique sur la base d'entretiens réalisés en région lilloise que pour « la quasi-totalité des enseignants […] rencontrés, [les] indicateurs statistiques permettant l'évaluation des élèves ou des établissements sont connus, mais ont peu de liens concrets avec leurs pratiques professionnelles » (2013, p. 83).

Afin de comprendre les raisons d'un tel constat, un certain nombre d'études se sont alors penchées sur la façon dont les enseignants percevaient et interprétaient les données issues de l'évaluation externe. À cet égard, s'ils ne déclarent pas en faire usage dans leur pratique quotidienne, les enseignants ne semblent pourtant pas fondamentalement opposés au principe de l'évaluation externe (Lafontaine, Soussi & Nidegger, 2009 ; Maier, 2009, 2010 ; Soussi, Nidegger, Ducrey *et al.*, 2006 ; Vanhoof, Verhaeghe & Van Petegem, 2012). Lorsque des questions très générales leur sont posées, les enseignants semblent en effet indiquer une attitude plutôt favorable au fait qu'une épreuve standardisée soit imposée à leurs élèves, ce qui ne semblait pas évident en raison de la perte d'autonomie de fait que supposent généralement les dispositifs d'évaluation externe. Ainsi, par exemple, 87 % des 120 enseignants luxembourgeois ayant répondu au questionnaire développé par Dierendonck et Fagnant (2010) considèrent qu'il est légitime pour les autorités publiques d'organiser des épreuves externes qui s'imposent à leurs élèves. Un constat similaire est établi dans une étude par entretiens avec 40 enseignants confrontés aux épreuves standardisées de la Communauté française de Belgique (Rosenwajn, à paraître), où l'attitude favorable à l'égard des épreuves paraît dans de nombreux cas justifiée par une meilleure égalité de traitement envers les élèves. C'est ainsi que Mons précise dans son article centré sur les épreuves standardisées que « 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 » (2009, p. 27).

Qu'est-ce qui permet alors d'expliquer ce contraste entre une acceptation plutôt favorable de l'évaluation externe et l'usage si faible qui est fait des résultats par les enseignants ? La question est d'autant plus pertinente que la plupart des dispositifs se donnent parmi leurs objectifs celui de la régulation des pratiques enseignantes. Certaines études se sont donc attachées à identifier les tensions susceptibles d'être produites par l'évaluation externe. Essentiellement menées dans le contexte francophone (Dierendonck, 2008 ; Dutercq & Lanéelle, 2013 ; Normand & Derouet, 2003), ces études soulignent qu'en dépit d'une attitude favorable au principe de l'évaluation standardisée, certains aspects de celle-ci posent problème : incohérence entre l'enseignement et le contenu de l'épreuve, calendrier de passation inapproprié, difficultés des questions, systèmes de notation, etc. Les problèmes soulevés par ces dimensions pratiques de l'évaluation externe peuvent alors limiter la volonté d'usage des résultats. Néanmoins, certaines données nous incitent à penser que l'opposition déclarée aux aspects concrets du dispositif d'évaluation externe ne peut pas constituer une explication suffisante au faible usage des résultats par les enseignants. En effet, sans nécessairement exprimer une opposition explicite vis-à-vis de l'évaluation externe, certains enseignants ne perçoivent tout simplement pas la pertinence des résultats comme source informative par rapport à leur activité professionnelle. Ainsi, les résultats au questionnaire de Dierendonck et Fagnant indiquent que « seuls 38 % des répondants déclarent que les graphiques illustrant la distribution des résultats obtenus par leurs élèves donnent une information intéressante et utile pour leur pratique quotidienne », tandis qu'ils « ne sont que 25 % à dire que les *feedbacks* leur ont permis de prendre

7 Cette affirmation se base sur un ensemble de 6 études que nous n'avons pas toutes reprises dans le cadre de cette revue de la littérature étant donné qu'elles n'étaient pas disponibles en anglais.

conscience que certaines facettes de leur enseignement devaient être modifiées et 19 % à dire que, suite aux *feedbacks*, ils voient comment améliorer certaines choses dans leur enseignement » (2010, p. 5). Dans le cadre du canton de Vaud, en Suisse, Longchamp et Gillièron se basent quant à eux sur des entretiens réalisés auprès des enseignants confrontés aux Épreuves cantonales de référence (ECR) pour conclure que les résultats confortent les observations quotidiennes des enseignants et ne donnent « que peu d'informations supplémentaires » (2009, p. 15). En France, de même, Dutercq et Lanéelle (2013), à partir d'entretiens réalisés auprès d'enseignants de CE1 et de CM2, indiquent que, si les résultats permettent de constater les difficultés des élèves, ils ne leur fournissent pas nécessairement les informations nécessaires pour y remédier. Les entretiens réalisés par Rozenwajn (à paraître) dans le cadre de la Communauté française de Belgique rapportent également des résultats proches, en ce sens qu'au moment de faire sens des résultats obtenus par leurs élèves, les enseignants mobilisaient essentiellement des attributions externes telles que l'origine sociale des élèves ou la responsabilité partagée avec leurs collègues les ayant précédés dans le parcours scolaire des élèves.

Influences sur les pratiques pédagogiques et non pédagogiques

Nous venons de voir que les enseignants semblent accorder peu de crédit à l'évaluation externe comme source de régulation de leurs pratiques. Est-ce à dire pourtant que les évaluations externes à faibles enjeux n'influencent pas ou très peu ces dernières ? Tout d'abord, soulignons que les pratiques stratégiques non pédagogiques destinées à l'amélioration des scores, largement étudiées dans les contextes anglophones, n'ont pas constitué une préoccupation centrale dans le contexte de l'Europe continentale. Seule l'étude de André pose le constat que certains enseignants augmentent le temps à disposition des élèves lors de la passation des épreuves, répondent aux questions des élèves ou font preuve d'une « certaine flexibilité dans les corrections » (2013, p. 53). Sa préoccupation ne se centre toutefois pas sur l'étendue du phénomène, comme dans le cas de la plupart des études nord-américaines relatives aux pratiques frauduleuses, mais sur les raisons subjectives qui amènent les enseignants à recourir à de telles pratiques. À partir d'entretiens avec 16 enseignants vaudois confrontés aux ECR, il en conclut que ces « ruses » mises en œuvre par les enseignants visent essentiellement à rétablir un sentiment de justice, soit par rapport aux efforts consentis par l'enseignant, soit pour contrecarrer des circonstances adverses subies par les élèves. Par ailleurs, même si l'objet de leur recherche ne concerne pas spécifiquement les pratiques frauduleuses, Dutercq et Lanéelle mentionnent aussi des rapports indiquant que certains enseignants « ne respectent pas forcément la consigne du strict respect de la formulation de la question », tandis que « d'autres préfèrent marquer "absent" à chaque fois qu'une compétence n'a pas été travaillée » (2013, p. 52-53).

Outre les pratiques non pédagogiques cependant, plusieurs observations laissent supposer que les dispositifs d'évaluation à faibles enjeux exercent aussi une influence sur les contenus enseignés. Les entretiens réalisés par Longchamp et Gilliéron permettent ainsi de constater que les enseignants confrontés aux épreuves externes ont tendance à orienter la planification de l'enseignement en fonction du contenu des épreuves : « Pour la plupart des enseignants interrogés (17), les compétences annoncées à l'avance ont une influence sur leur enseignement, particulièrement sur le choix des thèmes qui vont être travaillés » (2009, p. 16).

De plus, ces épreuves sont aussi reprises par certains enseignants, soit à l'identique soit comme source d'inspiration, pour la conception de leurs exercices de cours : « Certains maîtres (9) utilisent les épreuves cantonales comme référence ou comme réservoir d'exercices qu'ils modifient ou reprennent tels quels » (Longchamp & Gilliéron, 2009, p. 20).

Dans le contexte français, les épreuves standardisées semblent aussi affecter la planification des cours. Toujours selon Dutercq et Lanéelle par exemple, « de nombreux enseignants reconnaissent […] avoir revu l'organisation temporelle des séquences de l'année pour aborder

l'ensemble des notions avant les évaluations » (2013, p. 52). Mais si la planification des cours en fonction de l'épreuve externe affecte les matières testées, certains témoignages indiquent que le temps attribué aux matières non testées diminue également dans le contexte des évaluations à faibles enjeux. C'est ainsi que sur la base de ses entretiens, Baluteau souligne que les enseignants du primaire planifient stratégiquement leurs horaires de cours afin de favoriser les matières évaluées dans les épreuves nationales. Selon les termes de l'auteur :

> Face aux évaluations nationales, des enseignants sont amenés, avec certains publics, à finalement sortir du cadrage officiel et à construire un curriculum approprié. Ce qui revient à réordonner la culture scolaire aux dépens des disciplines dites « secondaires » dont le statut se trouve dévalué dans le dispositif d'évaluation nationale. Ainsi, l'évaluation standardisée influe sur le curriculum parfois plus que la prescription nationale (curriculum formel sous la forme de grilles horaires) (2013, p. 67).

En ce qui concerne l'influence des dispositifs à faibles enjeux sur les modalités de transmission des savoirs, les données limitées dans la littérature, à prendre avec prudence donc, semblent à nouveau indiquer que les méthodes pédagogiques utilisées par les enseignants ne sont pas significativement affectées par les épreuves standardisées. Ainsi, toujours sur la base des entretiens réalisés par Longchamp et Gillièron, ni les résultats ni les « cahiers d'accompagnement » qui y sont associés ne semblent influencer les pratiques pédagogiques des enseignants :

> La comparaison des résultats n'entraîne pas de changement dans l'enseignement de plus de la moitié des maîtres (12) qui préfèrent organiser leur travail en fonction des évaluations internes à la classe (2009, p. 23)

> Selon les enseignants consultés, le cahier pédagogique n'influence pas les pratiques d'enseignement. Un peu plus de la moitié (11) relèvent bel et bien des éléments qui ressortent des consignes, mais ils attribuent l'attention qu'ils portent à ces aspects à leur formation de base, à leur formation continue, en particulier au début d'EVM [École vaudoise en mutation, nom attribué à la réforme scolaire de 1998], ou aux échanges entre collègues lors de la préparation d'épreuves communes. Pour les enseignants le cahier d'accompagnement est en phase avec leurs pratiques et les conforte dans leur manière d'aborder les apprentissages (2009, p. 21).

Dans une certaine mesure donc, la similarité des observations en contexte d'évaluation externe à forts et faibles enjeux laisserait supposer que l'influence des épreuves standardisées sur les contenus enseignés est relativement indépendante des enjeux qui y sont associés. Qu'ils soient confrontés à des dispositifs à faibles ou à forts enjeux, les enseignants semblent tous dans une certaine mesure aligner leur enseignement sur le contenu de l'épreuve, réduire le temps attribué aux matières non testées et faire usage d'un matériel pédagogique similaire à celui de l'épreuve. Les études menées dans le contexte des dispositifs à faibles enjeux restent toutefois trop peu nombreuses pour pouvoir dépasser le stade de l'hypothèse et présentent certaines limites non négligeables. Tout d'abord en effet, elles se cantonnent à des comptes rendus de la part des enseignants (par questionnaires ou entretiens) à partir d'échantillons particulièrement restreints susceptibles de porter atteinte à la représentativité des observations. Ensuite, l'analyse de l'effet des épreuves à partir d'indicateurs quantitatifs reste peu explorée dans les études européennes. Nous ne savons donc pas dans quelle mesure les pratiques des enseignants sont susceptibles d'avoir un impact sur les scores des élèves. Nous ne savons pas non plus dans quelle mesure les épreuves externes à faibles enjeux renforcent le recours à des procédures telles que le redoublement. De même, peu d'observations de terrain semblent disponibles dans le cadre des dispositifs d'évaluation externe à faibles enjeux. Nous manquons donc d'informations plus précises quant à l'adéquation entre les comptes rendus de la part des enseignants et les pratiques effectives ainsi que sur d'autres effets potentiels non rapportés.

Discussion

L'objectif de cette revue de la littérature était de recenser les études s'étant penchées sur la question des pratiques enseignantes dans le contexte des dispositifs d'évaluation externe. Pour ce faire, nous avons privilégié une catégorisation des études en fonction de la distinction communément utilisée dans la littérature entre dispositifs à enjeux élevés (caractéristiques des régions anglophones) et faibles (davantage caractéristiques du continent européen). Rapidement il apparaît que plus d'études ont été menées pour comprendre les effets des dispositifs d'évaluation à enjeux élevés que pour saisir les effets de tels dispositifs à faibles enjeux (62 *vs* 15 études), mais aussi qu'elles ont été initiées bien plut tôt (1984 *vs* 1998).

Dans les deux cas cependant, l'intérêt pour l'influence de l'évaluation externe sur les contenus enseignés constitue la préoccupation principale. Certaines études nord-américaines indiquent à cet égard une importante tendance au séquençage et à l'accélération du rythme d'enseignement dans le cadre des dispositifs d'évaluation externe (Diamond, 2007 ; Valli & Buese, 2007). Cependant, les constats les plus récurrents qui semblent traverser l'ensemble de la littérature, indépendamment du type de dispositif étudié, concernent l'alignement de l'enseignement sur le contenu de l'épreuve, la réduction du temps attribué aux matières non testées et l'utilisation de matériel pédagogique similaire à l'épreuve. Ces phénomènes sont régulièrement désignés par la notion du *teaching to the test* (ou bachotage dans sa version française) et leur identification n'est d'ailleurs pas limitée aux études scientifiques. Plusieurs rapports de services d'inspection y font en effet également référence. Ainsi, un des rapports de l'OFSTED pour la House of Commons précise que lorsque des enjeux élevés sont associés aux résultats, une « distorsion se fait sous la forme du *teaching to the test*, réduisant le curriculum enseigné aux matières susceptibles d'être concernées par l'épreuve externe et d'une déviation inappropriée des ressources vers les élèves les plus proches de réussir l'épreuve au seuil exigé, au détriment tant des élèves hautement performants que de ceux n'ayant que peu ou pas du tout de chances d'atteindre le seuil de réussite, même avec du soutien » (House of Commons, 2008, p. 30). D'autre part, dans le contexte d'un dispositif d'évaluation à faibles enjeux comme la France, le rapport de l'IGEN-IGAENR précise qu'une « autre voie pour agir sur les contenus enseignés consiste à agir sur les évaluations 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 » (2005, p. 38).

Dans certains cas, de tels phénomènes d'alignement de l'enseignement sur le contenu de l'épreuve se voient attribuer une connotation négative les associant à des pratiques frauduleuses. Cannell (1989) lui-même utilise la notion du *teaching to the test* lorsqu'il dénonce les « tricheries » commises par les enseignants préparant intensivement les élèves aux épreuves standardisées. Cependant, l'alignement de l'enseignement sur le contenu de l'épreuve constitue aussi, dans une certaine mesure, un objectif de tout dispositif d'évaluation externe et n'est pas nécessairement perçu d'un mauvais œil par tous les acteurs. Comme le précisait déjà Shepard en 1991 en effet :

> La notion de *teaching to the test* est interpellante, mais elle sous-tend en fait de trop nombreuses significations pour être d'une utilité quelconque. En dépit qu'une connotation négative y soit associée pour la plupart, de nombreux enseignants y font référence pour désigner le fait d'enseigner les domaines couverts par l'épreuve (1991, p. 2).

Toutefois, au-delà des connotations morales associées à l'alignement de l'enseignement sur le contenu des épreuves, la question qui mérite d'être posée est celle de savoir si un tel phénomène permet de fournir une meilleure formation aux élèves. Or, les études indiquant le transfert de compétences limité d'une épreuve à une autre (Amrein & Berliner, 2002 ; Koretz, 1991) incitent à penser que les élèves confrontés à des dispositifs d'évaluation externe ne sont

pas nécessairement capables de généraliser les compétences acquises à d'autres contextes que celui de l'épreuve externe à laquelle ils ont été préparés.

Par ailleurs, si l'influence des épreuves standardisées sur les contenus enseignés semble assez bien établie, en particulier dans les contextes de reddition de comptes à enjeux élevés, il semble que ce soit nettement moins le cas en ce qui concerne le lien entre évaluation externe et innovation pédagogique. Un second résultat majeur de notre revue de la littérature, particulièrement forgé dans le cadre des études portant sur une évaluation externe à enjeux faibles, est en effet que les enseignants ne considèrent pas les résultats de l'évaluation externe comme une source d'information pertinente pour réguler leurs pratiques et les faire évoluer. Aux moments où se posent des questions d'ordre pédagogique, les enseignants continuent de se baser sur leur expérience personnelle, sur les collègues ou sur d'autres sources informatives telles que les manuels scolaires. Plus encore, s'il est attendu d'un dispositif d'évaluation externe qu'il stimule l'innovation pédagogique, les études menées dans le contexte nord-américain indiquent plutôt un recours plus prononcé aux pédagogies « traditionnelles » accompagné d'une réduction de la diversité des activités éducatives (Au, 2007 ; Firestone, Mayrowetz & Fairman, 1998). À l'instar donc des études menées sur la pression marchande et l'innovation pédagogique qui ont forgé le concept de « *marketization* » (voir Lubienski, 2003), il apparaît que les nouvelles modalités d'intervention publique *via* l'évaluation externe ne se traduisent pas non plus chez les enseignants par des innovations pédagogiques et une réflexivité forte quant à leurs pratiques d'enseignement.

Ensuite, certaines préoccupations semblent davantage caractéristiques du type de dispositif ou du contexte dans lequel sont menées les études. Ainsi, notamment en raison de la polémique concernant la validité des résultats de l'évaluation externe, de nombreuses études menées dans le cadre des dispositifs *high-stakes* se sont penchées sur toute une série de pratiques non pédagogiques susceptibles de rendre compte de l'inflation des scores aux épreuves externes. Certaines pratiques frauduleuses (présentation des épreuves avant passation, transmission d'informations aidant à la résolution de certaines questions pendant passation, modification des réponses pendant correction, etc.), certains « vices de procédure » (placements en éducation spécialisée, redoublements) ou des pratiques de gestion stratégique de l'hétérogénéité de la classe (tri éducatif) apparaissent alors comme des pratiques susceptibles de mettre à mal la validité des scores comme indicateur de l'amélioration de l'enseignement. Dans ce sens, Linn précisait déjà il y a une quinzaine d'années que la « corruption des indicateurs est un problème permanent des situations où des épreuves sont utilisées à des fins de reddition de comptes » (2000, p. 5).

Du côté des dispositifs à faibles enjeux, c'est davantage le constat d'une faible utilisation des résultats par les enseignants qui aura orienté l'attention des études sur les perceptions des enseignants à l'égard des dispositifs d'évaluation externe. Ces études tendent à indiquer que même en l'absence d'une opposition déclarée aux épreuves externes, les résultats de celles-ci ne sont la plupart du temps pas considérés comme une information relative à leurs propres compétences professionnelles, ni comme une information utile à l'orientation des pratiques pédagogiques.

En fin de compte, la recension de l'ensemble des effets des épreuves standardisées constatés dans les études prises en considération pour cette note de synthèse permet-elle de conclure à l'efficacité des dispositifs d'évaluation externe quant à leur capacité à contribuer à l'amélioration de l'enseignement par la régulation des pratiques enseignantes ? Des études supplémentaires sont sans doute nécessaires pour donner une réponse claire à une telle question. Néanmoins, les données actuellement disponibles paraissent tout d'abord indiquer assez clairement les problèmes de l'utilisation des résultats aux épreuves comme indicateur de la qualité de l'enseignement en raison de la diversité des stratégies susceptibles d'être mises en place pour gonfler artificiellement les scores. En outre, les résultats de l'épreuve externe ne sont généralement pas perçus par les enseignants comme une source d'information pertinente

pour alimenter la réflexion concernant leurs pratiques pédagogiques. L'« effet miroir » attendu de l'épreuve externe ne semble donc pas opérer. Ensuite, si l'alignement de l'enseignement sur le contenu de l'épreuve ne doit pas être considéré comme un effet pervers en soi, il semble qu'un tel phénomène ne permet tout de même pas aux élèves de mieux faire face à des problèmes qui dépassent le cadre spécifique de l'épreuve pour laquelle ils ont été préparés. La contribution des dispositifs d'évaluation externe à l'amélioration de la qualité de l'enseignement mérite donc d'être débattue.

Cette « inefficacité » des évaluations externes à produire les effets pour lesquels elles ont été pensées pourrait en partie s'expliquer par des limites propres aux dispositifs. Certaines études prescrivent ainsi des ajustements susceptibles d'optimiser les effets du dispositif dans le sens désiré. Schildkamp et Kuiper (2010) concluent ainsi leur article en précisant que la promotion de l'utilisation des données dans les écoles pourrait passer par la formation des enseignants à l'analyse des données et par un soutien des directions à leur utilisation. Cannell (1989) proposait quant à lui toute une série de mesures « sécuritaires » susceptibles de limiter les « tricheries » commises par les enseignants dans le cadre de l'épreuve externe.

Et pourtant, l'incapacité des dispositifs d'évaluation externe, et surtout des résultats aux épreuves, à contribuer à l'amélioration de la qualité de l'enseignement se heurte probablement à d'autres obstacles de nature plus théorique. En effet, les recherches concernant les effets des *feedbacks* de performances permettent de questionner la capacité des résultats à améliorer les compétences professionnelles des enseignants. L'évolution de ce domaine de recherches n'est d'ailleurs pas sans intérêt pour la question des dispositifs d'évaluation externe. À l'image des attentes formulées à l'égard des résultats aux épreuves externes en effet, les *feedbacks* de performances – considérés comme une information relative à la performance d'une personne à une tâche – ont pendant longtemps été traités comme une information inévitablement liée à l'amélioration des performances. Ce n'est qu'à la fin des années 1990 que Kluger et DeNisi (1996) remettent en question ce présupposé dans une méta-analyse en démontrant que, pour 38 % des 607 effets de taille récoltés, le *feedback* aurait un impact négatif sur la performance, tandis que la moyenne des effets ne suggère qu'un impact positif limité. Un tel constat pourrait en partie s'expliquer par des facteurs motivationnels liés à la préservation du Soi (Jordan & Audia, 2012) : confronté à des résultats subjectivement décevants, un individu aurait tendance à développer des justifications cognitives (attribution causale externe, comparaisons descendantes, etc.) susceptibles de limiter la volonté de mettre en place des comportements pouvant améliorer la performance. Cependant, même la volonté d'intensifier les efforts à partir d'un *feedback* décevant ne garantirait pas nécessairement l'amélioration des performances. Notamment parce que, en fonction de la complexité de la tâche à réaliser, le *feedback* ne fournit pas les informations nécessaires à l'amélioration des performances dans tous les cas. Il n'est ainsi pas étonnant que les résultats aux épreuves externes ne fournissent pas aux enseignants les informations dont ils ont besoin pour améliorer leurs compétences professionnelles, d'autant plus que l'enseignement peut être considéré comme un ensemble de tâches au caractère particulièrement complexe.

Il convient toutefois de signaler que les *feedbacks* utilisés dans le cadre des manipulations expérimentales diffèrent généralement des résultats de l'évaluation externe dans le sens où la responsabilité de la performance est plus ambiguë dans ce dernier cas. Cette particularité des résultats de l'épreuve externe ne constitue pas pour autant un « avantage » car elle offre plus de possibilités à un acteur de recourir à des attributions causales externes : face à des résultats décevants, les enseignants seraient en effet susceptibles de faire reposer la responsabilité de la performance sur les élèves plutôt que sur leurs pratiques d'enseignement. Néanmoins, la possibilité pour les enseignants de négliger des résultats décevants dépend du type de dispositif auquel ils sont confrontés. Dans le cas où il s'agit d'un dispositif d'évaluation à faibles enjeux, l'absence de conséquences associées aux résultats de l'évaluation externe

donne dans une certaine mesure la possibilité de les négliger sans répercussions majeures. Dans le cadre d'un dispositif d'évaluation à enjeux élevés pour les enseignants par contre, nous pouvons supposer que l'injonction à l'amélioration des performances limite la possibilité pour un enseignant d'adopter une attitude de négligence absolue vis-à-vis des résultats. L'injonction à l'amélioration des résultats combinée au manque de contrôlabilité perçue est alors susceptible d'aboutir à une augmentation de l'anxiété ou à de la démotivation (Valli & Buese, 2007).

Mais finalement, comment expliquer l'engouement pour les dispositifs d'évaluation externe en dépit des limites théoriques et empiriques auxquelles ils sont confrontés ? À cet égard, l'hypothèse selon laquelle l'amélioration de la qualité de l'enseignement ne constitue pas le seul motif du développement de l'évaluation externe peut être soulevée. En effet, tout d'abord, la tendance inflationniste des résultats de l'évaluation externe fournit des indicateurs bien utiles aux acteurs politiques soucieux de préserver leur image publique. Dans le contexte nord-américain, Linn déclare dans ce sens que « de mauvais résultats sont désirables pour les décideurs politiques voulant démontrer qu'ils ont eu un impact. Sur la base de l'expérience du passé, les décideurs politiques peuvent raisonnablement s'attendre à une augmentation des scores dans les premières années d'un dispositif […] avec ou sans amélioration réelle concernant la maîtrise des contenus que les tests aspirent à mesurer. La représentation édulcorée dépeinte par des gains à court terme observée dans la plupart des nouveaux dispositifs donne l'impression d'une amélioration tombant à pic pour la campagne électorale suivante » (2000, p. 4).

Il est toutefois important de souligner que des scores élevés ne constituent pas en soi une garantie pour la préservation du prestige politique. Ainsi, les évaluations externes certificatives de fin de primaire imposées en Communauté française de Belgique en 2009 ont fait l'objet de vives polémiques relatives au niveau de difficulté étant donné que plus de 95 % des élèves étaient parvenus à réussir l'épreuve. Considérant que ces chiffres ne pouvaient correspondre au niveau réel des élèves de la Communauté, certains acteurs sociaux ont alors avancé que ces épreuves certificatives produisaient un nivellement par le bas.

Ensuite, dans le cadre des contraintes budgétaires imposées par la plupart des États pour répondre aux difficultés économiques, les dispositifs d'évaluation externe présentent l'avantage d'une solution relativement peu coûteuse en comparaison avec d'autres mesures susceptibles d'améliorer la qualité de l'enseignement, comme la réduction du nombre d'élèves par classe par exemple. Dans son étude relative au coût des dispositifs d'*accountability* nord-américains, Hoxby précise que « le coût des dispositifs d'*accountability* est à ce point négligeable que même l'analyse la plus généreuse ne pourrait le faire paraître important, en relation avec le coût d'autres réformes éducatives » (2002, p. 2). Cette analyse la plus généreuse semblerait indiquer que les dispositifs d'évaluation externe ne coûteraient pas plus de 5,81 $ par élève pour une dépense moyenne totale de 8 157 $ (par élève), ce qui équivaudrait à 0,07 % des dépenses totales dans l'enseignement primaire et secondaire.

À notre connaissance, aucune étude du genre n'est disponible dans le cadre européen. Toutefois, sans prétendre offrir autre chose qu'une analyse superficielle, en sachant que la Communauté française de Belgique annonce la dépense d'un minimum de 300 000 € pour toutes les épreuves externes mises en place dans le primaire[8], le coût de celles-ci s'élèverait à 0,92 € par élève, pour une dépense moyenne totale de 4 277 € par élève[9]. Ces 300 000 € constituent aussi 0,04 % du budget total attribué à l'enseignement obligatoire (primaire et secondaire) dans l'année académique 2011-2012, ce qui semble constituer une dépense relativement négligeable.

Aussi, si l'amélioration de la qualité de l'enseignement ne semble pas garantie par les dispositifs d'évaluation externe, ils contribuent tout de même à la réalisation d'un objectif

8 Qui comprennent donc des épreuves non certificatives en 3e et 5e primaire, ainsi qu'une épreuve certificative en 6e primaire.

9 En prenant en compte les indicateurs statistiques de l'année académique 2011-2012.

longtemps poursuivi par les réformes des contenus d'enseignement : celui de l'alignement entre les exigences institutionnelles et les pratiques « locales ». En effet, les recherches, notamment ethnographiques, ont depuis plusieurs décennies mis en évidence le fait que l'autonomie qui caractérise généralement le travail enseignant génère une relative étanchéité entre les programmes officiels et ce qui se fait concrètement dans la salle de classe (Coburn, 2004 ; Meyer & Rowan, 1977). L'impact des dispositifs d'évaluation externe sur les contenus enseignés, au travers de l'alignement de l'enseignement sur les contenus de l'épreuve, semble alors réduire cet écart entre les exigences institutionnelles et les pratiques enseignantes en même temps qu'elle produit une homogénéisation des pratiques enseignantes. Hallett (2010) fournit quatre raisons principales susceptibles d'expliquer ce processus de « recouplage » (*recoupling*) induit par les dispositifs d'évaluation externe. En premier lieu, au travers de la simplification de l'information par des mesures quantitatives, l'évaluation externe produirait une situation de commensurabilité qui canalise l'attention et l'action des acteurs de terrain. En deuxième lieu, la standardisation permettrait la création de seuils facilitant l'évaluation du travail enseignant. En troisième lieu, l'évaluation externe créerait une situation de surveillance qui rompt en quelque sorte l'autonomie dont bénéficiait l'enseignant une fois la porte de la salle de classe refermée. En fin de compte, l'alignement des pratiques serait souvent forcé par des récompenses et sanctions matérielles ; condition qui, d'après les données dont nous disposons dans le cadre des dispositifs d'évaluation externe à faibles enjeux, ne semble pas indispensable.

Conclusion

Dans cette note de synthèse, nous avons voulu approfondir la question de l'impact des dispositifs d'évaluation externe sur les pratiques enseignantes. À cette fin, nous avons rassemblé un ensemble de 77 articles empiriques décrivant majoritairement des dispositifs à forts enjeux nord-américains. Qu'en retenir ?

D'une part, toute une série de comportements stratégiques non pédagogiques spécifiquement orientés vers l'amélioration des scores aux épreuves externes ont retenu l'attention d'un nombre conséquent d'études en contexte nord-américain. Pratiques frauduleuses avant, pendant et après passation de l'épreuve, vices de procédures tels que l'utilisation abusive du redoublement ou du placement en éducation spécialisée, et gestion stratégique de l'hétérogénéité de la classe destinée à diriger les ressources et l'attention vers les élèves les plus susceptibles de relever la moyenne des scores, sont autant de pratiques susceptibles de remettre en cause la validité des résultats des épreuves standardisées. D'autre part, l'évaluation standardisée, quels que soient les enjeux qui y soient associés, semble exercer une influence significative sur les contenus enseignés au travers de pratiques telles que l'alignement de l'enseignement sur le contenu de l'épreuve, l'utilisation de matériel pédagogique similaire à l'épreuve et l'organisation de périodes spécifiques de préparation à l'épreuve. Ces pratiques, parfois désignées par la notion du *teaching to the test* (ou bachotage dans la littérature francophone) semblent toutefois être d'une efficacité limitée pour transmettre aux élèves des compétences dont l'utilité dépasse le cadre de l'épreuve spécifique à laquelle ils ont été préparés.

Par ailleurs, alors que l'innovation pédagogique constitue un objectif régulièrement associé aux dispositifs d'évaluation externe, il semble que leur influence à cet égard soit limitée. Au moment de prendre des décisions quant à aux méthodes pédagogiques, les enseignants ne se basent que de manière limitée sur les données issues des épreuves externes. Les études menées dans le cadre nord-américain semblent en outre indiquer que les épreuves standardisées contribuent au renforcement d'une pédagogie « traditionnelle » davantage centrée sur la transmission de savoirs de l'enseignant vers l'élève accompagnée d'une réduction de la diversité des activités pédagogiques utilisées par les enseignants.

Ainsi donc, il semble que la contribution de l'évaluation standardisée à l'amélioration de la qualité de l'enseignement puisse être questionnée. Cependant, au travers de l'alignement de l'enseignement sur le contenu de l'épreuve, les dispositifs d'évaluation externe parviendraient tout de même à accomplir un objectif longtemps recherché par les réformes curriculaires, à savoir le couplage entre exigences institutionnelles et pratiques locales. Ce même couplage des pratiques enseignantes contribuerait d'ailleurs à un renforcement de l'homogénéité des enseignements fournis aux élèves.

Finalement, soulignons que l'étude des pratiques enseignantes dans le cadre des dispositifs d'évaluation présente encore de nombreuses limites. Tout d'abord, les effets des dispositifs d'évaluation externe à faibles enjeux demeurent encore relativement peu étudiés. Nous manquons encore de données empiriques permettant d'appréhender les spécificités liées aux variations des enjeux associés aux résultats de l'évaluation standardisée. Il en va de même pour la question de l'impact des multiples caractéristiques susceptibles de varier d'un dispositif à l'autre qui mérite également d'être étudiée : épreuves certificatives ou diagnostiques, âge et fréquence de présentation des épreuves, forme de présentation des résultats, matières couvertes par l'épreuve, etc.

Par ailleurs, quelle que soit la méthodologie privilégiée dans les études futures, elle gagnerait à veiller à la question du faible taux de participation qui caractérise généralement les études menées auprès des enseignants confrontés aux épreuves externes. Pour des raisons pratiques, les études qualitatives par entretiens ou étude de cas ne permettent la plupart du temps que de collecter des données à partir d'échantillons restreints. Ainsi, l'étude de Booher-Jennings (2005), aussi intéressante et rigoureuse soit-elle, ne se centrait que sur un seul établissement texan. Dans le même ordre d'idées, les entretiens réalisés par Schildkamp et Kuiper (2010) se limitaient à 11 enseignants à peine. Le problème vaut toutefois également pour les études par questionnaires. Pour l'enquête en ligne réalisée par Amrein-Beardsley, Berliner et Rideau (2010) par exemple, seuls 5 % des 59 597 enseignants contactés avaient contribué au questionnaire en fin de compte. Dierendonck et Fagnant (2010), en dépit d'une approche plus directe des enseignants par des enveloppes individualisées, n'ont quant à eux obtenu de réponses que de la part de 19 % des 634 enseignants sollicités. De tels échantillons portent ainsi atteinte à la généralisation des études empiriques qui pourraient se voir reprocher de ne représenter que le point de vue de la part des enseignants les plus intéressés par la question des épreuves standardisées. En dernier lieu, de nombreuses études ont été menées à une période relativement proche de la mise en place du dispositif. Nous ne disposons donc que de rares données quant à l'évolution des pratiques enseignantes en contexte d'évaluations standardisées sur le long terme, raison pour laquelle des études supplémentaires, éventuellement longitudinales, mériteraient d'être menées.

Esteban Rozenwajn
Université catholique de Louvain, GIRSEF
esteban.rozenwajn@uclouvain.be

Xavier Dumay
Université catholique de Louvain, GIRSEF
xavier.dumay@uclouvain.be

Annexes

Tableau 1. Études empiriques menées aux États-Unis et en Angleterre

N°	Année de Publication	Auteurs	Contexte	Méthodologie
1	1984	Savage	États-Unis	Non précisée*
2	1985	Darling-Hammond & Wise	États-Unis	Analyse d'indicateurs quantitatifs**
3	1985	Frary & Olson	États-Unis	Études de cas
4	1985	Ligon	États-Unis	Analyse d'indicateurs quantitatifs
5	1985	Perlman	États-Unis	Non précisée
6	1987	Bracey	États-Unis	Questionnaires
7	1989	Romberg, Zarinnia & Williamq	États-Unis	Entretiens
8	1990	Gay	États-Unis	Études de cas
9	1991	Haladyna, Nolen & Haas	États-Unis	Études de cas
10	1990	Smith, Edelsky, Draper *et al.*	États-Unis	Questionnaires
11	1991	Aiken	États-Unis	Entretiens
12	1990	Shepard	États-Unis	Études de cas
13	1991	Cohen & Ball	États-Unis	Études de cas
14	1991a	Smith	États-Unis	Études de cas
15	1991b	Smith	États-Unis	Questionnaires
16	1991	Shepard & Dougherty	États-Unis	Études de cas
17	1992	Zancanella	États-Unis	Analyse d'indicateurs quantitatifs
18	1993	McGill-Franzen & Allington	États-Unis	Questionnaires
19	1994	Herman, Adebi & Golan	États-Unis	Entretiens et questionnaires
20	1995	Lomax, West, Harmon *et al.*	États-Unis	Études de cas
21	1996	Brown, Taggart, McCallum *et al.*	Royaume-Uni	Entretiens
22	1998	Firestone, Mayrowetz & Fairman	États-Unis	Entretiens
23	2000	Barksdale-ladd & Thomas	États-Unis	Analyse d'indicateurs quantitatifs
24	2000	Carnoy, Loeb & Smith	États-Unis	Analyse d'indicateurs quantitatifs
25	2000	Haney	États-Unis	Études de cas
26	2000	Landman	États-Unis	Entretiens
27	2000	Perreault	États-Unis	Analyse d'indicateurs quantitatifs
28	2002	Amrein & Berliner	États-Unis	Analyse d'indicateurs quantitatifs
29	2002	Carnoy & Loeb	États-Unis	Entretiens
30	2002	Costigan	États-Unis	Analyse d'indicateurs quantitatifs
31	2002	Figlio & Getzler	États-Unis	Non précisée
32	2002	Grant, Gradwell, Lauricella *et al.*	États-Unis	Études de cas
33	2002	Groves	États-Unis	Analyse d'indicateurs quantitatifs
34	2002	Toenjes & Dworkin	États-Unis	Études de cas

Nº	Année de Publication	Auteurs	Contexte	Méthodologie
35	2002	Wolf & Wolf	États-Unis	Études de cas
36	2002	Smagorinsky, Lalkly & Johnson	États-Unis	Analyse d'indicateurs quantitatifs
37	2003	Amrein-Beardsley & Berliner	États-Unis	Analyse d'indicateurs quantitatifs
38	2003	Jacob & Levitt	États-Unis	Questionnaires
39	2003	Taylor, Shepard, Kinner *et al.*	États-Unis	Non précisée
40	2003	Vogler	États-Unis	Entretiens
41	2003	Clarke, Shore, Rhoades *et al.*	États-Unis	Études de cas
42	2003	Anagnostopoulos	États-Unis	Questionnaires
43	2003	Sturman	Royaume-Uni	Études de cas
44	2003	Rex	États-Unis	Entretiens
45	2003	Segall	États-Unis	Études de cas
46	2004	Agee	États-Unis	Entretiens et questionnaires
47	2004	Bol	États-Unis	Analyse d'indicateurs quantitatifs
48	2004	Braun	États-Unis	Analyse d'indicateurs quantitatifs
49	2004	Jacob & Levitt	États-Unis	Études de cas
50	2005	Boyle & Bragg	Royaume-Uni	Études de cas
51	2005	Booher-Jennings	États-Unis	Analyse d'indicateurs quantitatifs
52	2005	Hanushek & Raymond	États-Unis	Entretiens et questionnaires
53	2005	Kirkup, Sizmur, Sturman *et al.*	Royaume-Uni	Entretiens et questionnaires
54	2005	Wright & Choi	États-Unis	Entretiens
55	2005	Yeh	États-Unis	Analyse d'indicateurs quantitatifs
56	2006	Nichols, Glass & Berliner	États-Unis	Études de cas
57	2006	Rentner, Scott, Kober *et al.*	États-Unis	Études de cas
58	2007	Valli & Buese	États-Unis	Études de cas
59	2007	Diamond	États-Unis	Entretiens et questionnaires
60	2010	Amrein-Beardsley, Berliner & Rideau	États-Unis	Entretiens et questionnaires
61	2010	Collins, Reiss & Strobart	Royaume-Uni	Études de cas
62	2012	Diamond	États-Unis	Études de cas

Notes :
* : la catégorie « non précisée » s'applique dans les cas où le résumé ne fournissait pas assez d'informations pour en déterminer la méthodologie et que l'article n'était pas accessible aux auteurs ;
** : les indicateurs quantitatifs peuvent être constitués par les scores des épreuves, les taux d'inscription, de redoublement ou des placements en éducation spécialisée.

Tableau 2. Études empiriques menées sur le continent européen

Nº	Année de	Auteurs	Contexte	Méthodologie
1	1998	Rochex	France	Études de cas
2	2003	Normand & Derouet	France	Études de cas
3	2006	Soussi, Nidegger, Ducrey *et al.*	Suisse	Entretiens et questionnaires
4	2008	Dierendonck	Suisse	Entretiens
5	2009	Lafontaine, Soussi & Nidegger	Suisse	Questionnaires
6	2009	Soussi, Guilley, Guignard *et al.*	Suisse	Entretiens
7	2009	Maier	Allemagne	Questionnaires
8	2009	Longchamp & Gilliéron	Suisse	Entretiens
9	2010	Dierendonck & Fagnant	Suisse/Belgique francophone	Questionnaires
10	2010	Maier	Allemagne	Entretiens
11	2010	Schildkamp & Kuiper	Pays-Bas	Entretiens
12	2012	Vanhoof, Verhaege & Van Petegem	Belgique néerlandophone	Questionnaires
13	2013	André	Suisse	Entretiens
14	2013	Baluteau	France	Entretiens
15	2013	Dutercq & Lanéelle	France	Études de cas

Bibliographie

AGEE J. (2004). « Negotiating a teaching identity: An african american teacher's struggle to teach in test-driven contexts ». *The Teachers College Record*, nº 106(4), p. 747-774.

AIKEN L. R. (1991). « Detecting, understanding and controlling for cheating on tests ». *Research in Higher Education*, nº 32(6), p. 725-736.

AMREIN A. L. & BERLINER D. C. (2002). « High-Stakes Testing, Uncertainty, and Student Learning ». *Educacion Policy Analysis Archives*, nº 10(18), p. 1-74.

AMREIN-BEARDSLEY A. L. & BERLINER D. C. (2003). « Re-analysis of NAEP Math and Reading Scores in States with and without High-Stakes Tests: Response to Rosenshine ». *Education Policy Analysis Archives*, nº 11(25), p. 1-16.

AMREIN-BEARDSLEY A. L., BERLINER D. C. & RIDEAU S. (2010). « Cheating in the first, second and third degree: Educators' responses to high-stakes testing ». *Educacion Policy Analysis Archives*, nº 18(14), p. 1-36.

ANAGNOSTOPOULOS D. (2003). « The new accountability, student failure and teachers' work in urban high schools ». *Educational Policy*, nº 17(3), p. 291-316.

ANDRÉ B. (2013). « Les ruses des enseignants face à l'évaluation des établissements et du système scolaire ». In V. Dupriez & R. Malet (dir.), *L'évaluation dans les systèmes scolaires*. Bruxelles : De Boeck, p. 35-62.

AU W. (2007). « High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis ». *Educational Researcher*, nº 36(5), p. 258-267.

BALUTEAU F. (2013). « Les dispositifs d'évaluation nationale et les réorganisations locales autour du curriculum ». In V. Dupriez & R. Malet (dir.), *L'évaluation dans les systèmes scolaires*. Bruxelles : De Boeck, p. 63-76.

BARKSDALE-LADD M. A. & THOMAS K. F. (2000). « What's at Stake in High-Stakes Testing: Teachers and Parents Speak Out ». *Journal of Teacher Education*, nº 51(5), p. 384-397.

BOL L. (2004). « Teachers' assessment practices in a high-stakes testing environment ». *Teacher Education and Practice.*, nº 17(2), p. 162-181.

BOOHER-JENNINGS J. (2005). « Below the bubble: "Educational triage" and the Texas Accountability System ». *American Educational Research Journal*, nº 42(2), p. 231-268.

BOYLE B. & BRAGG J. (2005). « No science today: The demise of primary science ». *Curriculum Journal*, nº 16(4), p. 423-437.

BRACEY G. W. (1987). « Measurement-driven instruction: Catchy phrase, dangerous practice ». *Phi Delta Kappan*, nº 68(9), p. 683-686.

BRAUN H. (2004). « Reconsidering the impact of high-stakes testing ». *Educacion Policy Analysis Archives*, nº 12(1), p. 1-43.

BROWN M., TAGGART B., MCCALLUM B. & GIPPS C. (1996). « The impact of key stage 2 tests ». *Education*, nº 24(3), p. 3-7.

CANNELL J. J. (1989). *How public educators cheat on standardized achievement tests: The "Lake Wobegon" report.* Albuquerque (États-Unis) : Friends for Education.

CARNOY M. & LOEB S. (2002). « Does external accountability affect student outcomes? A cross-state analysis ». *Educational Evaluation and Policy Analysis*, nº 24(4), p. 305-331.

CARNOY M., LOEB S. & SMITH T. (2000). *Does higher state test scores in Texas make for better high school outcomes?*. New Orleans : American Educational Research Association Annual Meeting.

CATTONAR B., DUMAY X. & MANGEZ C. (2010). *Évaluations externes dans l'enseignement primaire en Belgique francophone. Réceptions et usages d'outils de régulation basés sur les connaissances*. Louvain-la-Neuve (Belgique) : Université catholique de Louvain.

CLARKE M., SHORE A., RHOADES K., ABRAMS L., MIAO J. & LI J. (2003). *Percieved effects of state-mandated testing programs on educators in low-, medium-, and high stakes states.* Chestnut Hill (États-Unis) : National Board on Educational Testing and Public Policy.

COBURN C. E. (2004). « Beyond Decoupling: Rethinking the relationship between the institutional environment and the classroom ». *Sociology of Education*, nº 77, p. 211-244.

COHEN D. K. & BALL D. L. (1990). « Policy and practice: An overview ». *Educational Evaluation and Policy Analysis*, nº 12(3), p. 233-239.

COLLINS S., REISS M. & STOBART G. (2010). « What happens when high-stakes testing stops? ». *Assessment in Education: Principles Policy & Practice*, nº 17(3), p. 273-286.

COSTIGAN A. T. (2002). « Teaching the culture of high-stakes testing: Listening to new teachers ». *Action in Teacher Education*, nº 23(4), p. 28-34.

DARLING-HAMMOND L. & WISE A. E. (1985). « Beyond standardization: State, standards and school improvement ». *The Elementary School Journal*, nº 85, p. 315-336.

DIAMOND J. B. (2007). « Where the rubber meets the road: Rethinking the connection between high-stakes testing policy and classroom instruction ». *Sociology of Education*, nº 80(4), p. 285-313.

DIAMOND J. B. (2012). « Accountability policy, school organization, and classroom practice: Partial recoupling and educational opportunity ». *Education and Urban Society*, nº 44(2), p. 151-182.

DIERENDONCK C. (2008). *Comment les évaluation externes des acquis des élèves sont-elles perçues par les enseignants du primaire dans les cantons de Neuchâtel, Vaud et Fribourg ? Enquête exploratoire.* Neuchâtel : IRDP.

DIERENDONCK C. & FAGNANT A. (2010). « Quelques réflexions autour des épreuves d'évaluation développées dans le cadre de l'approche par compétences ». *Le Bulletin de l'ADMEE-EUROPE*, nº 1.

DUPRIEZ V. & MALET R. (2013). *L'évaluation dans les systèmes scolaires.* Bruxelles : De Boeck.

DUTERCQ Y. & LANÉELLE X. (2013). « La dispute autour des évaluations des élèves dans l'enseignement français du premier degré ». *Sociologie*, nº 4, p. 43-62.

ELMORE R. F., ABELMANN C. H. & FUHRMAN S. H. (1996). « The new accountability in state education reform: From process to performance ». In H. F. Ladd (dir.), *Holding schools accountable: Performance-based reform in education*. Washington : The Brookings Institution, p. 65-98.

EURYDICE (2009). *Les évaluations standardisées des élèves en Europe : objectifs, organisation et utilisation des résultats*. Bruxelles : Eurydice.

FIGLIO D. N. & GETZLER L. S. (2002). *Accountability, ability and disability: Gaming the system*. Cambridge (États-Unis) : National Bureau of Economic Research.

FIRESTONE W. A., MAYROWETZ D. & FAIRMAN J. (1998). « Performance-based assessment and instructional change: The effects of testing in Maine and Maryland ». *Educational Evaluation and Policy Analysis*, nº 20(2), p. 95-113.

FRARY R. B. & OLSON G. H. (1985). « Detection of coaching and answer copying on standardized tests ». Chicago : American Educational Research Association Annual Meeting.

GAMORAN A. (2012). « Bilan et devenir de la loi *No Child Left Behind* aux États-Unis ». *Revue française de pédagogie*, nº 178, p. 13-26.

GAY G. H. (1990). « Standardized tests: Irregularities in administering of test affect test results ». *Journal of Instructional Psychology*, nº 17(2), p. 93-103.

GERWIN D. & VISONE F. (2006). « The freedom to teach: Contrasting history teaching in elective and state-tested courses ». *Theory & Research in Social Education*, nº 34(2), p. 259-282.

GILLBORN D. & YOUDELL D. (2000). *Rationing education: Policy, practice, reform and equity*. Buckingham : Open University Press.

GRANT S. G., GRADWELL J. M., LAURICELLA A. M., DERME-INSINNA A., PULLANO L. & TZETZO K. (2002). « When increasing stakes need not mean increasing standards: The case of the New York state global history and geography exama ». *Theory & Research in Social Education*, nº 30(4), p. 488-515.

GROVES P. (2002). « Doesn't feel morbid here? High-stakes testing and the widening of the equity gap ». *Educational Foudations*, nº 16(2), p. 15-31.

HALADYNA T. M., NOLAN S. B. & HAAS N. A. (1991). « Raising standardized test scores and the origins of test score pollution ». *Educational Researcher*, nº 20(5), p. 2-7.

HALLETT T. (2010). « The Myth Incarnate: Recoupling Processes, Turmoil, and Inhabited Institutions in an Urban Elementary School ». *American Sociological Review*, nº 75(1), p. 52-74.

HANEY W. (2000). « The Myth of the Texas Miracle in Education ». *Education Policy Analysis Archives*, nº 8(41).

HANUSHEK E. A. & RAYMOND M. E. (2002). « Sorting out accountability systems ». In W. M. Evers & H. J. Walberg (dir.), *School accountability*, p. 75-104. Stanford (États-Unis) : Hoover Institution Press.

HANUSHEK E. A. & RAYMOND M. E. (2005). « Does School Accountability Lead to Improved Student Performance? Does School Accountability Lead to Improved Student Performance? ». *Journal of Policy Analysis and Management*, nº 24(2), p. 297-327.

HARRIS D. N. & HERRINGTON C. D. (2006). « Accountability, standards, and the growing achievement gap: Lessons from the past half-century ». *American Journal of Education*, nº 112(2), p. 209-238.

HELLRUNG K. & HARTIG J. (2013). « Understanding and using feedback-A review of empirical studies concerning feedback from external evaluations to teachers ». *Educational Research Review*, nº 9, p. 174-190.

HERMAN J. L., ABEDI J. & GOLAN S. (1994). « Assessing the effects of standardized testing on schools ». *Educational and Psychological Measurement*, nº 54(2), p. 471-482.

HOUSE OF COMMONS (2008). *Testing and Assessment (Vol. I)*.Londres : The Stationary Office Limited.

HOUSE OF COMMONS (2009). *School accountability: First report of session 2009-10*. Londres : The Stationary Office Limited.

HOXBY C. M. (2002). *The cost of accountability*. Cambridge (États-Unis) : National Bureau of Economic Research.

IGEN-IGAENR (2005). *Les acquis des élèves, pierre de touche de la valeur de l'école ?* Paris : Ministère de l'Éducation nationale.

JACOB B. & LEVITT S. (2003). « Rotten apples: An investigation of the prevalence and predictors of teacher cheating ». *The Quarterly Journal of Economics*, nº 118(3), p. 843-877.

JACOB B. A. & LEVITT S. D. (2004). « To catch a cheat ». *Education Next*, nº 4(1), p. 68-75.

JORDAN A. H. & AUDIA P. G. (2012). « Self-Enhancement and Learning from Performance Feedback ». *Academy of Management Review*, nº 37(2), p. 211-231.

KEILLOR G. (1985). *Lake Wobegon Days*. New York : Viking.

KIRKUP C., SIZMUR J., STURMAN L. & LEWIS K. (2005). *Schools' use of data in teaching and learning.* Berkshire : National Foundation for Educational Research.

KLIEME E. (2004). *Le développement de standards nationaux de formation : Une expertise.* Bonn : Ministère fédéral de l'Éducation et de la Recherche.

KLUGER A. N. & DENISI A. (1996). « The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory ». *Psychological Bulletin*, nº 119(2), p. 254-284.

KORETZ D. M. (1991). *The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests*. Chicago :American Educational Research Association Annual Meeting, p. 85.

LAFONTAINE D., SOUSSI A. & NIDEGGER C. (2009). « Évaluations internationales et/ou épreuves nationales : tensions et changements de pratiques ». In M. L. Lopez & M. Crahay (dir.), *Évaluations en tension*. Bruxelles : De Boeck, p. 61-80.

LANDMAN J. (2000). *A state-mandated curriculum, a high-stakes test: One massachusetts high school history department's response to a very new policy context.* Travail de fin d'études pour la *Harvard Graduate School ofEducation*

LEE J. (2008). « Is Test-Driven External Accountability Effective? Synthesizing the Evidence From Cross-State Causal-Comparative and Correlational Studies ». *Review of Educational Research*, nº 78(3), p. 608-644.

LIGON G. (1985). *Opportunity knocked out: Reducing cheating by teachers on student tests*. Chicago :American Educational Research Association Annual Meeting.

LINN R. L. (2000). « Assessments and accountability ». *Educational Researcher*, nº 29(2), p. 4-16.

LOMAX R. G., WEST M. M., HARMON M. C., VIATOR K. A & MADAUS G. F. (1995). « The impact of mandated standardized testing on minority students ». *Journal of Negro Education*, nº 64(2), p. 171-185.

LONGCHAMP A.-L. & GILLIÉRON P. G. (2009). *Épreuves cantonales de référence de mathématiques en 6e année : le point de vue des enseignants sur l'utilisation du dispositif*. Lausanne : DFJC-URSP.

LOPEZ L. M. & CRAHAY M. (2009). *Évaluations en tension : entre la régulation des apprentissages et le pilotage des systèmes*. Bruxelles : De Boeck.

LUBIENSKI C. (2003). « Instrumentalist Perspectives on the "Public" in Public Education: Incentives and Purposes ». *Educational Policy*, nº 17(4), p. 478-502.

MAIER U. (2009). « Towards state mandated testing in Germany: how do teachers assess the pedagogical relevance of performance feedback information? » *Assessment in Education: Principles, Policy & Practice*, nº 16(2), p. 205-226.

MAIER U. (2010). « Accountability policies and teachers' acceptance and usage of school performance feedback: A comparative study ». *School Effectiveness and School Improvement*, nº 21(2), p. 145-165.

MAROY C. & VOISIN A. (2013). « Une typologie des politiques d'accountability en éducation : l'incidence de l'instrumentalisation et des théories de la régulation ».*Educação & Sociedade*, nº 124.

MCGILL-FRANZEN A. & ALLINGTON R. L. (1993). « Flunk'em or get them classified: The contamination of primary grade accountability data ». *Educational Researcher*, nº 22(1), p. 19-22.

MEYER J. W. & ROWAN B. (1977). « Institutionalized Organizations: Formal Structure as Myth and Ceremony ». *American Journal of Sociology*, nº 83(2), p. 340.

MONS N. (2009). « Les effets théoriques et réels de l'évaluation standardisée ». Bruxelles : Agence Exécutive Éducation, Audiovisuel et culture (EACEA).

MONS N. & PONS X. (2006). *Les standards en éducation dans le monde francophone : une analyse comparative*. Neuchâtel : IRDP.

NICHOLS S. & BERLINER D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools.* Cambridge (États-Unis) : Harvard Education Press.

NICHOLS S., GLASS G. & BERLINER D. (2006). « High-stakes testing and student achievement: Does accountability pressure increase student learning? ». *Educacion Policy Analysis Archives*, nº 14(1), p. 1-172.

NORMAND R. & DEROUET J.-L. (2003). *Le développement d'une culture de l'évaluation dans l'Éducation nationale : comment les enseignants utilisent-ils les résultats des évaluations nationales ?* Lyon : INRP.

PERLMAN C. L. (1985). *Results of a citywide testing program audit in Chicago*. Chicago : American Educational Research Association Annual Meeting.

PERREAULT G. (2000). « The classroom impact of high-stress testing ». *Education*, nº 120(4), p. 705-710.

PHELPS R. (2005). *Defending standardized testing.* Mahwah (États-Unis) : Lawrence Erlbaum.

POPHAM J. W. (1987). « The merits of measurement-driven instruction ». *Phi Delta Kappan*, nº 68(9), p. 679-682.

POPHAM J. W., CRUSE K. L., RANKIN S. C., SANDIFER P. D. & WILLIAMS P. L. (1985). « Measurement-driven instruction: It's on the road ». *Phi Delta Kappan*, nº 66(9), p. 628-634.

RAPPLE B. (1994). « Payment by results: An example of assessment in elementary education from nineteenth century Britain ». *Educacion Policy Analysis Archives*, nº 2(1).

RAYMOND M. E. & HANUSHEK E. A. (2003). « High Stakes research ». *Education Next*, nº 3(3), p. 48-55.

RENTNER D. S., SCOTT C., KOBER N., CHUDOWSKY N., CHUDOWSKY V., JOFTUS S. & ZABALA D. (2006). *From the capital to the classroom: Year 4 of the No Child Left Behind Act.* Washington : CEP. En ligne : <http://www.cep-dc.org/displayDocument.cfm?DocumentID=301> (consulté le 7 septembre 2015).

REX L. A. (2003). « Loss of the creature ». *Communication Education*, nº 52(1), p. 30-46.

ROCHEX J.-Y. (1998). « Culture de l'évaluation et activité enseignante ». *Publication numérique Société française*, nº 60. En ligne : <http://revuesshs.u-bourgogne.fr/societe_francaise/document.php?id=1765> (consulté le 7 septembre 2015).

ROMBERG T. A., ZARINNIA A. E. & WILLIAMS S. R. (1989). *The influence of mandated testing on mathematics instruction: Grade 8 teachers' perceptions*. Wisconsin : National Center for Research in Mathematical Sciences Education.

ROSENSHINE B. (2003). « High-Stakes Testing : Another Analysis Barak Rosenshine University of Illinois at Urbana , Champaign ». *Education Policy Analysis Archives*, nº 11(24), p. 1-8.

ROZENWAJN E. (à paraître). *Low-stakes testing and teacher practices : A mirror effect ?*

RUMBERGER R. W. (1995). « Dropping Out of Middle School: A Multilevel Analysis of Students and Schools ». *American Educational Research Journal*, nº 32(3), p. 583-625.

SAVAGE D. (1984). « Scrutinize students' test scores, and they might not look so rosy ». *American School Board Journal*, nº 171(8), p. 21-24

SCHILDKAMP K. & KUIPER W. (2010). « Data-informed curriculum reform: Which data, what purposes, and promoting and hindering factors ». *Teaching and Teacher Education*, nº 26(3), p. 482-496.

SEGALL A. (2003). « Teachers' perceptions of the impact of state-mandated standardized testing: The Michigan Educational Assessment Program (MEAP) as a case study of consequences ». *Theory & Research in Social Education*, nº 31(3), p. 287-325.

SHEPARD L. A. (1990). « Inflated test score gains: Is the problem old norms or teaching the test? ». *Educational Measurement: Issues and Practice*, nº 9(3), p. 15-22.

SHEPARD L. A. & DOUGHERTY K. C. (1991). « Effects of high-stakes testing on instruction ». In *American Educational Research Association Annual Meeting*. Washington : Office of Educational Research and Improvement.

SMAGORINSKY P., LAKLY A. & JOHNSON T. S. (2002). « Acquiesence, accomodation, and resistance in learning to teach within prescribed curriculum ». *English Education*, nº 34(3), p. 187-213.

SMITH L. M. (1991a). « Meanings of Test Preparation ». *American Educational Research Journal*, nº 28(3), p. 521-542.

SMITH L. M. (1991b). « Put to the test: The effects of external testing on teachers ». *Educational Researcher*, nº 20(5), p. 8-11.

SMITH L. M., EDELSKY C., DRAPER K., ROTTENBERG C. & CHERLAND M. (1990). *The role of testing in elementary schools.* Los Angeles : Center for Research on Evaluation, Standards and Student Testing.

SOUSSI A., GUILLEY E., GUIGNARD N. & NIDEGGER C. (2009). *Évaluation des acquis des élèves à l'école obligatoire*. Genève : SRED.

SOUSSI A., NIDEGGER C., DUCREY F., FERREZ E. & VIRY G. (2006). *Pratiques d'évaluation : ce qu'en disent les enseignants (à l'école obligatoire et dans l'enseignement postobligatoire général).* Genève : SERD.

STURMAN L. (2003). « Teaching to the test: Science or intuition? ». *Educational Research*, nº 45(3), p. 261-273.

TAYLOR G., SHEPARD L. A., KINNER F. & ROSENTHAL J. (2003). *A survey of teachers' perspectives on high-stakes testing in Colorado: What gets taught, what gets lost*. Los Angeles : National Center for Research on Evaluation, Standards and Student Testing.

THÉLOT C. (2002). « Évaluer l'école ». *Études*, nº 10(397), p. 323-334.

TOENJES L. A. & DWORKIN G. A. (2002). « Are increasing test scores in Texas really a myth? ». *Educational Policy Analysis Archives*, nº 10(17).

US DEPARTMENT OF EDUCATION (2009). *Race to the Top Program: Executive Summary.* Whasinghton : US Department of education. En ligne : <https://www2.ed.gov/programs/racetothetop/executive-summary.pdf> (consulté le 7 septembre 2015).

VALLI L. & BUESE D. (2007). « The changing roles of teachers in an era of high-stakes accountability ». *American Educational Research Journal*, nº 44(3), p. 519-558.

VANHOOF J., VERHAEGHE G. & VAN PETEGEM P. (2012). « Flemish primary teachers' use of school performance feedback and the relationship with school characteristics ». *Educational Research*, nº 54(4), p. 431-449.

VERDIÈRE J. (2013). « Les enseignants du collège en France. Faire face à un travail plus formalisé rendu plus visible ». In V. Dupriez & R. Malet (dir.), *L'évaluation dans les systèmes scolaires*. Bruxelles : De Boeck, p. 77–90.

VOGLER K. E. (2003). « An integrated curriculum using state standards in a high-stakes testing environment ». *Middle School Journal*, nº 34(4), p. 5-10.

WOESSMAN L. (2007). « International evidence on school competition, autonomy and accountability: A review ». *Peabody Journal of Education*, nº 82(2-3), p. 473-497.

WOLF S. A. & WOLF K. P. (2002). « Teaching true and to the test in writing ». *Language Arts*, nº 79(3), p. 229-240.

WRIGHT W. E. & CHOI D. (2005). *Voices from the classroom: A statewide survey of experienced third-grade english laguage learner teachers on the impact of laguage and high-stakes testing policies in Arizona.* Tempe (États-Unis) : Language Policy Research Unit.

YEH S. (2005). « Limiting the unintended consequences of high-stakes testing ». *Education Policy Analysis Archives*, nº 13(43).

ZANCANELLA D. (1992). « The influence of state-mandated testing on teachers of literature ». *Educational Policy and Policy Analysis*, nº 14(3), p. 283-295.